НАЦИОНАЛЬНАЯ АКАДЕМИЯ НАУК ТАДЖИКИСТАНА

Институт математики имени А. Джураева

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РЕСПУБЛИКИ ТАДЖИКИСТАН

Таджикский технический университет имени академика М.С. Осими

УДК 811::81’33::519.25

На правах рукописи

КОСИМОВ Абдунаби Абдурауфович

СТАТИСТИЧЕСКИЕ ЗАКОНОМЕРНОСТИ РАСПОЗНАВАНИЯ ОДНОРОДНОСТИ ТЕКСТОВ С ПОМОЩЬЮ γ-КЛАССИФИКАТОРА

ДИССЕРТАЦИЯ

на соискание ученой степени доктора технических наук

по специальности **05.13.11** – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»

Научный консультант: Усманов Зафар Джураевич

доктор физико-математических наук,

академик НАНТ, профессор

Душанбе – 2024

СОДЕРЖАНИЕ

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

АОТ – Автоматизированная обработка текстов

БД – База данных

ВМК МГУ – Факультет вычислительной математики и кибернетики Московского государственного университета имени М.В. Ломоносова

ВУЗ – Высшее учебное заведение

ДАН РТ – Доклады Академии наук Республики Таджикистан

ДНАНТ – Доклады Национальной академии наук Таджикистана

ЕЯ – Естественный язык

ИАН РТ – Известия Академии наук Республики Таджикистан

ИКТ – Информационно-коммуникационные технологии

ИМ НАНТ – Институт математики имени А. Джураева, Национальная академия наук Таджикистана

КЛ – Компьютерная лингвистика

МБС – Метод ближайшего соседа

МО – Машинное обучение

МП – Машинный перевод

ПП – Программный продукт

РТ – Республика Таджикистан

РТСУ – Российско-Таджикский (Славянский) университет

ТТУ – Таджикский технический университет имени академика М.С. Осими

ТУТ – Технологический университет Таджикистана

ТЯ – Таджикский язык

ЦАР – Центрально-Азиатский регион

ЦП – Цифровой портрет

ЦПА – Цифровой портрет автореферата

ЦПП – Цифровой портрет произведений

ЦПТ – Цифровой портрет текста

ЭВМ – Электронно-вычислительная машина

LSTM – Long short-term memory

THR – Text Homogeneity Recognition

URL – Uniform Resource Locator

ВВЕДЕНИЕ

Актуальность темы исследования. Настоящая диссертация является составной частью глобальной научной проблемы – автоматической обработки информации на естественном языке, признанной одной из актуальных проблем современной науки. С надеждами на успешное разрешение последней связан вопрос о способности сегодняшний цивилизации упорядочивать, контролировать, использовать и осмысливать лавинообразный приток знаний, порождаемый её собственной деятельностью.

Одной из граней этой проблемы является проектирование и разработка автоматических систем определения адресности и новизны информации, охватывающих такие вопросы, как плагиат, заимствование, компиляция, идентификация авторства, подобие произведения и его перевода и т.п. В связи с развитием информационных технологий исследования в этой области знания заметно интенсифицировались по всему миру. Многочисленные научные публикации во всех высокоразвитых странах показывают особую роль данной проблематики, её непосредственное влияние на развитие науки и техники, на прогресс в сфере искусственного интеллекта, на широкомасштабные приложения в мировой экономике.

Именно в этом заключается актуальность выбора темы настоящей диссертации, что подтверждается также и постановлением Правительства Республики Таджикистан «Об утверждении программы применения и развития информационных технологий в таджикском языке» от 06.06.2005, № 188, Указом Президента Республики Таджикистан об объявлении 2020-2040 гг. «Двадцатилетием изучения и развития естественных, точных и математических наук в сфере науки и образования» от 31.01.2020, №1445, и поручением, озвученным Президентом Республики Таджикистан, Лидером нации, уважаемым Эмомали Рахмоном в своем ежегодном Послании Маджлиси Оли о принятии и реализации Национальной стратегии развития искусственного интеллекта для разработки и широкого использования современных технологий в различных сферах экономики страны, 21 декабря 2021 года.

Степень научной разработанности изучаемой проблемы. Актуальность обозначенной научной проблемы подтверждается теоретическими и практическими работами таджикских и зарубежных исследователей. Теоретическая значимость проблемы связана с изучением комплекса вопросов формирования и исследования пригодности ЦП на основе распределения частотности различных алфавитных элементов текста для распознавания новизны, компиляции, плагиата, заимствования, идентификации авторства и шифров научных работ. Актуальность подобных работ связана с определением особых характеристик текста, которые, не будучи подконтрольны своим создателям,

содержат в себе косвенную информацию об авторском стиле и даже индивидуальных качествах автора. Практическая значимость проблемы имеет отношение к государственной административной деятельности, в которой на передний план выдвигается автоматическая обработка текстовой информации; к криминалистике, заинтересованной в установлении преступника по составу преступления и авторов анонимных текстов; к сфере образования и науки, в которых и студенческая молодежь и псевдонаучные работники не прочь воспользоваться компиляцией, заимствованиями, плагиатом при выполнении курсовых и дипломных проектов, представлении к защите кандидатских и докторских диссертаций.

Между тем, в дальнем зарубежье работы в этой области знания заметно интенсифицировались в связи с развитием информационных технологий. В подтверждение этого факта достаточно обратиться к трудам J. Rudman, J. Burrows, R. Zheng, P. Juola, A.Q. Morton, T.C. Mendenhall, A. Abbasi, J.J. Diederich, M.F. Amasyah, E. Stamatatos, D. Lowe, C. Apte, M. Corney, S. Argamon, F.J. Tweedie, R.H. Baayen, O. De Vel, C.E. Chaski, B. Allison, D. Guthrie, L. Guthrie, Y. Bengio, P. Simard, P. Frasconi, D. Russell, A. Gray, Q.D. Atkinson, W. Chang, Ch. Cathcart, D. Hall, A. Garrett, A. Kassian, A. Dybo, K. Calix, W.M. Hadi, J.R. Karr, J.J. Hughey, T.K. Lee, S. Hochreiter, J. Schmidhuber, T. Mikolov, S. Ioffe, C. Szegedy, B. Efron, J.M. Farringdon, T. Joachims, B. Kjell, R.D. Peng, M. Koppel, K. Luyckx, R. Matthews, F. Peng, W.J. Teahan и S. Waugh, [1-70].

В России подобным вопросам посвящены исследования А.А. Шелупанова, Р.В. Мещерякова, А.С. Романова, А.В. Куртуковой, А.В. Пруцкова, Л.С. Ломакиной, А.В. Мордвинова, А.С. Сурковой, Д.В. Ломакина, А.З. Панкратовой, В.Б. Родионова, С.С. Буденкова, М.С. Семенцова, М.Д. Ломакиной, А.А. Царева, С.С. Скорынина, И.Д. Чернобаева, А.А. Домнина, В.В. Поддубного, В.П. Фоменко, Т.Г. Фоменко, Н.А. Морозова, А.А. Маркова, Д.В. Хмелева, Е.И. Большаковой, А.А. Носкова, О.В. Песковой, Е.В. Ягуновой, В.В. Александрова, Л.Л. Иомдина, М.В. Арапова, В.К. Финна, А.А. Барсегяна, М.С. Куприянова, И.И. Холода, А.И. Башмакова, В.С. Белова, Г.Г. Белоногова, А.А. Хорошилова, Ю.Г. Зеленкова, А.П. Новоселова, Б.А. Кузнецова, М.Б. Болдина, Г.И. Симоновой, Ю.Н. Тюрина, А.А. Большакова, Р.Н. Каримова, А.А. Боровкова, И.И. Быстрова, Б.В. Тарасова, С.И. Радоманова, В.Н. Вапника, А.Я. Червоненкиса, Н.К. Верещагина, В.Н. Волковой, А.А. Денисова, Т.А. Гавриловой, А.С. Дмитриева, А.П. Еремеева, Н.Г. Загоруйко, Л.А. Заде, М. Кендалла, А. Стьюарта, А.Н. Кирдина, А.Ю. Новоходько, В.Г. Царегородцева, А.Н. Колмогорова, А.С. Костышина, В.Н. Кучуганова, И.В. Безсуднова, Д.В. Ландэ, Э. Лемана, А.В. Леоненкова, Н.Н. Леонтьевой, Н.В. Лукашевича, Г.Я. Мартыненко, А.С. Мельничука, Л.Н. Мурзина, А.С. Штерна, Г.В. Напреенко, В.А. Негуляева, А.А. Орлова, А.И. Орлова, А.А. Поликарпова, И.Н. Пономаренко, Д.М. Цыбулько, А.П.

Рыжова, Ю.Б. Сафроновой, И.П. Севбо, Э.Ф. Скороходько, Ю.Г. Сметанина, М.В. Ульянова, А.С. Пестовой, Г.Я. Солганика, В.М. Солнцева, А.А. Харкевича, Г. Хьетсо, Я.З. Цыпкина, И.Г. Чекунова, А.А. Рогова, Ю.В. Сидорова, А.Ю. Комиссарова, Е.В. Шараповой, Р.В. Шарапова, О.Г. Шевелева, М.А. Марусенко, Ю.Н. Павлова, А.В. Седова, Е.А. Тихомировой, В.В. Дягилева, А.А. Цхая, А.О. Шумской, С.В. Бутакова и З.И. Резановой, [71-254].

Вопросами распознавания однородности текста в Таджикистане, в частности занимались и занимаются З.Д. Усманов, Х.А. Тошхуджаев, Х.Т. Максудов, М.А. Умаров, М.А. Исмоилов, Х.А. Худойбердиев, О.М. Солиев, Ш.Н. Ашурова, Г.М. Довудов, А.А. Каримов, М.М. Каюмов, П.Э. Зульфикарова, Дж.Х. Баховудинов, С.М. Пиров, Н.М. Курбонов, М.Ё. Мухсинзода, Н.О. Косимова, О.А. Косимов, Б.Б. Иномов, Д.Э. Косимов, М.М. Фозилова, Ш.С. Саидов, Д.Н. Комилов и К.С. Бахтеев, [255-324, 1-А-73-А].

Все это говорит об актуальности избранной темы диссертации, в частности потому, что исследования в столь важном направлении находятся в Таджикистане на стадии становления и в ближайшем будущем напрямую будут связываться с разработкой государственной системы информационной безопасности.

Настоящая диссертация посвящена изучению проблемы распознавания однородности текстовых фрагментов на основе γ-классификатора.

Связь работы с научными программами (проектами), темами. Данное диссертационное исследование выполнено в рамках реализации следующих проектов научно-исследовательских работ института математики им. А. Джураева Национальной Академии наук Таджикистана:

– «Разработка и исследование математических моделей для решения прикладных и практических задач», ГР 0116TJ00533, с 2016 года по 2020 год;

– «Исследование актуальных задач прикладной математики и информатики», ГР 0121TJ1180, с 2021 года по 2025 год.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель работы – алгоритмизировать процесс распознавания однородности текстов и реализовать его в виде компьютерного программного комплекса.

Задачи исследования. Для достижения цели решаются следующие задачи:

1) сформировать две электронные коллекции текстов, из которых первая предназначена для предварительного тестирования, а вторая – для оценки перспективности применения γ-классификатора;

2) исследовать цифрового портрета текста (ЦПТ) для распознавания автора текста;

3) установить статистическую эффективность применения γ-классификатора для распознавания авторов произведений;

4) определить минимальный размер незнакомого текста, пригодного для распознавания его автора;

5) исследовать эффективность применения высокочастотных элементов ЦПТ для идентификации автора текста;

6) установить статистическую эффективность применения γ-классификатора и исследования пригодности ЦП на основе распределения частотности различных алфавитных элементов текста для распознавания других признаков однородности, таких как тематики текста, язык, группа языков, оригинал и его перевод, стиль произведений, шифры научных работ и т.д.;

7) исследовать статистические закономерности распознавания однородных текстов на корпусах художественных литературных произведений;

8) определить эффективность применения γ-классификатора для атрибуции искусственно сгенерированных произведений авторов;

9) исследовать влияние порядка ЦП текста на распознавание однородности произведения с помощью γ-классификатора;

10) спроектировать и реализовать компьютерный программный комплекс для распознавания (идентификации) однородности текста на основе различных ЦП текста и γ-классификатора.

Объект исследования – корпус печатных текстов и его характеристики на разных языках.

Предмет исследования – распознавание однородности произведения на основе γ-классификатора (математической триады) и частотности различных характеристик текста.

Методы исследования. Для решения задач, указанных в рубрике «Цель работы», использовались машинное обучение, кодирование информации, методы математической статистики, вычислительного эксперимента, теории множеств, системного анализа, распознавания и объектно-ориентированного программирования для разработки программных средств.

Научная новизна диссертации состоит в следующем:

1) исследована информативность нетрадиционных признаков на предмет количественного описания таджикских текстов;

2) установлена статистическая эффективность π математической модели опознавания авторов произведений таджикской классической поэзии (π = 1.00) на основе триграмм, современной поэзии (π = 0.98) с помощью униграмм и современной прозы (π = 0.96) на основе распределения длин предложений (в словах);

3) установлена 100%-ная статистическая эффективность путем применения метрического γ-классификатора и метода ближайшего (по расстоянию) соседа идентифицировать авторов произведений – убывающих по размерам последовательности текстовых фрагментов от величины в 7000 слов (40000 символов) вплоть до 20 слов (100 символов);

4) для целей существенного сокращения объёма вычислительных процедур установлена возможность эффективного использования не всех, а только высокочастотных элементов ЦП текстов;

5) установлена статистическая эффективность применения γ-классификатора и исследована пригодность ЦП на основе распределения частотности различных алфавитных элементов текста для распознавания других признаков однородности, таких как тематики текста, язык, группа языков, оригинал и его перевод, стиль произведений и шифры научных работ;

6) исследованы статистические закономерности опознавания авторов и языков произведений на корпусах художественных литературных произведений с помощью γ-классификатора;

7) γ-классификатор и метод ближайшего соседа были протестированы на случайных выборках текстов, распознаются с достаточно высокой точностью признаки однородности произведений различных модельных коллекций и корпусов;

8) установлена эффективность применения γ-классификатора для атрибуции искусственно сгенерированных поэм «Шахнаме» А. Фирдоуси по обучению рекуррентных нейронных сетей LSTM (Long short-term memory);

9) исследовано влияние порядка ЦП текста на распознавание однородности произведения с помощью γ-классификатора;

10) впервые в Таджикистане создан объектно-ориентированный компьютерный программный комплекс распознавания (идентификации) однородности текста на основе различных ЦП текста и γ-классификатора среди сколь угодно большого числа текстов.

Теоретическая значимость работы состоит в том, что в ней экспериментально опробован новый метод классификации дискретных случайных величин и установлена эффективность его применения для целей распознавания

авторства и для самых разных типов «однородностей» произведений художественной литературы для любых естественных языков на основе различных ЦП текста.

Практическая ценность работы состоит в том, что она нацелена на применение созданного в ней компьютерного программного комплекса *в государственной административной деятельности* для автоматизации процесса обработки текстовой информации, *в сфере криминалистики* для установления авторства анонимных текстов, *в области образования и науки* для обнаружения плагиата в курсовых и дипломных проектах, а также в представленных к защите кандидатских и докторских диссертациях.

Комплекс программ под названием **«THR»** (text homogeneity recognition) применён в следующих организациях:

1. Академия Министерства внутренних дел Республики Таджикистан.
2. Государственный комитет национальной безопасности Республики Таджикистан.
3. Институт языка и литературы имени Рудаки НАНТ.
4. Институт математики имени А.Джураева НАНТ.
5. ТТУ имени академика М.С. Осими (см. Приложение 1).

Построенный с широким использованием математических моделей и высокого уровня программирования комплекс, в частности, предназначен для развития таджикского языка с использованием возможностей информационных технологий.

Данный комплекс программы является важным как с точки зрения компьютерной лингвистики, так и с точки зрения литературоведения, и направлен на оказание практической помощи исследователям в области языка, литературы, математики и информационных технологий. Среди них призвано определить и распознать стиль каждого автора, особенности отдельных произведений разных авторов, частоту встречаемости букв, слогов, слов, словосочетаний, состав слов в отдельных произведениях, создание различных математических моделей.

Положения, выносимые на защиту: экспериментальное доказательство эффективности применения γ-классификатора с помощью различных ЦП текста для распознавания однородности текстовой информации.

Достоверность и обоснованность полученных результатов подтверждены сериями вычислительных экспериментов, в которых посредством γ-классификатора и метода ближайшего соседа распознаются с достаточно высокой точностью самых разных типов «однородностей» произведения различных модельных коллекций и корпусов.

Соответствие диссертации паспорту научной специальности. Содержание исследования данной диссертации соответствует пунктам 1, 3, 4, 5 и 7 по

специальности 05.13.11 - «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»:

- модели, методы и алгоритмы проектирования и анализа программ и программных систем, их эквивалентных преобразований, верификации и тестирования;

- модели, методы, алгоритмы, языки и программные инструменты для организации взаимодействия программ и программных систем;

- системы управления базами данных и знаний;

- программные системы символьных вычислений;

- человеко-машинные интерфейсы; модели, методы, алгоритмы и программные средства машинной графики, визуализации, обработки изображений, систем виртуальной реальности, мультимедийного общения.

Личный вклад соискателя учёной степени. Диссертационная работа является результатом более 10-летних исследований автора, проведенных в Таджикском техническом университете имени академика М.С. Осими, и на научно-исследовательских базах Института математики имени А. Джураева НАНТ. Постановка задачи осуществлялась совместно с научным консультантом. Основные результаты диссертационной работы получены автором самостоятельно.

Апробация и реализация результатов диссертации. Основные материалы и результаты диссертации получили положительные отзывы и обсуждены на:

– научно-исследовательских семинарах Института математики имени А. Джураева НАНТ, Политехнического института Таджикского технического университета имени академика М.С. Осими в городе Худжанд и Российско-Таджикского (Славянского) университета 2011-2024 гг.;

– международной научно-практической конференции «Подготовка конкурентоспособных специалистов рынка труда в условиях интеграции высших учебных заведений зарубежных стран и РТ», 2013 г., Душанбе;

– международной конференции «Памир: актуальные проблемы и научно-техническое развитие», 2013 г., Хорог;

– I международном круглом столе «Проблемы духовных и социальных ценностей современной молодежи России и Центральной Азии и пути их решения», 2013 г., Абакан;

– научно-практических семинарах «Новые информационные технологии в автоматизированных системах», 2014 г., 2016 г., 2018 г., 2019 г., Москва;

– международной научно-практической конференции «Перспективы развития науки и образования», 2016 г., Душанбе;

– международной конференции «Kamal Khujandi: Development of literary study and literary relations», 28-29 октября 2016 г., Худжанд;

– международной научно-практической конференции «Роль ИКТ в инновационном развитии экономики Республики Таджикистан», 2017г., Душанбе;

– международной научной конференции «Современные проблемы математики и их приложения», 14-15 июня 2017 г., Душанбе, Куляб;

– всероссийской научно-практической конференции «Состояние и перспективы развития ИТ-образования», 2019 г., Чувашская Республика;

– ежегодной межвузовской научно-технической конференции студентов, аспирантов и молодых специалистов имени Е.В. Арменского, МИЭМ НИУ ВШЭ, Секция №1 «Математика и компьютерное моделирование», 2020 г., Москва;

– proceedings of the 8th International Scientific and Practical Conference «Science and practice: implementation to modern society», 26-28.12.2020, Manchester, Great Britain;

– XVI международной конференции по компьютерной и когнитивной лингвистике TEL-2020, 12-13 ноября 2020 г., Казань, Россия;

– республиканской научно-теоретической конференции «Цифровая экономика и необходимость внедрения новой системы национальных счетов», 17 февраля 2021 г., Душанбе;

– XI международной научно-технической конференции «Открытые семантические технологии проектирования интеллектуальных систем», Open Semantic Technologies for Intelligent Systems (OSTIS-2021), 16-18 сентября 2021 г., Минск, Республика Беларусь;

– международной научно-практической конференции «Технические науки и инженерное образование для устойчивого развития», 12-13 ноября 2021 г., Таджикский технический университет имени академика М.С. Осими, Душанбе;

– международной конференции, посвящённой памяти профессора А.А. Тарасова и О.В. Казарина, по теме «Взаимодействие вузов, научных организаций и учреждений культуры в сфере защиты информации и технологий безопасности», 19 и 20 апреля 2022 г., Москва;

– международной научно-практической конференции «XII Ломоносовские чтения», посвященной 30-летию установления дипломатических отношений между Республикой Таджикистан и Российской Федерацией, Филиал МГУ имени М.В. Ломоносова в г. Душанбе, 29-30 апреля 2022 г., Душанбе;

– VI международной научно-практической конференции «Global and regional aspects of sustainable development», 26-28 февраля 2022 г., Копенгаген, Дания;

– XXII международной конференции «Информатика: проблемы, методы, технологии» (IPMT-2022), Воронежский государственный университет, 10-12 февраля 2022 г., Воронеж;

– международной научно-практической конференции “Цифровизация и искусственный интеллект”, посвященной «Двадцатилетию изучения и развития естественных, точных и математических наук в сфере науки и образования (2020-

2040 годы)», Таджикский технический университет имени академика М.С. Осими, 2023, Душанбе;

– международной конференции “Современные проблемы математики”, посвящённой 50-летию Института математики им. А.Джураева Национальной академии наук Таджикистана, 26-27 мая 2023 г., Душанбе;

– лучший педагог – 2023: IV международная книжная коллекция научно-педагогических работников, 2023, Астана;

– международной научно-практической конференции «Новые достижения в области естественных наук и информационных технологий», посвящённой «Двадцатилетию изучения и развития естественных, точных и математических наук на 2020-2040 гг.», 2023, Душанбе, РТСУ.

Публикации по теме диссертации. По теме диссертации опубликовано 73 работы, из них 34 (14 без соавторов) статьи в журналах из перечня, рекомендованных ВАК при Президенте Республики Таджикистан, 30 докладов в сборниках трудов и международных конференций, две монографии и два учебных пособия, а также пять баз данных и программ для ЭВМ, зарегистрированных в качестве объектов интеллектуальной собственности, [1-А-73-А].

Структура диссертации и объём. Диссертация состоит из введения, шести глав, заключения и приложений. Библиографический список включает 397 наименований. Основная часть диссертации изложена на 271 странице. Диссертация содержит 9 рисунков и 107 таблиц.

Автор выражает свою особую признательность и благодарность научным консультантам – доктору физ.-мат. наук, профессору, академику НАНТ, глубоко-уважаемому Усманову З.Д. и доктору физ.-мат. наук, профессору, академику НАНТ Рахмонову З.Х., а также сотрудникам Политехнического института Таджикского технического университета имени академика М.С. Осими в городе Худжанд, Института математики имени А. Джураева НАНТ и Таджикского технического университета имени академика М.С. Осими.

ГЛАВА 1. ОСНОВНЫЕ ПОНЯТИЯ И ОПРЕДЕЛЕНИЯ

Как сказано во введении, объектом диссертационного исследования является распознавание однородности произведений художественной литературы. Тот факт, что материалом для обработки служат тексты, не имеет особого значения. Принципиальным моментом является использование в качестве количественного описания текста его, так называемого, цифрового портрета, основанного на учёте распределения частотности различных алфавитных элементов текста, а также применение в качестве математической модели принятия решений, так называемого, γ-классификатора, предложенного З.Д. Усмановым. И то, и другое в равной мере применимы к решению аналогичных проблем для любых естественных языков с буквенным алфавитом.

В настоящей главе приводится обзор литературы (статей и публикаций), формулируется постановка задачи по автоматическому распознаванию однородности текста, вводятся понятия, широко используемые в дальнейшем, приводится подробное описание алгоритма γ-классификатора и дается краткое описание тех задач, которые будут исследованы в других главах. Сразу же отметим, что работа γ-классификатора демонстрируется сначала на модельных коллекциях. Вначале маленькие размеры используются для проведения предварительных исследований и лишь после того, как на модельной коллекции удаётся получить обнадеживающие результаты, использованные методы обработки исследуются на корпусах текстов, см. главу 4.

§ 1.1. Обзор исследования

Применение методов математического моделирования к идентификации однородности текстов опирается в своей основе на модель текста, то есть количественное описание объекта исследования. В настоящее время по подсчетам J.Rudman используется около 1000 групп характеристик [3] в качестве текстовых моделей, среди которых – морфологические, лексические, идиосинкразические, синтаксические, структурные, контентно-специфические и другие характеристики. В дополнение к сказанному уместно отметить, что в монографии А.А. Шелупанова, А.С. Романова и Р.В. Мещерякова [227] представлен обширный обзор работ по распознаванию однородности текста на основе разнообразных ЦП текстов и применяемых методов классификации.

В Таджикистане исследования, непосредственно относящиеся к автоматическому распознаванию новизны информации, компиляции, плагиата, заимствования, к идентификации авторства и т.п., берут своё начало нашим предыдущим исследованием совместно с научным консультантом З.Д. Усмановым. С помощью обобщенной формулы золотого сечения для поэмы А. Фирдоуси «Шахнаме» нами предложены 3 параметра, один из которых

характеризует само произведение, а два других – творчество самого автора [278, 283]. В другой работе мы применили в качестве цифрового кода пять натуральных единиц измерения текста для распознавания произведений А. Фирдоуси [283]. В дальнейшем мы доказали пригодность в качестве ЦП таджикского текста использования распределения частотности N-грамм (N=1,2,3) для идентификации авторов произведений [283-285] и своими многоплановыми вычислениями подтвердили также высокую эффективность и конкурентоспособность γ-классификатора дискретных случайных величин З.Д. Усманова для решения различных задач распознавания новизны информации.

В последние годы работы по автоматической обработке информации на таджикском языке сосредотачиваются на тестировании разнообразных ЦП текстовой информации: распределении частотностей словоформных N-грамм (Ш.Н. Ашурова, [262-264]), длин предложений (К.С. Бахтеев, [314-316]) длин слов (А.А. Каримов, [256]), анаграмм (М.М. Каюмов, [317-324]), частотностей слогов (Х.А. Худойбердиев, [14-А, 54-А]) и др.

Выполненный в работе анализ современного состояния проблемы позволил сделать вывод, что в области обработки текстов не существует единой системы и методологии, определяющих возможность решения различных задач для текстов разного типа. Поэтому создание новых методов и алгоритмов распознавания однородности текстов представляется важным и актуальным для решения прикладных задач, таких как задача классификации документов по тематическим категориям, идентификация авторства, плагиата, язык, содержание текстов и т.д. В следующей подглаве формулируется постановка задачи, решаемой в настоящей диссертации.

§ 1.2. Постановка проблемы

Можно выделить семь основных описаний проблемы, возникающей при распознавании однородности объектов, показанных на рисунке 1.1.

1. Проблема или объект исследования – это процесс или явление, которое берется исследователем для изучения или как часть научного познания, которое исследователь постигает. Объектами изучения бывают текст, изображения, звук, формула, код программ и т.д. В настоящей работе объектом исследования является текст.

2. Признаки однородности – так как проблема связана с текстом, рассматриваются следующие признаки однородности: распознавание авторства, тематики текста, язык, группа языков, оригинал и его перевод, стиль произведений и шифры научных работ. Эта задача изучается в главах 2-4.

3. Элементы – примерами элементов текста могут служить буквы алфавита естественного языка, буквенные N-граммы и слоги, знаки пунктуации, морфемы, словоформы, длины слов, предложений и абзацев (в символах и словах), анаграмм

и др.

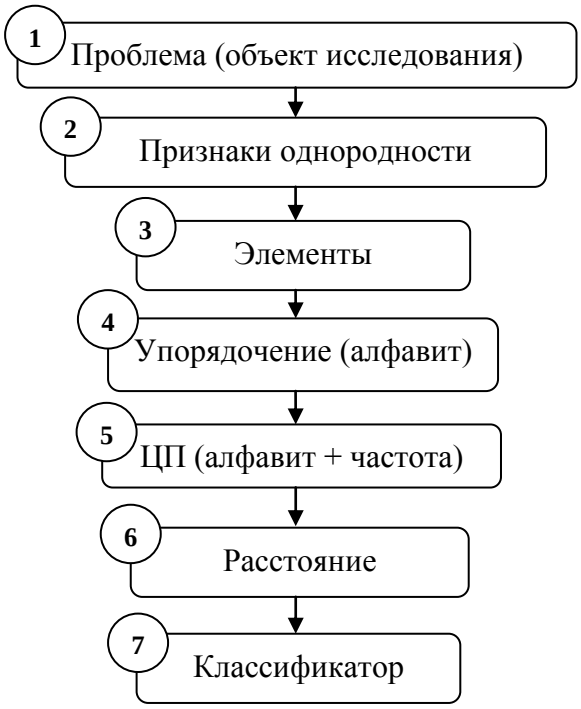


Рисунок 1.1. – Описание проблемы

4. Упорядочение (алфавит) – если элементы фиксированы (т.е. выбраны), то результат зависит от порядка расположения элементов (т.е. от выбора алфавита). Эту задачу изучаем в главе 5.

5. ЦП – это количественное описание объекта исследования, его математическая модель, назовём распределение частотности элементов алфавита. Примерами ЦП текста являются распределения частотностей символьных, буквенных и словоформных *N*-грамм, длин слов и предложений и т.д.

6. Расстояния между текстами – в широком смысле, степень (мера) удалённости или близости текстов друг от друга. Формул нахождения расстояния очень много, например: γ-классификатор, евклидово расстояние, коэффициент корреляции, Смирнов-Колмогоров, Фишер-Синдекор и т.д.

7. Классификатор – для распознавания однородных текстов помимо ЦП используются математические модели принятия решений, среди которых особо успешными являются нейронные сети, машина опорных векторов и недавно разработанный в Институте математики имени А. Джураева НАНТ γ-классификатор, [259, 260]. В следующих §§ 1.3 и 1.4 дается описание существа γ-классификатора З.Д. Усманова.

§ 1.3. Терминология и понятие

1.3.1. Задача распознавания авторства произведения.

Пусть $\mathbb{A} = \{A_i\}$ – список авторов A_i, $i = \overline{1,\alpha}$, и $\mathbb{T} = \{T_j\}$ – некоторое множество принадлежащих им текстов T_j, $j = \overline{1,\beta}$. Предположим, что $\mathbb{T}$

разделено на две части, $\mathbb{T} = \mathbb{T}_1 + \mathbb{T}_2$, из которых $\mathbb{T}_1$ предназначается для разработки правила соответствия (отображения) «текст → автор» (**задача 1** – обучение математической модели), а $\mathbb{T}_2$ – для проверки эффективности разработанного правила (**задача 2** – тестирование математической модели).

Существование взаимосвязи между текстом и его автором составляет основу современной стилеметрии. С позиции статистики авторский стиль – это вероятностное явление. По существу, любые элементы или же признаки, обнаруживаемые в текстах, появляются с какими-то частотами, которые не подконтрольны автору и тем не менее несут информацию, характеризующую своего создателя.

В задаче распознавания автора текста приходится иметь дело с парой математических моделей: количественным описанием (образом) текста и моделью принятия решения (классификацией). И тех и других моделей – необозримые множества. В настоящее время описаны разнообразные пары моделей, использованные для исследовательских целей. Обилие возможных комбинаций элементов пары является причиной, по которой исследователи в настоящее время не затрагивают вопросы построения общей теории, ограничиваясь подбором высоко эффективных пар для решений конкретных задач распознавания авторства.

Обсуждаемая задача является частным случаем общей проблемы построения систем распознавания образов, состоящей в разработке оптимальных решающих процедур для классификации образов и идентификации объектов, как единичных реализаций образов. Поэтому все достижения в развитии распознающих систем находят применение в решении задач идентификации авторства.

1.3.2. ЦП печатного текста.

Введем ряд определений, которыми будем пользоваться в дальнейшем.

Определение 1.3.1. ***Алфавит*** **–** *упорядоченное множество элементов текста,* см. § 1.2.

Примерами элементов текста являются буквы естественного языка, символы и знаки препинания, буквенные *N*-граммы и слоги, леммы и морфемы, корни и основы слов, словоформы, тематические ключевые слова и ключевые *N*-граммы, длины слов и предложений и многое другое. Совокупность элементов, *упорядоченных каким-либо образом*, образует алфавит.

Определение 1.3.2. ***ЦП текста*** *будем называть распределением частотности элементов алфавита.*

Следовательно, ЦПТ – это пара, составленная, с одной стороны, из упорядоченных элементов текста и, с другой стороны, из информации об относительной частоте встречаемости в тексте самих элементов. Таковыми примерами являются распределения частотностей упорядоченных символьных, буквенных и словоформных *N*-грамм, длин слов и предложений и т.д.

ЦП текста T записывается в табличном виде:

$$\begin{array}{lccccc} N: & 1 & 2 & . & . & . & m \\ P: & p_1 & p_2 & . & . & . & p_m, \end{array} \qquad (1.1)$$

в которой первая строка – порядковые номера (индексы) алфавитных элементов (m – число элементов), а вторая – их относительные частоты встречаемости в T, причём $\sum_{k=1}^{m} p_k = 1$.

ЦП представляется также в виде дискретной функции

$$F(s) = \sum_{k=1}^{s} p_k \qquad (s = 1, \cdots, m). \qquad (1.2)$$

1.3.3. Расстояния между ЦПТ.

Пусть T_1, T_2 – произвольная пара текстов, характеризуемых на основе единого алфавита, и

$$F^{(\alpha)}(s) = \sum_{k=1}^{s} p_k^{(\alpha)} - \qquad (1.3)$$

соответствующие им ЦП, представленные дискретными функциями, $\alpha = 1, 2,$ и $s = 1, \cdots, m$.

Определение 1.3.3. *Расстоянием между текстами* T_1 *и* T_2 *называется положительное число* $\rho(T_1, T_2)$*, определяемое формулой*

$$\rho(T_1, T_2) = \sqrt{m/2}\ max_s \left| F^{(1)}(s) - F^{(2)}(s) \right|, \qquad (1.4)$$

то есть расстояние между двумя текстами вычисляется как максимальное расстояние по оси ординат между их дискретными функциями $F^{(1)}(s)$ и $F^{(2)}(s)$, помноженное на весовой коэффициент $\sqrt{m/2}$. Отметим также, что равенство $\rho(T_1, T_2) = 0$ означает совпадение ЦП T_1 и T_2, но не самих текстов.

1.3.4. Гипотеза $\mathbb{H}$ «однородности» особенностей авторского стиля.

Обнаруживаемые в творчестве авторов «однородности» тех или иных особенностей стилей проявляются в их произведениях, словоупотреблениях, синтаксисе, композиции, интонациях, ритмах и многом другом. Не уточняя этого понятия, ограничимся тем, что сопоставим ему синонимы «похожий», «одинаковый», «сходный», «однотипный», «родственный» и т.п. Все они привязываются к понятию авторского стиля, который индивидуализирует творчество автора на фоне его коллег из писательского сообщества.

Гипотеза $\mathbb{H}$, связываемая с содержательным смыслом изучаемого вопроса, используется для решения задачи 1 путем подбора и последующей настройки математической модели. Наиболее естественной представляется следующая:

ГИПОТЕЗА $\mathbb{H}$. *Произведения одного автора – «однородные», а разных авторов – «неоднородные».*

Произведение – широкое понятие. Оно характеризуется набором признаков. Но тогда свойство «однородности» произведений можно интерпретировать как «однородность» отдельных признаков или же их совокупностей. Следовательно, обсуждаемая гипотеза может быть высказана в следующем видоизменённом виде:

ГИПОТЕЗА $\mathbb{H}^*$. *Конкретные признаки «однородны» во всех произведениях одного и того же автора и «неоднородны» в произведениях разных авторов.*

С такой точки зрения становится понятным, почему исследователи, занятые распознаванием авторства текста, имеют дело с его отдельными характеристиками, а не с текстами в целом. Так, например, распределения буквенных униграмм, биграмм, триграмм (с пробелом и без пробела), слогов, морфем, словоформных *N*-грамм, длин предложений и абзацев и многие другие признаки также успешно распознают авторов текстовых фрагментов.

В литературе можно указать много примеров нарушения этой гипотезы, однако она принимается к исполнению, как первое приближение к реальной ситуации, позволяющей преобразовать гипотезу в математическую модель.

§ 1.4. γ-классификатор, [271]

γ-классификатор – это математическая триада, состоящая из ЦП текста, формулы расстояний между текстами и алгоритма обучения по прецедентам.

1.4.1. Математическая модель $\mathbb{H}$-гипотезы.

Пусть γ – некоторое положительное число.

Определение 1.4.1. *Тексты* T_1, T_2 *называются* ***γ- однородными****, если*

$$\rho\,(T_1\,, T_2) \leq \gamma, \tag{1.5}$$

и ***γ- неоднородными****, если*

$$\rho\,(T_1\,, T_2) > \gamma. \tag{1.6}$$

Неравенства (1.5) и (1.6) являются математической интерпретацией (моделью) гипотезы $\mathbb{H}$.

Определение 1.4.2. *γ-классификатор – алгоритм, зависящий от одного вещественного параметра γ и сопоставляющий тексту из $\mathbb{T}_1$ его автора из списка $\mathbb{A}$.*

Очевидно, что от значения γ зависит однородность или неоднородность любой пары текстов, следовательно, и степень выполнимости гипотезы. Однородность всех текстов одного автора в рамках математической модели означает справедливость неравенства (1.5), а неоднородность любых двух текстов разных авторов – справедливость неравенства (1.6). Гипотеза $\mathbb{H}$ может нарушаться для каких-то пар текстов одного и того же автора в случае, когда вместо неравенства (1.5) имеет место неравенство (1.6), а также в случае, когда

какие-то два текста двух различных авторов удовлетворяют неравенство (1.5) вместо того, чтобы выполнялось неравенство (1.6).

Пусть $\tau = \tau(\gamma)$ – суммарное количество нарушений гипотезы $\mathbb{H}$ одновременно в двух случаях: невыполнение неравенства «однородности» в случае двух текстов, принадлежащих одному автору, и невыполнение неравенства «неоднородности» в случае двух текстов, принадлежащих разным авторам. Тогда для фиксированного γ *показатель выполнения гипотезы будет определяться величиной* π, задаваемой формулой

$$\pi = 1 - \tau(\gamma)/L\,, \qquad (1.7)$$

где L – число взаимных расстояний между всеми парами текстов из подколлекции $\mathbb{T}_1$. Из этой формулы следует, что π может принимать значения из отрезка [0, 1], причём $\pi = 0$, если $\tau = L$, и $\pi = 1$, если $\tau = 0$. В первом случае гипотезу $\mathbb{H}$ следует признать непригодной, а во втором – полностью согласованной с обучающей выборкой.

В связи с тем, что эффективность γ-классификатора зависит от значения параметра γ, представляет интерес найти такое его значение, при котором π принимает максимальное значение. *Именно в этом и заключается суть настройки γ-классификатора на данных обучающей выборки*. Если такая настройка будет приемлемой, то можно говорить о решении **задачи 1** – обучения γ-классификатора**.**

1.4.2. Множество текстов $\mathbb{T}_1$, предназначенное для настройки γ-классификатора, предполагается разделенным на n непересекающихся подмножеств $\mathbb{T}_1{}^{(k)}$, состоящих из $q^{(k)}$ текстов, принадлежащих одному и тому же автору $A^{(k)}$, $k = 1, \cdots, n$,.

Для настройки классификатора требуется знать:

- $Q = \sum_{k=1}^{n} q_k$ – суммарное количество текстов множества $\mathbb{T}_1$,
- $L = C_Q^2 = Q(Q-1)/2$ – общее число L пар текстов на $\mathbb{T}_1$,
- $L_1 = \frac{1}{2}\sum_{k=1}^{n} q^{(k)}(q^{(k)} - 1)$ – суммарное число всех пар авторских текстов (принадлежащих одним и тем же авторам),
- $L_2 = L - L_1$ – число пар между текстами различных авторов.

1.4.3. Алгоритм настройки γ-классификатора. Предположим, что обучающая выборка $\mathbb{T}_1$ со всеми текстами, привязанными к своим авторам, задана, и необходимые величины n, Q, L, L_1 и L_2 либо известны заранее, либо уже вычислены.

Алгоритм включает в себя следующие основные процедуры.

1. По ЦП (1.1) или (1.2) всех текстов обучающей выборки объёма Q с помощью формул (1.2) – (1.4) из § 1.3 подсчитать L расстояний между её текстами.

2. Полученный набор расстояний разделить на два множества $X = \{x_i\}$ и $Y = \{y_j\}$, в которых $x_i,\ i = 1, \cdots, h_1,$ и $y_j, j = 1, \cdots, h_2$, являются *упорядоченными по возрастанию расстояниями между парами текстов*, принадлежащих в первом случае подмножествам $\mathbb{T}_1{}^{(k)},\ k = 1, \cdots, n,$ а во втором случае – разным подмножествам.

3. Подсчитать $\lambda(x_i)$ и $\lambda(y_j)$ – частотности чисел x_i и y_j. Очевидно, что

$$L_1 = \sum_{i=1}^{h_1} \lambda(x_i), \qquad L_2 = \sum_{j=1}^{h_2} \lambda(y_j) \qquad \text{и} \qquad L = L_1 + L_2.$$

4. Сформировать множество $Z = X \cup Y = \{z_k\},\ k = 1, \cdots, h \quad (h \le h_1 + h_2)$, элементы которого z_k пронумерованы в порядке возрастания их значений, $z_1 \le z_2 \le \cdots \le z_h$. Очевидно, что в связи с принятыми обозначениями z_k есть либо число $x_{i_0} \in X$ с частотой $\lambda(x_{i_0})$, либо число $y_{j_0} \in Y$ с частотой $\lambda(y_{j_0})$, либо число $x_{i_{00}} = y_{j_{00}}$, из которых $x_{i_{00}}$ с частотой $\lambda(x_{i_{00}})$, а $y_{j_{00}}$ с частотой $\lambda(y_{j_{00}})$.

Числа $z_k,\ k = 1, \cdots, h,$ разделяют числовую полуось $z > 0$ на интервал $(0, z_1)$ и h полуинтервалов $[z_1, z_2)$, ... , $[z_{h-1}, z_h)$ и $[z_h, \infty)$. Функция $\tau(\gamma)$, определенная на вещественной полуоси $(0\ , \infty)$, принимает целочисленные постоянные значения в интервале $(0, z_1)$ и на h полуинтервалах $[z_1, z_2)$, ... , $[z_{h-1}, z_h)$ и $[z_h, \infty)$.

Эти значения таковы:

- $\tau(\gamma) \equiv \tau_1 = L_1$ при $\gamma \in (0, z_1)$;
- $\tau(\gamma) \equiv \tau_2$ при $\gamma \in [z_1, z_2)$, причём $\tau_2 = \tau_1 + \Delta_1$,

где $$\Delta_1 = \begin{cases} -\lambda(x_1) & \text{при } z_1 = x_1, \\ \lambda(y_1) & \text{при } z_1 = y_1, \\ \lambda(y_1) - \lambda(x_1) & \text{при } z_1 = x_1 = y_1; \end{cases}$$

- $\tau(\gamma) \equiv \tau_k$ при $\gamma \in [z_{k-1}, z_k)$, причём $\tau_k = \tau_{k-1} + \Delta_{k-1}$,

где $$\Delta_{k-1} = \begin{cases} -\lambda(x_{i_0}) & \text{при } z_{k-1} = x_{i_0}, \\ \lambda(y_{j_0}) & \text{при } z_{k-1} = y_{j_0}, \\ \lambda(y_{j_{00}}) - \lambda(x_{i_{00}}) & \text{при } z_{k-1} = x_{i_{00}} = y_{j_{00}}. \end{cases}$$

и $k = 3, \cdots, h$;

- и, наконец, $\tau(\gamma) \equiv \tau_{h+1} = L_2$ на полуинтервале $[z_h, \infty)$.

5. Вычислить значения $\tau_1, \tau_2, \cdots, \tau_h, \tau_{h+1}$ по формулам предыдущего пункта и выделить минимальное из них.

Пусть это будет τ_{k^*}, где $k^* = \arg\min_k \tau_k$ и $k = 1, \cdots, h+1$. В таком случае эффективность кластеризатора будет характеризоваться величиной

$$\pi(\tau_{k^*}) = 1 - \tau_{k^*} / L,$$

а область оптимального значения γ определяться из условия

$$\gamma^{\text{опт}} \in \begin{cases} (0, z_1), & \text{если } k^* = 1, \\ [z_{k-1}, z_k), & \text{если } k^* = k = 2, \cdots, h, \\ [z_h, \infty), & \text{если } k^* = h+1. \end{cases}$$

Вопрос о приемлемости полученного решения зависит от величины $\pi(\tau_{k^*})$. Если эта величина оказывается в определенном смысле близкой к единице, то можно признать, что предложенная математическая модель удачно настроена на данных обучающей выборки, и тем самым откалиброванный кластеризатор можно использовать в качестве классификатора, подготовленного к выполнению своих функций (в частности для тестирования математической модели, **задачи 2**).

1.4.4. Пояснения к описанию алгоритма. Как сказано ранее, обучение математической модели распознаванию авторства текста эквивалентно настройке γ-классификатора на данных обучающей выборки. Настройка производится за счет выбора оптимального значения γ, обеспечивающего достижение максимально возможного уровня выполнения гипотезы $\mathbb{H}$. Соответствующим показателем этого уровня является величина π, вычисляемая по формуле (1.7). Эта величина, в свою очередь, связана со значением $\tau = \tau(\gamma)$ – суммарным числом случаев нарушений гипотезы $\mathbb{H}$, которое складывается из нарушений *условий однородности* пары текстов, принадлежащих одному автору, и нарушений *условий неоднородности* пары текстов, принадлежащих двум разным авторам.

Так как показатель $\tau = \tau(\gamma)$ зависит от γ, то было бы желательно иметь явный вид искомой зависимости. Однако такой зависимости нет, и приводимый в п. 1.4.3 алгоритм является, по существу, набором процедур последовательного вычисления значений функции $\tau = \tau(\gamma)$.

Предварительный анализ свойств этой функции подсказывает, что она определена для значений γ на полуоси (0, ∞) и является кусочно-гладкой с разрывами в точках $z_1 \le z_2 \le \cdots \le z_h$, см. п. 4 алгоритма. Указанные значения определяются по данным обучающей выборки, более точно, совокупностью L расстояний между текстами множества $\mathbb{T}_1$.

В п. 2 совокупность L расстояний разделяется на две части. В одной части с числом элементов L_1 собираются все расстояния между собственными текстами самих авторов, которые должны быть *однородными* в согласии с гипотезой $\mathbb{H}$. В другой части с числом элементов L_2 – все расстояния между текстами различных авторов, которые должны быть *неоднородными* в согласии с той же гипотезой $\mathbb{H}$. Полученные наборы расстояний обозначаются через $X = \{x_i\}$ и $Y = \{y_j\}$, в которых x_i, $i = 1, \cdots, h_1$, и y_j, $j = 1, \cdots, h_2$, являются *упорядоченными по возрастанию расстояниями между парами текстов*, принадлежащих в первом случае подмножествам ${\mathbb{T}_1}^{(k)}$, $k = 1, \cdots, n$, а во втором случае – разным подмножествам.

В п. 4 формируется множество $Z = X \cup Y = \{z_k\}$, $k = 1, \cdots, h$ $(h \le h_1 + h_2)$, элементы которого z_k нумеруются в порядке возрастания их значений. Очевидно, что значение z_k есть либо число $x_{i_0} \in X$ с частотой $\lambda(x_{i_0})$, либо число $y_{j_0} \in Y$ с частотой $\lambda(y_{j_0})$, либо число $x_{i_{00}} = y_{j_{00}}$, из которых $x_{i_{00}}$ с частотой

$\lambda(x_{i_{00}})$, а $y_{j_{00}}$ с частотой $\lambda(y_{j_{00}})$.

Числа z_k, $k = 1, \cdots, h$, разделяют числовую полуось $z > 0$ на интервал $(0, z_1)$ и h полуинтервалов $[z_1, z_2)$, ... , $[z_{h-1}, z_h)$ и $[z_h, \infty)$. Функция $\tau(\gamma)$ принимает целочисленные постоянные значения в интервале $(0, z_1)$ и на h полуинтервалах $[z_1, z_2)$, ... , $[z_{h-1}, z_h)$ и $[z_h, \infty)$. Скачки значений функции $\tau(\gamma)$ происходят в точках $z_1, z_2, \cdots, z_h$ и, как устанавливается в п. 4, имеем

$$\tau(\gamma) = \begin{cases} L_1 \text{ при } \gamma \in (0, z_1), \\ \tau(z_1) \text{ при } \gamma \in [z_1, z_2), \\ \tau(z_{k-1}) \text{ при } \gamma \in [z_{k-1}, z_k) \text{ и } k = 3, \cdots, h, \\ L_2 \text{ при } \gamma \in [z_h, \infty). \end{cases}$$

Остается определить минимальное значение $\tau(\gamma)$ и далее поступать так, как указано в п. 5 алгоритма.

1.4.5. Замечание. Обратим внимание на то, что гипотезы $\mathbb{H}$ и $\mathbb{H}^*$, настроенные на идентификацию авторства и особенности авторского стиля, могут быть переориентированы также и на другие цели.

К примеру, если различать произведения по различным тематикам, то $\mathbb{H}^{**}$ - гипотезу для настройки γ-классификатора естественно формулировать в следующем виде*: любые произведения по одной тематике «однородны», а по разным – «неоднородны»*. И опять-таки неравенства (1.5) и (1.6) можно рассматривать в качестве математической интерпретации (модели) $\mathbb{H}^{**}$-гипотезы.

Другой пример – распознавание языков произведений. В этом случае $\mathbb{H}^{**}$ - гипотеза формулируется в слегка видоизмененном виде*: любые произведения, написанные на одном языке, «однородны», а на разных – «неоднородны»*. И опять неравенства (1.5) и (1.6) выступают в качестве математической интерпретации $\mathbb{H}^{**}$-гипотезы.

Важно отметить, что плодотворность гипотез зависит не только от γ-классификатора, но также и от тщательно подобранного ЦП объекта исследования.

В следующих четырех главах настоящей диссертации изучаются вопросы распознавания однородности текста на основе различных ЦПТ и γ-классификатора среди сколь угодно большого числа текстов.

ГЛАВА 2. ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ РАСПОЗНАВАНИЯ ОДНОРОДНОСТИ ТЕКСТОВ НА ПРИМЕРАХ МОДЕЛЬНЫХ КОЛЛЕКЦИЙ ХУДОЖЕСТВЕННЫХ ПРОИЗВЕДЕНИЙ

Для распознавания авторского стиля или же идентификации однородности текстов важнейшая роль отводится выбору ЦПТ. Как отмечено в [3] и § 1.1 главы 1, подсчёты Рудмана показали, что для указанных целей различными исследователями использованы порядка тысячи количественных признаков, которые можно разделить на *лексические* (уровень символов и слов), *морфологические* (грамматические классы, N-граммы грамматических классов, леммы, морфемы), *синтаксические* (функциональные слова, словоформные N-граммы, пунктуация, синонимия, признаки предложения), *идиосинкразические* (орфографические и грамматические ошибки, текстовые аномалии), *контентно-специфические* (ключевые слова по тематике, ключевые N-граммы, эмотиконы, сокращения и акронимы, слова на другом языке), *структурные* (структура текста, форматирование и оформление) и, наконец, *метаданные* (структура данных и стеганография).

В настоящей главе тестируются количественные признаки высокого уровня на предмет возможности их использования в качестве информативных признаков для распознавания автора на примере модельных коллекций художественных произведений таджикского языка, а также узбекского языка, и в роли исследовательского аппарата применялись γ-классификатор З.Д. Усманова и метод ближайшего соседа. Наша цель будет состоять не только в том, чтобы выявить различия в размерах и расположениях оптимальных полуинтервалов γ, но также и в определении числа нарушений гипотезы однородности, вычислении коэффициента эффективности распознавания авторов по их произведениям в целом и возможно минимальным фрагментам. Фрагменты могут извлекаться из любого произведения, из его любой части текста. В качестве таковых частей выступают «начало», «середина» и «конец» произведения, «в пределах» которых кусочки текста различных размеров выбирались бессистемно, случайным образом.

§ 2.1. Применение буквенных N-граммных единиц измерения текста

§ 2.1.1. Об определение автора текста на основе частотности символьных униграмм

Решается задача распознавания авторов произведений по отдельности для классической и современной поэзий, а также современной прозы. Произведениям сопоставляется ЦП, характеризуемый распределением в них частотности буквенных униграмм. Устанавливается эффективность применения γ-

классификатора для идентификации авторов произведений.

В настоящем параграфе мы продолжаем тестирование количественных описаний текстов, начатое в работах [1-А-10-А], на предмет их пригодности для идентификации авторов произведений. В качестве таковых в [255, 6-А] рассматривались частотности букв таджикского алфавита (униграммы), в [7-А, 8-А] – буквенных биграмм и триграмм, в [283] – набора из пяти натуральных единиц измерения текста, в [256, 257] – частотности длин слов и знаков препинаний, в [14-А] – частотности слогов, в [258, 13-А, 17-А] – частотности длин предложений. Существенным моментом в сравнении с нашим предыдущим исследованием [6-А] является изучение вопроса о распознавании авторов текстов, относящихся к произведениям классической и современной поэзии, а также к современной прозе. Следуя [6-А], будем называть *цифровым портретом текста* распределение в нём частотности букв таджикского алфавита (униграммы). В параграфе изучается вопрос об эффективности применения такого показателя для распознавания авторов поэтических и прозаических произведений.

1. Состав модельной коллекции текстов представлен следующими произведениями.

Классическая поэзия

– А. Рӯдакӣ: «Абёти пароканда» (АР, АП, 22,2 *Кб*), «Қасоид» (АР, Қ, 49,9 *Кб*);

– А. Фирдавсӣ: «Достони Рустам ва Сӯҳроб» (АФ, Р&С, 164 *Кб*), «Достони Бежан бо Манижа» (АФ, Б&М, 149 *Кб*);

– С. Шерозӣ: «Ғазалиёт қисми 1» (СШ, Ғ1, 165 *Кб*), «Ғазалиёт қисми 2» (СШ, Ғ2, 130 *Кб*);

– Ҳ. Шерозӣ: «Ғазалиёт қисми1» (ҲШ, Ғ1, 340 *Кб*), «Ғазалиёт қисми 2» (ҲШ, Ғ2, 295 *Кб*);

– Ҷ. Румӣ: «Маснавии Маънавӣ Дафтари Аввал» (ҶР, ММ1, 486 *Кб*), «Маснавии Маънавӣ Дафтари Дуввум» (ҶР, ММ2, 414 *Кб*).

Современная поэзия

– А. Суруш: «Дафтари 1» (АС, Д1, 107 *Кб*), «Дафтари 2» (АС, Д2, 130 *Кб*);

– А. Шукӯҳӣ: «Баргҳои тиллоӣ» (АШ, БТ, 327 *Кб*), «Шохаи райҳон» (АШ, ШР, 131 *Кб*);

– Г. Сафиева: «Офтоб дар соя» (ГС, О, 138 *Кб*), «Шӯъла дар санг» (ГС, Ш, 569 *Кб*);

– И. Фарзона: «101-Ғазал» (ИФ, 101Ғ, 105 *Кб*), «Мӯҳри гули мино» (ИФ, МГМ, 496 *Кб*);

– М. Турсунзода: «Қиссаи Ҳиндустон» (МТ, ҚҲ, 64,9 *Кб*), «Ҳасани аробакаш» (МТ, ҲА, 92,2 *Кб*).

Современная проза

– А. Зоҳир: «Бозгашт» (АЗ, Б, 784 *Кб*), «Завол» (АЗ, З, 877 *Кб*);

– Г. Муҳаммадиева: «Бӯи модар» (ГМ, БМ, 525 *Кб*), «Сафинаи муҳаббат» (ГМ, СМ, 561 *Кб*);

– М. Шакурӣ: «Садри Бухоро» (МШ, СБ, 1308 *Кб*), «Хуросон аст ин ҷо» (МШ, Х, 1057 *Кб*);

– С. Турсун: «Нисфирӯзӣ» (СТ , Н, 108 *Кб*), «Повести Камони Рустам» (СТ , ПКР, 43,7 *Кб*);

– С. Айнӣ: «Дохунда» (СА, Д, 751 *Кб*), «Марги судхӯр» (СА, МС, 523 *Кб*).

Для авторов и их произведений приняты обозначения, указываемые в скобках: первые две буквы – это инициалы авторов, вторые – сокращенные шифры текстов, третьи – информация о объёмах произведений в килобайтах.

2. Обработка статистического материала включала в себя 4 этапа.

Этап 1. Создание компьютерный программы и вычисления с её помощью ЦП произведений – распределений частотности униграмм по отдельности для всех текстов, упомянутых в п. 1.

Этап 2. Создание компьютерный программы и вычисления с её помощью парных расстояний между ЦПП по формуле, предложенной в § 1.3.

Этап 3. Настройка γ-классификатора. Существо настройки заключается в том, чтобы определить такое значение вещественного параметра γ, при котором достигается максимальное значение критерия «γ-однородности» произведений, см. § 1.4.

Этап 4. Установление эффективности применения настроенного γ-классификатора для распознавания авторов произведений.

На этапе 1 цифровые портреты произведений представляются в табличном виде:

$$\begin{array}{lccccc} \bar{N}: & 1 & 2 & . & . & . \quad m \\ P: & p_1 & p_2 & . & . & . \quad p_m, \end{array}$$

где первая строка – список униграмм; m – общее число униграмм; вторая строка – частоты p_i встречаемости в пределах произведений буквенных униграмм $i(i = 1,2,\dots,m)$, причём

$$\sum_{i=1}^{m} p_k = 1.$$

На этапе 2 вычисления расстояний $\rho(T_1, T_2)$ между текстами T_1 и T_2 производились по формуле T_1 и T_2

$$\rho(T_1, T_2) = \sqrt{\frac{m}{2}} \max_s \left| \sum_{k=1}^{s} (p_k^{(1)} - p_k^{(2)}) \right|,$$

в которой m (= 35) – количество униграмм; $p_k^{(1)}$ и $p_k^{(2)}$ – частоты встречаемости в текстах T_1 и T_2 суммарные количества буквенных униграмм k, k = 1, … , m, и $(\mathrm{s} = 1, \cdots, \mathrm{m})$.

Результаты вычислений показаны в таблицах 2.1-2.3.

На этапе 3 качество классификатора при фиксированном γ оценивается величиной π, вычисляемой по формуле

$$\pi = 1 - \tau/L\,, \tag{2.1}$$

где L (= 45) – суммарное число взаимных расстояний между 10 текстами исходной коллекции; $\tau = \tau(\gamma)$ – число нарушений неравенств

$$\rho(T_1, T_2) \le \gamma, \tag{2.2}$$

$$\rho(T_1, T_2) > \gamma. \tag{2.3}$$

Первое проверяется на 5 парах текстов одних и тех же авторов, второе – на 40 парах текстов различных авторов.

На этапе 4 производится настройка γ-классификатора на основе вполне естественной гипотезы о том, что произведения одного автора «однородны», а разных авторов «неоднородны». На языке ЦП, характеризующих распределения частотности длин 10 пар произведений, определение γ сводится к отысканию такого его значения, при котором общее число τ нарушений неравенств (2.2), (2.3) по отдельности на текстах 3-х модельных коллекций становится минимальным. Для нахождения таких γ используется алгоритм, предложенный в § 1.4.

3. Результаты вычислений расстояний между 10 произведениями классической поэзии представлены в табл. 2.1.

Таблица 2.1. – Расстояния между произведениями ***классической поэзии***

Автор (Произ.)		**АР**		**АФ**		**СШ**		**ХШ**		**ЧР**	
		АП	**К**	**Р&С**	**Б&М**	**Ғ1**	**Ғ2**	**Ғ1**	**Ғ2**	**ММ1**	**ММ2**
		2248	**5054**	**16355**	**14799**	**16261**	**13001**	**33724**	**28923**	**48713**	**41661**
АР	**АП**										
	К	**0.0419**									
АФ	**Р&С**	0.0503	0.0715								
	Б&М	0.0620	0.0670	**0.0278**							
СШ	**Ғ1**	0.1482	0.1402	0.1812	0.1896						
	Ғ2	0.1155	0.1157	0.1485	0.1569	**0.0653**					
ХШ	**Ғ1**	0.1193	0.1114	0.1523	0.1607	0.0920	0.0636				
	Ғ2	0.1414	0.1335	0.1744	0.1828	0.1117	0.0938	**0.0301**			
ЧР	**ММ1**	0.1003	0.0964	0.1307	0.1371	0.1843	0.1516	0.1554	0.1775		
	ММ2	0.1011	0.0804	0.1238	0.1301	0.1765	0.1437	0.1419	0.1640	**0.0208**	

Для классической поэзии оптимальное значение γ оказалось следующим

$$\gamma^{opt} \in [0.0420;\ 0.0502),$$

Применять этот факт для выяснения метрической близости пары произведений T_1 и T_2 необходимо следующим образом, см. [16-А, 44-А]:

– если $\rho(T_1,\ T_2) \leq 0.0420$, то T_1 и T_2 однородны;

– если $\rho(T_1,\ T_2) > 0.0502$, то T_1 и T_2 неоднородны;

– если $0.0420 < \gamma \leq 0.0502$, то ситуация – неопределенная.

Из данных таблицы следует, что только одно расстояние, именно 0.0653, соответственно между ЦП двух произведений С. Шерозӣ «Ғазалиёт қисми 1» и «Ғазалиёт қисми 2» нарушает сформулированную гипотезу. Эти пары согласно (2.3) утверждают неоднородность указанных двух произведений С. Шерозӣ, хотя принадлежат одним авторам.

Желтым цветом в таблице 2.1 отмечен 1 случай нарушения гипотезы однородности.

4. Результаты вычислений расстояний между 10 произведениями современной поэзии представлены в табл. 2.2.

Таблица 2.2. – Расстояния между произведениями в ***современной поэзии***

Автор (Произ.)		**АС**		**АШ**		**ГС**		**ИФ**		**МТ**	
		Д1	**Д2**	**БТ**	**ШР**	**О**	**Ш**	**101Г**	**МГМ**	**ҚҲ**	**ҲА**
		7890	**9322**	**32036**	**12810**	**12103**	**51434**	**9841**	**41217**	**8463**	**6118**
АС	**Д1**										
	Д2	**0.0301**									
АШ	**БТ**	0.1060	0.0783								
	ШР	0.1003	0.0726	**0.0172**							
ГС	**О**	0.0798	0.0585	0.0571	0.0565						
	Ш	0.0743	0.0466	0.0675	0.0661	**0.0255**					
ИФ	**101Г**	0.1495	0.1219	0.1019	0.1006	0.0697	0.0752				
	МГМ	0.1422	0.1145	0.0960	0.0958	0.0624	0.0679	**0.0208**			
МТ	**ҚҲ**	0.1193	0.0916	0.0839	0.0851	0.0562	0.0470	0.0478	0.0439		
	ҲА	0.1018	0.0821	0.0375	0.0402	0.0723	0.0827	0.1171	0.1049	**0.0871**	

Для современной поэзии оптимальное значение γ оказалось следующим

$$\gamma^{opt} \in [0.0302;\ 0.0374).$$

Применять этот факт для выяснения метрической близости пары произведений T_1 и T_2 необходимо следующим образом, см. [16-А, 44-А]:

– если $\rho(T_1,\ T_2) \leq 0.0302$, то T_1 и T_2 однородны;

– если $\rho(T_1,\ T_2) > 0.0374$, то T_1 и T_2 неоднородны;

– если $0.0302 < \gamma \leq 0.0374$, то ситуация – неопределенная.

И здесь в табл. 2.2 закрашенные жёлтым цветом ячейки (в данном случае их – 1) показывают нарушение сформулированной гипотезы для соответствующих пар произведений.

5. Результаты вычислений расстояний между 10 произведениями

современной прозы представлены в табл. 2.3.

Таблица 2.3. – Расстояния между произведениями в ***современной прозе***

Автор (Произ.)		**АЗ**		**ГМ**		**МШ**		**СТ**		**СА**	
		Б	**З**	**БМ**	**СМ**	**СБ**	**Х**	**Н**	**ПКР**	**Д**	**МС**
		70804	**79431**	**46608**	**50368**	**113592**	**91202**	**9936**	**4041**	**71134**	**48801**
АЗ	**Б**										
	З	**0.0464**									
ГМ	**БМ**	0.0313	0.0383								
	СМ	0.0351	0.0430	**0.0186**							
МШ	**СБ**	0.1322	0.1073	0.1117	0.1289						
	Х	0.1557	0.1307	0.1351	0.1524	**0.0394**					
СТ	**Н**	0.0733	0.0894	0.0724	0.0564	0.1818	0.2053				
	ПКР	0.0903	0.1101	0.0797	0.0721	0.1856	0.2091	**0.0535**			
СА	**Д**	0.0543	0.0725	0.0563	0.0600	0.1531	0.1766	0.0748	0.0547		
	МС	0.0565	0.0798	0.0689	0.0697	0.1448	0.1682	0.1000	0.0673	**0.0315**	

Для современной прозы оптимальное значение γ оказалось следующим

$$\gamma^{\text{opt}} \in [0.0316;\ 0.0350).$$

Применять этот факт для выяснения метрической близости пары произведений T_1 и T_2 современной прозы необходимо следующим образом, см. [16-А, 44-А]:

– если $\rho(T_1,\ T_2) \leq 0.0316$, то T_1 и T_2 однородны;

– если $\rho(T_1,\ T_2) > 0.0350$, то T_1 и T_2 неоднородны;

– если $0.0316 < \gamma \leq 0.0350$, то ситуация - неопределенная.

И здесь закрашенные в табл. 2.3 жёлтым цветом ячейки (в данном случае их – 4) показывают нарушение сформулированной гипотезы для соответствующих пар произведений.

6. Вычисления по формуле (2.1) коэффициента эффективности π:

– для классической поэзии выдаёт значение π = 98%,

– для современной поэзии выдаёт значение π = 98%,

– для современной прозы выдаёт значение π = 91%.

распознавания автора по цифровому портрету его произведений.

Полученные значения показывают, что распознавание автора текста по цифровому портрету (распределению частотности буквенных униграмм) для поэтических произведений (в сравнении с прозаическими) более успешно.

Результаты данного параграфа опубликованы в [58-А].

§ 2.1.2. Об идентификации автора текстового фрагмента на основе частотности символьных униграмм

На примере модельной коллекции таджикских литературных произведений изучается задача о возможности определения авторства фрагмента текста

минимального размера, извлеченного из коллекции.

В настоящем параграфе, используя γ-классификатор [259, 260] и цифровой текстовый портрет, предложенный в [255] и характеризующий распределение частотности буквенных унограмм, мы занимаемся идентификацией авторов произведений. Существенным моментом в сравнении с нашим предыдущим исследованием § 2.1.1 является изучение вопроса о минимально допустимом размере текстового фрагмента, для которого ещё удается получить удовлетворительный результат решения рассматриваемой задачи.

Модельная коллекция текстов, на которой разворачивается наше исследование, та же самая, что и в § 2.1.1.

Обработка коллекционного материала происходила в 4 этапа.

Этап 1 состоял из выбора двух произведений различных авторов, каковыми оказались «Рустам ва Сӯҳроб» А. Фирдоуси и «Дохунда» С. Айни. Из каждого произведения извлекались по 9 случайных выборок текстовых фрагментов, размеры которых в словах и символах показаны в таблице 2.4.

Таблица 2.4. – Информация о размерах фрагментов в словах и символах

Номер фрагмента	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Число слов	7000	3500	1700	900	450	250	130	60	20
Число символов	40000	20000	10000	5000	2500	1200	600	300	100

Как видно из таблицы, размеры фрагментов уменьшаются от номера к номеру.

Замечание. Формальное деление числа символов на соответствующее ему число слов не будет означать среднюю длину слова, исчисляемую количеством букв, поскольку к символам помимо букв относятся также знаки препинания и арифметические операции, цифры, обозначения типа «№», «@», «$» и т.п.

Этап 2. Для каждого фрагмента выбранных произведений строится ЦП, который определяется распределением частотности буквенных унограмм, содержащихся в рассматриваемом фрагменте.

ЦП представляется в табличном виде:

$$N: \quad 1 \quad 2 \quad \dots \quad 35$$

$$P: \quad p_1 \quad p_2 \quad \dots \quad p_{35},$$

в котором первая строка – номера символов, расположенных в алфавитном порядке, а вторая – относительные частоты встречаемости символов в тексте ***T***, причём $\sum_{i=1}^{35} p_i = 1.$

Этап 3. Вычисления расстояний $\rho(T_1, T_2)$ между ЦП 9 фрагментов $\boldsymbol{T_1}$ и 12

произведениями $\boldsymbol{T_2}$ рассматриваемой коллекции текстов. Соответствующие вычисления производились по формуле

$$\rho(T_1, T_2) = \sqrt{\frac{35}{2}} \max_s \left| \sum_{i=1}^{s} p_i^{(1)} - p_i^{(2)} \right| \qquad (2.4)$$

где $p_i^{(1)}$ и $p_i^{(2)}$ – частоты встречаемости в фрагментах $\boldsymbol{T_1}$ и в произведениях $\boldsymbol{T_2}$ буквенных униграмм i (i = 1, … , 35) и (s=1, … , 35).

Этап 4. Определение автора текстового фрагмента производится методом ближайшего соседа, см., например, [248]. Существо метода заключается в том, что классифицируемый фрагмент $\boldsymbol{T_1}$ объявляется принадлежащим тому автору, чьё произведение $\boldsymbol{T_2}$ в сравнении с другими произведениями оказывается наиболее «сходным» с фрагментом. Иными словами, рассматриваемые объекты являются ближайшими соседями, и расстояния между их ЦП минимальны в сравнении с прочими расстояниями.

Полученные результаты сведены в две группы таблиц соответственно с номерами 2.5.1, 2.5.2, 2.5.3 и 2.6.1, 2.6.2, 2.6.3. В ячейках таблиц представлены расстояния от 12 произведений модельной коллекции текстов до 9 фрагментов из романа С. Айни «Дохунда» (в 1-й группе) и до 9 фрагментов из поэмы А. Фирдоуси «Рустам ва Сӯҳроб» (во 2-й группе).

В этих таблицах используются те же сокращения для имен авторов и названий произведений, что и в § 2.1.1, а именно: А. Фирдоуси «Бежан бо Манижа» (АФ, Б&М, 14799) и «Рустам ва Сӯҳроб» (АФ, Р&С, 16355); Дж. Руми «Маснавии Маънавӣ, Дафтари 1» (ҶР, ММ1, 48713) и «Маснавии Маънавӣ, Дафтари 2» (ҶР, ММ2, 41661); А. Суруш «Дафтари 1» (АС, Д1, 7890) и «Дафтари 2» (АС, Д2, 9322); С. Айни «Одина» (СА, О, 25446), «Аҳмади Девбанд» (СА, АД, 7480), «Дохунда» (СА, Д, 71134) и «Марги судхӯр» (СА, МС, 48801); С. Турсун «Нисфирӯзӣ» (СТ, Н, 9936) и «Повести Камони Рустам» (СТ, ПКР, 4041). Для авторов и их произведений приняты обозначения, указываемые в скобках: первые две буквы – это инициалы авторов, вторые – сокращенные шифры текстов, третьи – число слов в произведениях.

В последующих таблицах первые 2 колонки указывают авторов и их произведения, а ячейки 9-ти других колонок – значения расстояний фрагментов различных длин до соответствующих произведений, вычисленные по формуле (2.4). Отметим также, что в названиях таблиц используются фразы о фрагментах, извлеченных из ***«начала»***, ***«середины»*** и ***«конца»*** произведений, тем самым в определенном смысле подчеркивается случайный характер выбора фрагментов из текстов произведений.

Таблица 2.5.1. – Расстояния между ЦПП из коллекции текстов и фрагментами, извлеченными из ***«начала»*** романа С. Айни «Дохунда»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**Р&С**	0.1533	0.1752	0.1697	0.2057	0.1719	0.1635	0.1216	0.1301	0.1408
	Б&М	0.1368	0.1587	0.1531	0.1891	0.1553	0.1455	0.1124	0.1322	0.1395
ҶР	**ММ1**	0.1871	0.1956	0.1782	0.2139	0.1893	0.2082	0.1834	0.1493	0.2026
	ММ2	0.1818	0.1956	0.1891	0.2238	0.1972	0.2207	0.1959	0.1593	0.2151
АС	**Д1**	0.1088	0.1091	0.0899	0.1078	0.1361	0.1619	0.2204	0.2395	0.1681
	Д2	0.1155	0.1158	0.0706	0.1061	0.1129	0.1387	0.1972	0.2162	0.1469
СТ	**Н**	0.0746	0.0964	0.0909	0.1269	0.1374	0.1418	0.1704	0.2035	0.2021
	ПКР	0.0634	0.0645	0.0944	0.0942	0.1514	0.1559	0.1849	0.2045	0.2031
СА	**О**	**0.0529**	**0.0416**	**0.0851**	**0.0803**	**0.1435**	**0.1481**	**0.1986**	**0.2177**	**0.1906**
	АД	**0.0714**	**0.0805**	**0.1391**	**0.1344**	**0.1975**	**0.2021**	**0.2076**	**0.2545**	**0.2531**
	Д	**0.0123**	**0.0301**	**0.0696**	**0.0649**	**0.1281**	**0.1326**	**0.1726**	**0.1917**	**0.1757**
	МС	**0.0357**	**0.0552**	**0.0727**	**0.0651**	**0.1282**	**0.1327**	**0.1762**	**0.1952**	**0.1621**

Закрашенные ячейки этой таблицы показывают, что из 5 фрагментов, взятых из «начала» романа С. Айни «Дохунда», все 5 оказались ближайшими соседями для самого произведения. Кроме того, всего лишь 1 фрагмент (размером в 2500) оказался ближайшим соседом поэмы А. Суруш «Дафтари 2». Интересно, что 3 самых маленьких фрагмента (в 600, 300 и 100 символов) оказались ближайшими соседями с поэмами А. Фирдоуси «Бежан бо Манижа» и «Рустам ва Сӯҳроб».

Таблица 2.5.2. – Расстояния между ЦПП из коллекции текстов и фрагментами, извлеченными из ***«середины»*** романа С. Айни «Дохунда»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**Р&С**	0.1406	0.1571	0.1473	0.1946	0.1829	0.1839	0.1956	0.2651	0.2574
	Б&М	0.1491	0.1589	0.1382	0.1781	0.1664	0.1673	0.2017	0.2712	0.2716
ҶР	**ММ1**	0.1437	0.1746	0.1655	0.1982	0.1838	0.1935	0.2351	0.3045	0.2701
	ММ2	0.1495	0.1871	0.1715	0.2107	0.1896	0.1951	0.2365	0.3059	0.2715
АС	**Д1**	0.1046	0.1172	0.1298	0.1069	0.1051	0.1458	0.1956	0.2247	0.2538
	Д2	0.0836	0.0962	0.1088	0.1136	0.0839	0.1247	0.1746	0.2095	0.2328
СТ	**Н**	0.0621	0.1095	0.0993	0.1158	0.1048	0.1333	0.1466	0.2447	0.2973
	ПКР	0.0684	0.0771	0.0784	0.0831	0.0719	0.0992	0.1442	0.2153	0.3223
СА	**О**	**0.0982**	**0.1108**	**0.1234**	**0.0711**	**0.0986**	**0.1393**	**0.1892**	**0.2183**	**0.2891**
	АД	**0.1127**	**0.1222**	**0.1299**	**0.0776**	**0.1051**	**0.1458**	**0.1957**	**0.2247**	**0.3222**
	Д	**0.0425**	**0.0513**	**0.0605**	**0.0564**	**0.0526**	**0.0938**	**0.1293**	**0.1866**	**0.2804**
	МС	**0.0579**	**0.0581**	**0.0707**	**0.0477**	**0.0485**	**0.0867**	**0.1365**	**0.1885**	**0.2881**

Как явствует из этой таблицы, для 8 фрагментов, взятых из «середины» романа «Дохунда», 5 оказались ближайшими соседями для самого произведения, а 3 – ближайшими соседями для «Марги судхӯр». Кроме того, всего лишь 1 фрагмент (размером в 100) оказался ближайшим соседом поэмы А. Суруш «Дафтари 2».

Таблица 2.5.3. – Расстояния между ЦПП из коллекции текстов и фрагментами, извлеченными из ***«конца»*** романа С. Айни «Дохунда»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**Р&С**	0.1792	0.1833	0.1721	0.1794	0.2258	0.2237	0.1889	0.2657	0.2344
	Б&М	0.1627	0.1668	0.1554	0.1878	0.2311	0.2289	0.1941	0.2744	0.2524
ЧР	**ММ1**	0.2029	0.2158	0.2176	0.2554	0.2923	0.3035	0.2596	0.2951	0.3006
	ММ2	0.1979	0.2102	0.2121	0.2457	0.2826	0.2938	0.2499	0.2886	0.2944
АС	**Д1**	0.0996	0.1375	0.1393	0.1646	0.1889	0.2028	0.1531	0.2691	0.3212
	Д2	0.1158	0.1442	0.1461	0.1713	0.2031	0.2192	0.1694	0.2651	0.3228
СТ	**Н**	0.1004	0.1046	0.0932	0.1043	0.1412	0.1523	0.1451	0.2776	0.3044
	ПКР	0.0677	0.0718	0.0605	0.0896	0.1339	0.1377	0.0998	0.2323	0.3383
СА	**О**	**0.0334**	**0.0312**	**0.0301**	**0.0574**	**0.0908**	**0.1068**	**0.1208**	**0.2605**	**0.4085**
	АД	**0.0742**	**0.0421**	**0.0366**	**0.0511**	**0.1022**	**0.1001**	**0.1237**	**0.2693**	**0.4172**
	Д	**0.0314**	**0.0411**	**0.0431**	**0.0823**	**0.1217**	**0.1302**	**0.1062**	**0.2299**	**0.3705**
	МС	**0.0479**	**0.0565**	**0.0561**	**0.0922**	**0.1331**	**0.1442**	**0.1026**	**0.2503**	**0.3983**

Для фрагментов из «конца» романа «Дохунда» (размерами не менее 1200 и 300 символов) основной результат – тот же, что и в 2-х предыдущих случаях: ближайшими для них соседями служат только произведения С. Айни. Интересно, что всего лишь 1 фрагмент (размером в 600) оказался ближайшим соседом с произведением С. Турсуна «Повести Камони Рустам». Особый интерес представляет собой «выброс», который указывает на то, что фрагмент размером в 100 символов из «конца» романа «Дохунда» выступает в качестве ближайшего соседа поэмы «Рустам ва Сӯҳроб», см. соответствующие ячейки.

Таблица 2.6.1. – Расстояния между ЦПП из коллекции текстов и фрагментами, извлеченными из ***«начала»*** поэмы А. Фирдоуси «Рустам ва Сӯҳроб»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**Р&С**	**0.0189**	**0.0275**	**0.0506**	**0.0732**	**0.0787**	**0.0851**	**0.1367**	**0.1601**	**0.4186**
	Б&М	**0.0338**	**0.0465**	**0.0686**	**0.0825**	**0.0967**	**0.0891**	**0.1313**	**0.1589**	**0.4377**
ЧР	**ММ1**	0.1219	0.1235	0.0995	0.1058	0.1341	0.1224	0.1372	0.1348	0.3759
	ММ2	0.1148	0.1153	0.0933	0.0988	0.1271	0.1238	0.1368	0.1516	0.3927
АС	**Д1**	0.2039	0.2124	0.2355	0.2537	0.2533	0.2699	0.3113	0.2978	0.5257
	Д2	0.1806	0.1892	0.2123	0.2304	0.2301	0.2467	0.2881	0.2746	0.4981
СТ	**Н**	0.1539	0.1625	0.1855	0.2037	0.2033	0.2201	0.2613	0.2479	0.4862
	ПКР	0.1683	0.1769	0.2001	0.2181	0.2177	0.2344	0.2758	0.2623	0.4931
СА	**О**	0.1969	0.2163	0.2272	0.2521	0.2682	0.2482	0.2895	0.3002	0.5442
	АД	0.1998	0.2191	0.2301	0.2551	0.2711	0.2403	0.2817	0.2959	0.5507
	Д	0.1561	0.1728	0.1877	0.2087	0.2248	0.2221	0.2635	0.2661	0.5115
	МС	0.1788	0.1981	0.2091	0.2339	0.2501	0.2257	0.2671	0.2769	0.5431

В этой и двух следующих таблицах 9 фрагментов выбираются из поэмы А. Фирдоуси «Рустам ва Сӯҳроб». Закрашенные ячейки показывают, что для 7 фрагментов ближайшим соседом является именно «Рустам ва Сӯҳроб». Интересно, что 2 самых маленьких фрагмента (в 300 и 100 символов) оказались ближайшими соседями с поэмой Дж. Руми «Маснавии Маънавӣ, Дафтари 1».

Таблица 2.6.2. Расстояния между ЦПП из коллекции текстов и фрагментами, извлеченными из *«**середины**»* поэмы А. Фирдоуси «Рустам ва Сӯҳроб»

Авторы (произв.)		Длины фрагментов (в символах)								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**P&C**	**0.0118**	**0.0237**	**0.0498**	**0.1316**	**0.0914**	**0.1241**	**0.1505**	**0.3616**	**0.1927**
	Б&М	**0.0242**	**0.0328**	**0.0555**	**0.1151**	**0.0755**	**0.1182**	**0.1446**	**0.3562**	**0.1872**
ЧР	**ММ1**	0.1339	0.1272	0.1429	0.1968	0.1937	0.2475	0.2739	0.3465	0.2334
	ММ2	0.1271	0.1202	0.1359	0.1871	0.1871	0.2407	0.2671	0.3474	0.2401
АС	**Д1**	0.1862	0.1905	0.1844	0.1327	0.1576	0.1682	0.1946	0.4902	0.3492
	Д2	0.1631	0.1673	0.1612	0.1199	0.1344	0.1741	0.2005	0.4671	0.3261
СТ	**Н**	0.1363	0.1406	0.1344	0.1242	0.1228	0.1297	0.1228	0.4403	0.2992
	ПКР	0.1507	0.1551	0.1489	0.1327	0.1313	0.1382	0.1313	0.4547	0.3137
СА	**О**	0.1882	0.1755	0.1626	0.1109	0.1359	0.1296	0.1536	0.4849	0.3274
	АД	0.1911	0.1784	0.1548	0.1157	0.1281	0.1301	0.1564	0.4897	0.3207
	Д	0.1448	0.1427	0.1366	0.0849	0.1098	0.1036	0.1101	0.4569	0.3014
	МС	0.1701	0.1573	0.1402	0.0922	0.1134	0.1091	0.1354	0.4642	0.3051

Закрашенные ячейки этой таблицы показывают, что из 5 фрагментов, взятых из «середины» поэмы А. Фирдоуси «Рустам ва Сӯҳроб», 3 оказались ближайшими соседями для самой поэмы, а 2 – ближайшими соседями для «Бежан бо Манижа». Кроме того, всего лишь 1 фрагмент (размером в 300) оказался ближайшим соседом поэмы Дж. Руми «Маснавии Маънавӣ, Дафтари 1». Интересно, что три других фрагмента (размерами в 5000, 1200 и 600) оказались ближайшими соседями произведения С. Айни «Дохунда».

Таблица 2.6.3. – Расстояния между ЦПП из коллекции текстов и фрагментами, извлеченными из *«**конца**»* поэмы А. Фирдоуси «Рустам ва Сӯҳроб»

Авторы (произв.)		Длины фрагментов (в символах)								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**P&C**	**0.0166**	**0.0156**	**0.0274**	**0.0568**	**0.0489**	**0.0782**	**0.0967**	**0.2033**	**0.3178**
	Б&М	**0.0251**	**0.0322**	**0.0383**	**0.0748**	**0.0655**	**0.0866**	**0.0975**	**0.2095**	**0.3145**
ЧР	**ММ1**	0.1362	0.1242	0.1399	0.1257	0.1311	0.1883	0.2116	0.2429	0.2843
	ММ2	0.1293	0.1181	0.1329	0.1195	0.1243	0.1814	0.2054	0.2479	0.2852
АС	**Д1**	0.1684	0.1751	0.1883	0.1878	0.1546	0.1068	0.1423	0.1596	0.4034
	Д2	0.1451	0.1519	0.1651	0.1646	0.1551	0.1211	0.1587	0.1726	0.3733
СТ	**Н**	0.1184	0.1251	0.1383	0.1378	0.1277	0.0788	0.1194	0.2044	0.4018
	ПКР	0.1329	0.1396	0.1528	0.1623	0.1604	0.0941	0.0809	0.1703	0.4057
СА	**О**	0.1919	0.2039	0.2131	0.2478	0.2459	0.1796	0.1003	0.1251	0.4226
	АД	0.1947	0.2068	0.2159	0.2506	0.2488	0.1825	0.1032	0.1341	0.4476
	Д	0.1484	0.1605	0.1696	0.2043	0.2025	0.1362	0.0688	0.1463	0.3947
	МС	0.1737	0.1857	0.1949	0.2296	0.2277	0.1615	0.0935	0.1482	0.4019

Как явствует из этой таблицы, для 6 фрагментов, взятых из «конца» поэмы А. Фирдоуси «Рустам ва Сӯҳроб», ближайшим соседом является именно «Рустам ва Сӯҳроб». Кроме того, всего лишь 1 фрагмент (размером в 100) оказался ближайшим соседом поэмы Дж. Руми «Маснавии Маънавӣ, Дафтари 1», а 2 других фрагмента (размерами в 600 и 300) оказались ближайшими соседями с произведениями С. Айни «Дохунда» и «Одина».

Заключение. Итак, результаты, представленные в таблицах, показывают, что ближайшими соседями по отношению к выбранным фрагментам являются, в основном, произведения именно того автора, из произведения которого извлекались сами фрагменты. В иной интерпретации это значит, что методом ближайшего соседа путем вычисления расстояний по формуле (2.4) представляется возможным установить авторство достаточно малого кусочка литературного произведения, причём для прозаических произведений (в сравнении с поэтическими) более успешно.

Для художественных текстов можно предложить оценку эффективности применяемого метода, опираясь на вполне естественную гипотезу, согласно которой фрагмент, извлекаемый из какого-либо произведения, должен быть «однородным» с любыми произведениями одного и того же автора. На языке «расстояний» этому соответствует утверждение о том, что ближайшими соседями искомого фрагмента являются, прежде всего, произведения его автора.

Для прозаического произведения, по данным таблиц 2.5.1, метод ближайшего соседа безошибочно определяет автора фрагментов размерами не менее 5000 символов.

По данным таблицы 2.5.2, метод безошибочно определяет автора 8 фрагментов из 9 и для 1 фрагмента размером 100 допускает ошибку.

По данным таблицы 2.5.3, метод ближайшего соседа безошибочно определяет автора 7 фрагментов из 9 и для 2-х фрагментов размерами 600 и 100 допускает ошибку.

Для поэтического произведения, по данным таблиц 2.6.1, метод ближайшего соседа безошибочно определяет автора 7 фрагментов из 9 и для 2-х фрагментов размерами 300 и 100 допускает ошибку.

По данным таблицы 2.6.2, метод безошибочно определяет автора 5 фрагментов из 9 и для 4-х фрагментов размерами 5000, 1200, 600 и 300 допускает ошибку.

По данным таблицы 2.6.3, метод ближайшего соседа безошибочно определяет автора 6 фрагментов из 9 и для 3-х фрагментов размерами 600, 300 и 100 допускает ошибку.

Результаты данного параграфа опубликованы в [63-А].

§ 2.1.3. Об определение автора текста на основе частотности символьных биграмм

Решается задача распознавания авторов произведений по отдельности для классической и современной поэзий, а также современной прозы. Произведениям сопоставляется ЦП, характеризуемый распределением в них частотности буквенных биграмм. Устанавливается эффективность применения γ-классификатора для идентификации авторов произведений.

В настоящем параграфе мы продолжаем тестирование количественных описаний текстов, начатое в работах [1-А-10-А], на предмет их пригодности для идентификации авторов произведений. В качестве таковых в [255, 6-А] рассматривались частотности букв таджикского алфавита (униграммы), в [7-А, 8-А] – буквенных биграмм и триграмм, в [283] – набора из пяти натуральных единиц измерения текста, в [256, 257] – частотности длин слов и знаков препинаний, в [14-А] – частотности слогов, в [258, 13-А, 17-А] – частотности длин предложений. Существенным моментом в сравнении с нашим предыдущим исследованием [7-А] является изучение вопроса о распознавании авторов текстов, относящихся к произведениям классической и современной поэзии, а также к современной прозе. Следуя [7-А], будем называть *цифровым портретом текста* распределение в нём частотности буквенных биграмм. В данном параграфе изучается вопрос об эффективности применения такого показателя для распознавания авторов поэтических и прозаических произведений.

1. Обработка статистического материала включала в себя 4 этапа.

Этап 1. Создание компьютерной программы и вычисление с её помощью ЦП произведений – распределений частотности биграмм по отдельности для всех текстов, упомянутых в § 2.1.1.

Этап 2. Создание компьютерной программы и вычисление с её помощью парных расстояний между ЦПП по формуле, предложенной в § 1.3.

Этап 3. Настройка γ-классификатора. Существо настройки заключается в том, чтобы определить такое значение вещественного параметра γ, при котором достигается максимальное значение критерия «γ-однородности» произведений, см. § 1.4.

Этап 4. Установление эффективности применения настроенного γ-классификатора для распознавания авторов произведений.

На этапе 1 цифровые портреты произведений представляются в табличном виде:

$$\begin{array}{lllllll} \overline{\mathrm{N}} & : & 1 & 2 & . & . & . \quad \mathrm{m} \\ \mathrm{P} & : & \mathrm{p}_1 & \mathrm{p}_2 & . & . & . \quad \mathrm{p}_\mathrm{m}, \end{array}$$

где первая строка – список биграмм; m – общее число биграмм; вторая строка – частоты p_i встречаемости в пределах произведений буквенных биграмм $i(i = 1,2, \dots, m)$, причём

$$\sum_{i=1}^{m} p_k = 1.$$

На этапе 2 вычисления расстояний $\rho(T_1, T_2)$ между текстами T_1 и T_2 производились по формуле T_1 и T_2

$$\rho(T_1,\, T_2) = \sqrt{\frac{m}{2}}\, \max_s \left| \sum_{k=1}^{s} (p_k^{(1)} - p_k^{(2)}) \right|,$$

в которой m (= 35^2) – количество биграмм; $p_k^{(1)}$ и $p_k^{(2)}$ – частоты встречаемости в текстах T_1 и T_2 суммарные количества буквенных биграмм k, k = 1, … , m, и $(\mathrm{s} = 1, \cdots, \mathrm{m})$.

Результаты вычислений показаны в таблицах 2.7-2.9.

На этапе 3 качество классификатора при фиксированном γ оценивается величиной π, вычисляемой по формуле

$$\pi = 1 - \tau / L\,, \tag{2.5}$$

где L (= 45) – суммарное число взаимных расстояний между 10 текстами исходной коллекции; $\tau = \tau(\gamma)$ – число нарушений неравенств

$$\rho(T_1,\, T_2) \le \gamma, \tag{2.6}$$

$$\rho(T_1,\, T_2) > \gamma. \tag{2.7}$$

Первое проверяется на 5 парах текстов одних и тех же авторов, второе – на 40 парах текстов различных авторов.

На этапе 4 производится настройка γ-классификатора на основе вполне естественной гипотезы о том, что произведения одного автора «однородны», а разных авторов «неоднородны». На языке ЦП, характеризующих распределение частотности буквенных биграмм 10 пар произведений, определение γ сводится к отысканию такого его значения, при котором общее число τ нарушений неравенств (2.6), (2.7) по отдельности на текстах 3-х модельных коллекций становится минимальным. Для нахождения таких γ используется алгоритм, предложенный в § 1.4.

2. Результаты вычислений расстояний между 10 произведениями классической поэзии представлены в табл. 2.7.

Таблица 2.7. – Расстояния между произведениями ***классической поэзии***

Автор (Произ.)		**Число слов**	**АР**		**АФ**		**СШ**		**ХШ**		**ЧР**	
			АП	**К**	**Р&С**	**Б&М**	**Ғ1**	**Ғ2**	**Ғ1**	**Ғ2**	**ММ1**	**ММ2**
			2248	**5054**	**16355**	**14799**	**16261**	**13001**	**33724**	**28923**	**48713**	**41661**
АР	**АП**	**2248**										
	К	**5054**	**0.3384**									
АФ	**Р&С**	**16355**	0.4535	0.4229								
	Б&М	**14799**	0.5212	0.4704	**0.1740**							
СШ	**Ғ1**	**16261**	0.9981	0.9486	1.2794	1.3945						
	Ғ2	**13001**	0.7989	0.7494	1.0801	1.1952	**0.4027**					
ХШ	**Ғ1**	**33724**	0.7262	0.6683	0.9839	1.0990	0.5490	0.3873				
	Ғ2	**28923**	0.8389	0.7921	1.0955	1.2105	0.6680	0.5697	**0.2135**			
ЧР	**ММ1**	**48713**	0.6093	0.5837	0.8461	0.9645	1.3058	1.0949	0.9987	1.1142		
	ММ2	**41661**	0.6217	0.5121	0.8036	0.9220	1.2439	1.0016	0.9054	1.0297	**0.1233**	

Для классической поэзии оптимальное значение γ оказалось следующим

$$\gamma^{opt} \in [0.3385;\ 0.3872),$$

Применять этот факт для выяснения метрической близости пары произведений T_1 и T_2 необходимо следующим образом, см. [16-А, 44-А]:

– если $\rho(T_1,\ T_2) \leq 0{,}3385$, то T_1 и T_2 однородны;

– если $\rho(T_1,\ T_2) > 0{,}3872$, то T_1 и T_2 неоднородны;

– если $0{,}3385 < \gamma \leq 0{,}3872$, то ситуация – неопределенная.

Из данных таблицы следует, что только одно расстояние, именно 0.4027 соответственно между ЦП двух произведений С. Шерозӣ «Ғазалиёт қисми 1» и «Ғазалиёт қисми 2» нарушает сформулированную гипотезу. Эти пары согласно (2.7) утверждают неоднородность указанных двух произведений С. Шерозӣ, хотя принадлежат одному автору.

Желтым цветом в таблице 2.7 отмечен 1 случай нарушения гипотезы однородности.

3. Результаты вычислений расстояний между 10 произведениями современной поэзии представлены в табл. 2.8.

Таблица 2.8. – Расстояния между произведениями в ***современной поэзии***

Автор (Произ.)		Число слов	АС		АШ		ГС		ИФ		МТ	
			Д1	Д2	БТ	ШР	О	Ш	101Г	МГМ	ҚҲ	ҲА
			7890	9322	32036	12810	12103	51434	9841	41217	8463	6118
АС	Д1	7890										
	Д2	9322	**0.1887**									
АШ	БТ	32036	0.6275	0.4638								
	ШР	12810	0.5935	0.4292	**0.1093**							
ГС	О	12103	0.5520	0.3840	0.3468	0.3552						
	Ш	51434	0.5329	0.3705	0.4451	0.4535	**0.1836**					
ИФ	101Г	9841	0.9230	0.7859	0.6071	0.6209	0.4388	0.4744				
	МГМ	41217	0.8663	0.7238	0.5736	0.5692	0.3843	0.4199	**0.1558**			
МТ	ҚҲ	8463	0.7179	0.6019	0.5080	0.5316	0.3571	0.2963	0.3094	0.2886		
	ҲА	6118	0.6048	0.5047	0.2678	0.3015	0.4346	0.4946	0.6975	0.6349	**0.5715**	

Для современной поэзии оптимальное значение γ оказалось следующим

$$\gamma^{opt} \in [0.1888;\ 0.2677).$$

Применять этот факт для выяснения метрической близости пары произведений T_1 и T_2 необходимо следующим образом, см. [16-А, 44-А]:

– если $\rho(T_1,\ T_2) \leq 0.1888$, то T_1 и T_2 однородны;

– если $\rho(T_1,\ T_2) > 0.2677$, то T_1 и T_2 неоднородны;

– если $0.1888 < \gamma \leq 0.2677$, то ситуация – неопределенная.

И здесь в табл. 2.8 закрашенные жёлтым цветом ячейки (в данном случае их – 1) показывают нарушение сформулированной гипотезы для соответствующих пар произведений.

4. Результаты вычислений расстояний между 10 произведениями современной прозы представлены в табл. 2.9.

Таблица 2.9. – Расстояния между произведениями в ***современной прозе***

Автор (Произ.)		Число слов	АЗ		ГМ		МШ		СТ		СА	
			Б	З	БМ	СМ	СБ	Х	Н	ПКР	Д	МС
			70804	79431	46608	50368	113592	91202	9936	4041	71134	48801
АЗ	Б	70804										
	З	79431	0.3042									
ГМ	БМ	46608	0.2222	0.2410								
	СМ	50368	0.2409	0.2931	0.1400							
МШ	СБ	113592	0.7955	0.6454	0.6775	0.7799						
	Х	91202	0.9333	0.7807	0.8147	0.9189	0.4087					
СТ	Н	9936	0.5109	0.6367	0.4290	0.3437	1.0852	1.2205				
	ПКР	4041	0.6867	0.8125	0.5801	0.5195	1.1189	1.2565	0.3467			
СА	Д	71134	0.4140	0.5384	0.3772	0.3581	0.9234	1.0592	0.5229	0.3364		
	МС	48801	0.5129	0.6387	0.4345	0.4512	0.9377	1.0108	0.6492	0.4421	0.1897	

Для современной прозы оптимальное значение γ оказалось следующим

$$\gamma^{opt} \in [0.1898;\ 0.2221).$$

Применять этот факт для выяснения метрической близости пары произведений T_1 и T_2 современной прозы необходимо следующим образом, см. [16-А, 44-А]:

– если $\rho(T_1, T_2) \leq 0.1898$, то T_1 и T_2 однородны;

– если $\rho(T_1, T_2) > 0.2221$, то T_1 и T_2 неоднородны;

– если $0.1898 < \gamma \leq 0.2221$, то ситуация – неопределенная.

И здесь закрашенные в табл. 2.9 жёлтым цветом ячейки (в данном случае их – 3) показывают нарушение сформулированной гипотезы для соответствующих пар произведений.

5. Вычисления по формуле (2.5) коэффициента эффективности π*:*

– для классической поэзии выдает значение π = 98%,

– для современной поэзии выдаёт значение π = 98%,

– для современной прозы выдаёт значение π = 93%.

распознавания автора по цифровому портрету его произведений.

Полученные значения показывают, что распознавание автора текста по цифровому портрету (распределению частотности буквенных биграмм) для поэтических произведений (в сравнении с прозаическими) более успешно.

Результаты данного параграфа опубликованы в [49-А].

§ 2.1.4. Об идентификации автора текстового фрагмента на основе частотности символьных биграмм

На примере модельной коллекции таджикских литературных произведений изучается задача о возможности определения авторства фрагмента текста

минимального размера, извлеченного из коллекции. Рассматривается модельная коллекция текстов таджикского языка, составленная из произведений классической поэзии и современной прозы на кириллической графике. Каждому произведению сопоставлен ЦП – распределение частотностей символьных биграмм. Для решения проблемы идентификации авторов текстов биграммы являются вполне приемлемыми количественными характеристиками.

В этом пункте, используя γ-классификатор §§ 1.3-1.4 и цифровой текстовый портрет из § 2.1.3, характеризующий распределение частотности буквенных биграмм, приводится описание процесса идентификации авторов произведений таджикского текста. Отметим, что ранее аналогичный вопрос изучался именно для символьных (буквенных) униграмм, биграмм и триграмм с учетом пробела [10-А]. Существенным моментом в сравнении с нашим предыдущим исследованием § 2.1.3 является изучение вопроса о минимально допустимом размере текстового фрагмента, для которого удается получить удовлетворительный результат решения рассматриваемой задачи. Отметим также, что в сравнении с исследованием [10-А] выбора фрагментов из текстов, извлеченных из *«начала»*, *«середины»* и *«конца»* произведений, решается задача по отдельности для поэзий и прозы.

Модельная коллекция текстов, на которой разворачивается наше исследование, та же самая, что и в § 2.1.1.

Обработка коллекционного материала происходила в 4 этапа.

Этап 1 состоял из выбора двух произведений различных авторов, каковыми оказались «Рустам ва Сухроб» А. Фирдоуси и «Дохунда» С. Айни. Из каждого произведения извлекались по 9 случайных выборок текстовых фрагментов, размеры которых в словах и символах показаны в таблице 2.10.

Таблица 2.10. Информация о размерах фрагментов в словах и символах

Номер фрагмента	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Число слов	7000	3500	1700	900	450	250	130	60	20
Число символов	40000	20000	10000	5000	2500	1200	600	300	100

Как видно из таблицы, размеры фрагментов уменьшаются от номера к номеру.

Замечание. Формальное деление числа символов на соответствующее ему число слов не будет означать среднюю длину слова, исчисляемую количеством букв, поскольку к символам помимо букв относятся также знаки препинания и арифметических операций, цифры, обозначения типа «№», «@», «$» и т.п.

Этап 2. Для каждого фрагмента выбранных произведений строится ЦП, который определяется распределением частотности буквенных биграмм, содержащихся в рассматриваемом фрагменте.

ЦП представляется в табличном виде:

$$N: \quad 1 \quad 2 \quad \ldots \quad 1225$$
$$P: \quad p_1 \quad p_2 \quad \ldots \quad p_{1225},$$

в котором первая строка – номера биграмм, расположенных в алфавитном порядке, а вторая – относительные частоты встречаемости буквенных биграмм в тексте $\boldsymbol{T}$, причём $\sum_{i=1}^{1225} p_i = 1$.

Этап 3. Вычисление расстояний $\rho(T_1, T_2)$ между ЦП 9 фрагментов $\boldsymbol{T_1}$ и 12 произведениями $\boldsymbol{T_2}$ рассматриваемой коллекции текстов. Соответствующие вычисления производились по формуле

$$\rho(T_1, T_2) = \sqrt{\frac{1225}{2}} \max_s \left| \sum_{i=1}^{s} p_i^{(1)} - p_i^{(2)} \right| \qquad (2.8)$$

где $p_i^{(1)}$ и $p_i^{(2)}$ – частоты встречаемости в фрагментах $\boldsymbol{T_1}$ и в произведениях $\boldsymbol{T_2}$ буквенных биграмм i (i = 1, … , 1225) и (s=1, … , 1225).

Этап 4. Определение автора текстового фрагмента производится методом ближайшего соседа [248, 249]. Сущность метода заключается в том, что классифицируемый фрагмент $\boldsymbol{T_1}$ объявляется принадлежащим тому автору, чьё произведение $\boldsymbol{T_2}$ в сравнении с другими произведениями оказывается наиболее «сходным» с фрагментом. Иными словами, рассматриваемые объекты являются ближайшими соседями, и расстояния между их ЦП минимальны в сравнении с прочими расстояниями.

Полученные результаты сведены в две группы таблиц соответственно с номерами 2.11.1, 2.11.2, 2.11.3 и 2.12.1, 2.12.2, 2.12.3. В ячейках таблиц представлены расстояния от 12 произведений модельной коллекции текстов до 9 фрагментов из романа С. Айни «Дохунда» (в 1-й группе) и до 9 фрагментов из поэмы А. Фирдоуси «Рустам ва Сухроб» (во 2-й группе). В этих таблицах используются те же сокращения для имен авторов и названий произведений, что и в § 2.1.1.

В последующих таблицах первые 2 колонки указывают авторов и их произведения, а ячейки 9-и других колонок – значения расстояний фрагментов различных длин до соответствующих произведений, вычисленные по формуле (2.8). Отметим также, что в названиях таблиц используются фразы о фрагментах, извлеченных из ***«начала»***, ***«середины»*** и ***«конца»*** произведений, тем самым в определенном смысле подчеркивается случайный характер выбора фрагментов из текстов произведений.

Таблица 2.11.1. – Расстояния между ЦП произведений из коллекции текстов и фрагментами, извлеченными из ***«начала»*** романа С. Айни «Дохунда»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**Р&С**	0.5573	0.5982	0.6717	0.7503	1.0922	1.2611	1.6813	1.5618	2.3129
	Б&М	0.6544	0.6716	0.6682	0.7212	1.1462	1.3140	1.7159	1.6335	2.2528
ЧР	**ММ1**	0.7584	0.8357	0.7952	1.0484	0.9381	1.1166	1.4725	1.7174	2.4685
	ММ2	0.8130	0.8841	0.8559	1.1104	1.0265	1.2275	1.5834	1.7827	2.5338
АС	**Д1**	0.4965	0.5156	0.9310	0.9761	1.2956	1.4943	1.8959	2.1686	2.9197
	Д2	0.3919	0.4851	0.7735	0.7849	1.1836	1.3003	1.7000	1.9239	2.6749
СТ	**Н**	0.4577	0.4234	0.7571	0.7876	1.2746	1.4555	1.8483	2.1166	2.8677
	ПКР	0.5441	0.5874	0.8607	0.8800	1.2978	1.4787	1.8927	2.2428	2.9939
СА	**О**	**0.3130**	**0.3175**	**0.6645**	**0.7096**	**1.0283**	**1.2285**	**1.6214**	**2.0498**	**2.8009**
	АД	**0.5506**	**0.6049**	**1.0013**	**0.9192**	**1.3278**	**1.4543**	**1.8821**	**2.2721**	**3.0232**
	Д	**0.1232**	**0.1482**	**0.6078**	**0.6529**	**0.9750**	**1.1686**	**1.5704**	**1.9388**	**2.6899**
	МС	**0.2562**	**0.2456**	**0.6564**	**0.7016**	**1.0170**	**1.2144**	**1.5983**	**1.9371**	**2.6882**

Закрашенные ячейки этой таблицы показывают, что из 4 фрагментов, взятых из «начала» романа С. Айни «Дохунда», все 4 оказались ближайшими соседями для самого произведения. Кроме того, ещё 3 фрагмента (размерами в 2500, 1200 и 600) оказались ближайшим соседом с поэмой Дж. Руми «Маснавии Маънави, Дафтари 1». Интересно, что 2 самых маленьких фрагмента (в 300 и 100 символов) оказались ближайшими соседями с поэмами А. Фирдоуси «Рустам ва Сухроб» и «Бежан бо Манижа».

Таблица 2.11.2. – Расстояния между ЦП произведений из коллекции текстов и фрагментами, извлеченными из ***«середины»*** романа С. Айни «Дохунда»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**Р&С**	0.5732	0.8167	0.7396	1.0583	0.9825	1.0452	1.1623	1.7480	1.6993
	Б&М	0.7022	0.9456	0.8792	1.1253	1.1185	1.1427	1.2877	1.7797	1.6941
ЧР	**ММ1**	0.7356	1.0244	0.9692	1.1792	1.0078	1.1868	1.3460	2.0187	1.9645
	ММ2	0.7907	1.0878	1.0124	1.2333	1.0619	1.2722	1.3939	2.0629	2.0087
АС	**Д1**	0.5232	0.6087	0.6365	0.7520	0.6976	0.8247	0.9005	1.5000	1.6306
	Д2	0.4695	0.6444	0.6177	0.8370	0.8024	0.9197	0.9955	1.5950	1.5978
СТ	**Н**	0.4576	0.6337	0.5411	0.7787	0.7073	0.8797	0.9555	1.5550	1.6652
	ПКР	0.5479	0.5887	0.5823	0.6048	0.6376	0.6899	0.7863	1.4214	1.8266
СА	**О**	**0.2970**	**0.2827**	**0.3377**	**0.3843**	**0.3892**	**0.5474**	**1.2148**	**1.2685**	**1.4150**
	АД	**0.6741**	**0.7069**	**0.6588**	**0.5567**	**0.7036**	**0.6747**	**0.9854**	**1.4066**	**1.9524**
	Д	**0.1766**	**0.4375**	**0.3604**	**0.5681**	**0.5244**	**0.7299**	**1.0766**	**1.4519**	**1.4266**
	МС	**0.2846**	**0.3704**	**0.3081**	**0.4759**	**0.4246**	**0.6376**	**1.0314**	**1.3129**	**1.4253**

Как явствует из этой таблицы, для 8 фрагментов, взятых из «середины» романа «Дохунда», 1 фрагмент оказался ближайшим соседом для самого произведения, 1 фрагмент – ближайшим соседом для «Марги судхур» и 6 фрагмент – ближайшим соседом для «Одина». Кроме того, всего лишь 1 фрагмент (размером в 600) оказался ближайшим соседом произведения С. Турсуна «Повести Камони Рустам».

Таблица 2.11.3. – Расстояния между ЦП произведений из коллекции текстов и фрагментами, извлеченными из ***«конца»*** романа С. Айни «Дохунда»

Авторы (произв.)		Длины фрагментов (в символах)								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**P&C**	0.5628	0.6208	0.5537	0.6202	0.6871	0.9209	0.8734	1.7209	1.9824
	Б&М	0.6697	0.7422	0.6638	0.7522	0.8067	0.9392	0.8917	1.7392	2.0540
ЧР	**ММ1**	0.7293	0.7597	0.7709	0.8398	0.9750	1.1842	1.1367	1.9842	1.8407
	ММ2	0.7834	0.8120	0.8192	0.8353	1.0301	1.1866	1.1438	1.9587	1.8086
АС	**Д1**	0.5978	0.5670	0.7384	0.5152	0.8140	1.0133	0.9722	1.8133	2.1899
	Д2	0.4811	0.3955	0.4918	0.5168	0.8001	1.0124	0.9649	1.8124	2.0716
СТ	**Н**	0.5052	0.4906	0.6405	0.4039	0.6616	0.9098	0.9120	1.7098	2.2713
	ПКР	0.6484	0.6176	0.7889	0.5160	0.4772	0.6742	1.0508	1.8593	2.3876
СА	**О**	**0.3654**	**0.3270**	**0.4934**	**0.4218**	**0.4954**	**0.7064**	**0.9380**	**1.9927**	**2.5028**
	АД	**0.5793**	**0.4900**	**0.6613**	**0.5182**	**0.5652**	**0.8035**	**1.0971**	**2.2048**	**2.7042**
	Д	**0.2219**	**0.1888**	**0.3601**	**0.3788**	**0.5171**	**0.7438**	**0.7492**	**1.8338**	**2.3522**
	МС	**0.2999**	**0.2691**	**0.4404**	**0.4816**	**0.5185**	**0.7634**	**0.9229**	**1.9314**	**2.4918**

Для фрагментов из «конца» романа «Дохунда» (размерами не менее 5000 и 600 символов) основной результат – тот же, что и в 2-х предыдущих случаях: ближайшими для них соседями служат только произведения С. Айни. Кроме того, всего лишь 1 фрагмент (размером в 100) оказался ближайшим соседом поэмы Дж. Руми «Маснавии Маънави, Дафтари 2». Интересно, что три других фрагмента (размерами в 2500, 1200 и 300) оказались ближайшими соседями произведений С. Турсуна «Повести Камони Рустам» и «Нисфирузи».

Таблица 2.12.1. – Расстояния между ЦП произведений из коллекции текстов и фрагментами, извлеченными из ***«начала»*** поэмы А. Фирдоуси «Рустам ва Сухроб»

Авторы (произв.)		Длины фрагментов (в символах)								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**P&C**	**0.1546**	**0.1875**	**0.4082**	**0.4992**	**0.5481**	**0.8051**	**1.1439**	**1.3656**	**4.5971**
	Б&М	**0.2449**	**0.2787**	**0.4306**	**0.4558**	**0.5225**	**0.7216**	**0.9941**	**1.3299**	**4.6489**
ЧР	**ММ1**	0.4687	0.5032	0.6811	0.8072	0.8467	1.1006	1.3661	1.5885	4.2512
	ММ2	0.5336	0.5681	0.7458	0.8851	0.9092	1.1616	1.4271	1.6495	4.2329
АС	**Д1**	0.9457	0.9818	1.1572	1.2691	1.3285	1.5824	1.8478	2.0703	4.8321
	Д2	0.7414	0.7701	0.9494	1.0648	1.1243	1.3782	1.6435	1.8661	4.6062
СТ	**Н**	0.8734	0.9063	1.0663	1.2237	1.2471	1.4468	1.7121	1.9346	4.7704
	ПКР	1.0088	1.0431	1.2001	1.3605	1.3809	1.5375	1.8316	2.0216	4.8087
СА	**О**	0.8587	0.8447	1.0526	1.1802	1.2274	1.2817	1.6181	1.7696	5.0472
	АД	0.9738	0.9976	1.1677	1.3137	1.3856	1.5231	1.8619	1.9381	5.1542
	Д	0.6547	0.6429	0.8486	0.9774	1.0246	1.2165	1.5178	1.7043	4.8671
	МС	0.7363	0.7583	0.9302	1.0455	1.1163	1.2878	1.5532	1.7757	4.9891

В этой (табл. 2.12.1) и двух следующих таблицах (табл. 2.12.2, 2.12.3) 9 фрагментов выбираются из поэмы А. Фирдоуси «Рустам ва Сухроб». Закрашенные ячейки этой таблицы показывают, что из 8 фрагментов, взятых из «начала» поэмы А. Фирдоуси «Рустам ва Сухроб», 3 оказались ближайшими соседями для самой поэмы, а 5 – ближайшими соседями для «Бежан бо Манижа». Кроме того, всего лишь 1 фрагмент (размером в 100) оказался ближайшим соседом поэмы Дж. Руми «Маснавии Маънави, Дафтари 2».

Таблица 2.12.2. – Расстояния между ЦП произведений из коллекции текстов и фрагментами, извлеченными из ***«середины»*** поэмы А. Фирдоуси «Рустам ва Сухроб»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**P&C**	**0.0801**	**0.2232**	**0.3465**	**0.7885**	**0.6762**	**0.6125**	**0.7236**	**2.4113**	**1.5991**
	Б&М	**0.2058**	**0.2326**	**0.3856**	**0.7948**	**0.6072**	**0.5566**	**0.7761**	**2.3231**	**1.7191**
ЧР	**ММ1**	0.5624	0.5867	0.7125	1.1195	1.1087	0.9869	1.1738	2.5881	1.3655
	ММ2	0.6155	0.6345	0.7721	1.1617	1.1349	1.0401	1.2224	2.6382	1.4321
АС	**Д1**	0.8553	0.8839	0.7788	0.6378	0.7845	0.8601	1.0698	3.0289	1.8648
	Д2	0.6511	0.6599	0.5745	0.7457	0.7249	0.7512	0.8533	2.8102	1.6723
СТ	**Н**	0.7526	0.8194	0.6739	0.6135	0.6888	0.7802	0.9861	2.9363	1.7576
	ПКР	0.8871	0.9548	0.8106	0.6946	0.8351	0.9169	1.1228	3.0518	1.8553
СА	**О**	0.8573	0.8623	0.6692	0.4569	0.5918	0.7255	0.9591	2.9402	1.5882
	АД	0.9826	0.9774	0.7842	0.5721	0.7092	0.8406	1.0741	3.1706	1.9954
	Д	0.6533	0.6583	0.4652	0.5612	0.4701	0.5329	0.7551	2.8259	1.5417
	МС	0.7349	0.7399	0.5467	0.4473	0.4866	0.6031	0.8366	2.8478	1.5855

Закрашенные ячейки этой таблицы показывают, что из 5 фрагментов, взятых из «середины» поэмы А. Фирдоуси «Рустам ва Сухроб», 4 оказались ближайшими соседями для самой поэмы, а 1 фрагмент – ближайшим соседом для «Бежан бо Манижа». Кроме того, всего лишь 1 фрагмент (размером в 100) оказался ближайшим соседом поэмы Дж. Руми «Маснавии Маънави, Дафтари 1». Интересно, что три других фрагмента (размерами в 5000, 2500 и 1200) оказались ближайшими соседями произведений С. Айни «Марги судхур» и «Дохунда».

Таблица 2.12.3. – Расстояния между ЦП произведений из коллекции текстов и фрагментами, извлеченными из ***«конца»*** поэмы А. Фирдоуси «Рустам ва Сухроб»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**P&C**	**0.1246**	**0.1531**	**0.1984**	**0.4345**	**0.4677**	**0.4627**	**0.6423**	**1.3389**	**4.0598**
	Б&М	**0.2475**	**0.2311**	**0.2388**	**0.4905**	**0.3773**	**0.4905**	**0.6784**	**1.3789**	**3.9715**
ЧР	**ММ1**	0.6055	0.5878	0.5971	0.7289	0.5801	0.9481	1.0529	1.5929	4.2365
	ММ2	0.5417	0.5269	0.5818	0.7849	0.6361	0.9959	0.9873	1.6322	4.2867
АС	**Д1**	0.7065	0.7092	0.8487	0.9051	0.8217	0.6754	1.0173	1.1852	4.6725
	Д2	0.5022	0.4627	0.6021	0.6585	0.7147	0.5423	0.8131	1.2518	4.5101
СТ	**Н**	0.6137	0.6284	0.7771	0.8261	0.7274	0.5776	0.9403	1.2659	4.5848
	ПКР	0.7473	0.7637	0.9105	0.9556	0.8722	0.7261	1.0225	1.1005	4.7781
СА	**О**	0.7797	0.7396	0.7767	1.0118	0.9397	0.8122	1.0347	1.3253	4.8335
	АД	0.9465	0.9064	0.9214	1.1816	1.1531	0.9964	1.0721	1.0502	4.9397
	Д	0.6013	0.5612	0.5762	0.8271	0.7557	0.7469	1.0266	1.3011	4.6307
	МС	0.7167	0.6766	0.6916	0.9486	0.8765	0.7351	0.9432	1.2235	4.6988

Как явствует из этой таблицы, для 8 фрагментов, взятых из «конца» поэмы А. Фирдоуси «Рустам ва Сухроб», 6 оказались ближайшими соседями для самой поэмы, а 2 – ближайшими соседями для «Бежан бо Манижа». Кроме того, всего лишь 1 фрагмент (размерам в 300) оказался ближайшим соседом произведения С. Айни «Ахмади Девбанд».

Заключение. Таким образом, результаты, представленные в таблицах, показывают, что ближайшими соседями по отношению к выбранным фрагментам

являются, в основном, произведения именно того автора, из произведения которого извлекались сами фрагменты. В иной интерпретации это значит, что методом ближайшего соседа путем вычисления расстояний по формуле (2.8) представляется возможным установить авторство достаточно малого кусочка литературного произведения, причём для поэтических произведений (в сравнении с прозаическими) более успешно.

Для художественных текстов можно предложить оценку эффективности применяемого метода, опираясь на вполне естественную гипотезу, согласно которой фрагмент, извлекаемый из какого-либо произведения, должен быть «однородным» с любыми произведениями одного и того же автора. На языке «расстояний» этому соответствует утверждение о том, что ближайшими соседями искомого фрагмента являются, прежде всего, произведения его автора.

Для прозаического произведения, по данным таблицы 2.11.1, метод ближайшего соседа безошибочно определяет автора фрагментов, состоящих из не менее 5000 символов.

По данным таблицы 2.11.2, метод безошибочно определяет автора 8 фрагментов из 9 и для 1 фрагмента размером 600 символов допускает ошибку.

По данным таблицы 2.11.3, метод ближайшего соседа безошибочно определяет автора 5 фрагментов из 9 и для 4-х фрагментов размерами 2500, 1200, 300 и 100 символов допускает ошибку.

Для поэтического произведения, по данным таблицы 2.12.1, метод ближайшего соседа безошибочно определяет автора 8 фрагментов из 9 и для 1 фрагмента размером 100 символов допускает ошибку.

По данным таблицы 2.12.2, метод безошибочно определяет автора 5 фрагментов из 9 и для 4-х фрагментов размерами 5000, 2500, 1200 и 300 символов допускает ошибку.

По данным таблицы 2.12.3, метод ближайшего соседа безошибочно определяет автора 8 фрагментов из 9 и для 1 фрагмента размером 300 символов допускает ошибку.

Результаты данного параграфа опубликованы в [28-А].

§ 2.1.5. Об определение автора текста на основе частотности символьных триграмм

Решается задача распознавания авторов произведений по отдельности для классической и современной поэзий, а также современной прозы. Произведениям сопоставляется ЦП, характеризуемый распределением в них частотности буквенных триграмм. Устанавливается эффективность применения γ-классификатора для идентификации авторов произведений. Сконструированы ЦП и метрическое пространство произведений. В предположении уникальности авторского творчества устанавливаются пороговые значения метрики, на основе

которых определяются классы «однородных» произведений. γ-классификатор дискретных случайных величин, подтвердивший высокую эффективность при идентификации авторства текстовых фрагментов в произведениях классической и современной поэзий, а также в современной прозе таджикского языка, тестируется на предмет приспособляемости к распознаванию авторства по отдельности.

В этом параграфе мы продолжаем тестирование количественных описаний текстов, начатое в работах [1-А-10-А], на предмет их пригодности для идентификации авторов произведений. В качестве таковых в [255, 6-А] рассматривались частотности букв таджикского алфавита (униграммы), в [7-А, 8-А] – буквенных биграмм и триграмм, в [283] – набора из пяти натуральных единиц измерения текста, в [256, 257] – частотности длин слов и знаков препинаний, в [14-А] – частотности слогов, в [258] – частотности длин предложений. Существенным моментом в сравнении с нашим предыдущим исследованием [8-А] является изучение вопроса о распознавании авторов текстов, относящихся к произведениям классической и современной поэзии, а также к современной прозе. Следуя [8-А], будем называть *цифровым портретом текста* распределение в нём частотности буквенных триграмм. В настоящем параграфе изучается вопрос об эффективности применения такого показателя для распознавания авторов поэтических и прозаических произведений.

1. Обработка статистического материала включала в себя 4 этапа.

Этап 1. Создание компьютерной программы и вычисление с её помощью ЦП произведений – распределений частотности триграмм по отдельности для всех текстов, упомянутых в § 2.1.1.

Этап 2. Создание компьютерной программы и вычисление с её помощью парных расстояний между ЦПП по формуле, предложенной в § 1.3.

Этап 3. Настройка γ-классификатора. Существо настройки заключается в том, чтобы определить такое значение вещественного параметра γ, при котором достигается максимальное значение критерия «γ-однородности» произведений, см. § 1.4.

Этап 4. Установление эффективности применения настроенного γ-классификатора для распознавания авторов произведений.

На этапе 1 цифровые портреты произведений представляются в табличном виде:

$$\begin{array}{lccccc} \bar{N}: & 1 & 2 & \ldots & m \\ P: & p_1 & p_2 & \ldots & p_m, \end{array}$$

где первая строка – список триграмм; m – общее число триграмм; вторая строка – частоты p_i встречаемости в пределах произведений буквенных триграмм $i(i =$

$1,2,\ldots,m$), причём

$$\sum_{i=1}^{m} p_k = 1.$$

На этапе 2 вычисления расстояний $\rho(T_1, T_2)$ между текстами T_1 и T_2 производились по формуле T_1 и T_2

$$\rho(T_1, T_2) = \sqrt{\frac{m}{2}} \max_s \left| \sum_{k=1}^{s} (p_k^{(1)} - p_k^{(2)}) \right|,$$

в которой m (= 35^3) – количество триграмм; $p_k^{(1)}$ и $p_k^{(2)}$ – частоты встречаемости в текстах T_1 и T_2 суммарные количества буквенных триграмм k, k = 1, … , m, и $(s = 1, \cdots, m)$.

Результаты вычислений показаны в таблицах 2.13-2.15.

На этапе 3 качество классификатора при фиксированном γ оценивается величиной π, вычисляемой по формуле

$$\pi = 1 - \tau/L, \qquad (2.9)$$

где L (= 45) – суммарное число взаимных расстояний между 10 текстами исходной коллекции; $\tau = \tau(\gamma)$ – число нарушений неравенств

$$\rho(T_1, T_2) \le \gamma, \qquad (2.10)$$

$$\rho(T_1, T_2) > \gamma. \qquad (2.11)$$

Первое проверяется на 5 парах текстов одних и тех же авторов, второе – на 40 парах текстов различных авторов.

На этапе 4 производится настройка γ-классификатора на основе вполне естественной гипотезы о том, что произведения одного автора «однородны», а разных авторов «неоднородны». На языке ЦП, характеризующих распределение частотности буквенных триграмм 10 пар произведений, определение γ сводится к отысканию такого его значения, при котором общее число τ нарушений неравенств (2.10), (2.11) по отдельности на текстах 3-х модельных коллекций становится минимальным. Для нахождения таких γ используется алгоритм, предложенный в § 1.4.

2. Результаты вычислений расстояний между 10 произведениями классической поэзии представлены в табл. 2.13.

Таблица 2.13. – Расстояния между произведениями ***классической поэзии***

Автор (Произ.)		Число слов	АР		АФ		СШ		ҲШ		ЧР	
			АП	Қ	Р&С	Б&М	Ғ1	Ғ2	Ғ1	Ғ2	ММ1	ММ2
			2248	5054	16355	14799	16261	13001	33724	28923	48713	41661
АР	АП	2248										
	Қ	5054	1.9960									
АФ	Р&С	16355	2.6926	2.5109								
	Б&М	14799	3.1829	2.7928	1.0377							
СШ	Ғ1	16261	5.9236	5.6347	7.5673	8.2500						
	Ғ2	13001	4.7421	4.4532	6.3881	7.0708	2.3930					
ҲШ	Ғ1	33724	4.3074	3.9610	5.8515	6.5131	3.2466	2.3004				
	Ғ2	28923	4.9709	4.6896	6.5248	7.1877	3.9504	3.3762	1.2768			
ЧР	ММ1	48713	3.7705	3.4513	5.0336	5.7404	7.7241	6.4930	5.9458	6.6181		
	ММ2	41661	3.8045	3.0253	4.7719	5.4950	7.3588	5.9720	5.3721	6.1058	0.7373	

Для классической поэзии оптимальное значение γ оказалось следующим

$$\gamma^{opt} \in [1.9961;\ 2.3003).$$

Применять этот факт для выяснения метрической близости пары произведений T_1 и T_2 необходимо следующим образом, см. [16-А, 44-А]:

– если $\rho(T_1,\ T_2) \leq 1.9961$, то T_1 и T_2 однородны;

– если $\rho(T_1,\ T_2) > 2.3003$, то T_1 и T_2 неоднородны;

– если $1.9961 < \gamma \leq 2.3003$, то ситуация – неопределенная.

Из данных таблицы следует, что только одно расстояние, именно 2.3930 соответственно между ЦП двух произведений С. Шерозӣ «Ғазалиёт қисми 1» и «Ғазалиёт қисми 2» нарушают сформулированную гипотезу. Эти пары согласно (2.11) утверждают неоднородность указанных двух произведений С. Шерозӣ, хотя принадлежат одному автору.

Желтым цветом в таблице 2.13 отмечен 1 случай нарушения гипотезы однородности.

3. Результаты вычислений расстояний между 10 произведениями современной поэзии представлены в табл. 2.14.

Таблица 2.14. – Расстояния между произведениями в ***современной поэзии***

Автор (Произ.)		Число слов	АС		АШ		ГС		ИФ		МТ	
			Д1	Д2	БТ	ШР	О	Ш	101Г	МГМ	ҚҲ	ҲА
			7890	9322	32036	12810	12103	51434	9841	41217	8463	6118
АС	Д1	7890										
	Д2	9322	1.1480									
АШ	БТ	32036	3.7376	2.7482								
	ШР	12810	3.5669	2.5489	0.6535							
ГС	О	12103	3.3618	2.3521	2.0539	2.1014						
	Ш	51434	3.2085	2.3274	2.6350	2.6850	1.1145					
ИФ	101Г	9841	5.6466	4.7739	3.6036	3.6775	2.6091	2.8322				
	МГМ	41217	5.1367	4.3999	3.3941	3.3685	2.2717	2.4922	0.9363			
МТ	ҚҲ	8463	4.2518	3.5704	3.0097	3.1498	2.1315	1.7561	1.8301	1.7112		
	ҲА	6118	3.5973	2.9893	1.5866	1.7859	2.5738	2.9507	4.1421	3.7573	3.3886	

Для современной поэзии оптимальное значение γ оказалось следующим

$$\gamma^{opt} \in [1.1481;\ 1.5865).$$

Применять этот факт для выяснения метрической близости пары произведений T_1 и T_2 необходимо следующим образом, см. [16-А, 44-А]:

– если $\rho(T_1,\ T_2) \leq 1.1481$, то T_1 и T_2 однородны;

– если $\rho(T_1,\ T_2) > 1.5865$, то T_1 и T_2 неоднородны;

– если $1.1481 < \gamma \leq 1.5865$, то ситуация – неопределенная.

И здесь в табл. 2.14 закрашенные жёлтым цветом ячейки (в данном случае их – 1) показывают нарушение сформулированной гипотезы для соответствующих пар произведений.

4. Результаты вычислений расстояний между 10 произведениями современной прозы представлены в табл. 2.15.

Таблица 2.15. – Расстояния между произведениями в ***современной прозе***

Автор (Произ.)		**Число слов**	**АЗ**		**ГМ**		**МШ**		**СТ**		**СА**	
			Б	**З**	**БМ**	**СМ**	**СБ**	**Х**	**Н**	**ПКР**	**Д**	**МС**
			70804	**79431**	**46608**	**50368**	**113592**	**91202**	**9936**	**4041**	**71134**	**48801**
АЗ	**Б**	**70804**										
	З	**79431**	**1.8158**									
ГМ	**БМ**	**46608**	1.3139	1.4810								
	СМ	**50368**	1.4288	1.7519	**0.8282**							
МШ	**СБ**	**113592**	4.7213	3.8220	4.0112	4.6244						
	Х	**91202**	5.5497	4.6478	4.8292	5.4612	**2.4375**					
СТ	**Н**	**9936**	3.0341	3.7945	2.6130	2.0497	6.4312	7.2691				
	ПКР	**4041**	4.0711	4.8315	3.4370	3.0868	6.6482	7.4816	**2.0557**			
СА	**Д**	**71134**	2.4625	3.1858	2.2538	2.1185	5.4750	6.3106	3.0951	2.0052		
	МС	**48801**	3.0353	3.7807	2.5715	2.6712	5.5688	6.0156	3.8667	2.6434	**1.1227**	

Для современной прозы оптимальное значение γ оказалось следующим

$$\gamma^{\mathrm{opt}} \in [1.1228;\ 1.3138).$$

Применять этот факт для выяснения метрической близости пары произведений T_1 и T_2 современной прозы необходимо следующим образом, см. [16-А, 44-А]:

– если $\rho(T_1,\ T_2) \leq 1.1228$, то T_1 и T_2 однородны;

– если $\rho(T_1,\ T_2) > 1.3138$, то T_1 и T_2 неоднородны;

– если $1.1228 < \gamma \leq 1.3138$, то ситуация – неопределенная.

И здесь закрашенные в табл. 2.15 жёлтым цветом ячейки (в данном случае их – 3) показывают нарушение сформулированной гипотезы для соответствующих пар произведений.

5. Вычисления по формуле (2.9) коэффициента эффективности π:

– для классической поэзии выдает значение π = 98%,

– для современной поэзии выдаёт значение π = 98%,

– для современной прозы выдаёт значение $\pi = 93\%$.

распознавания автора по цифровому портрету его произведений.

Полученные значения показывают, что распознавание автора текста по цифровому портрету (распределению частотности буквенных триграмм) для поэтических произведений (в сравнении с прозаическими) более успешно.

Результаты данного параграфа опубликованы в [46-А].

§ 2.1.6. Об идентификации автора текстового фрагмента на основе частотности символьных триграмм

На примере модельной коллекции таджикских литературных произведений изучается задача о возможности определения авторства фрагмента текста минимального размера, извлеченного из коллекции. Рассматривается модельная коллекция текстов таджикского языка, составленная из произведений классической поэзии и современной прозы на кириллической графике. Каждому произведению сопоставлен ЦП – распределение частотностей символьных триграмм.

В данном пункте, используя γ-классификатор §§ 1.3-1.4 и цифровой текстовый портрет § 2.1.5, характеризующий распределение частотности буквенных триграмм, приводится описание процесса идентификации авторов произведений таджикского текста. Отметим, что ранее аналогичный вопрос изучался именно для символьных (буквенных) униграмм, биграмм и триграмм с учетом пробела [10-А]. Существенным моментом в сравнении с нашим предыдущим исследованием § 2.1.5 является изучение вопроса о минимально допустимом размере текстового фрагмента, для которого удается получить удовлетворительный результат решения рассматриваемой задачи. Отметим также, что в сравнении с исследованием [10-А] выбора фрагментов из текстов, извлеченных из *«начала», «середины»* и *«конца»* произведений, решается задача по отдельности для поэзий и прозы.

Модельная коллекция текстов, на которой разворачивается наше исследование, та же самая, что и в § 2.1.1.

Обработка коллекционного материала происходила в 4 этапа.

Этап 1 состоял из выбора двух произведений различных авторов, каковыми оказались «Рустам ва Сӯҳроб» А. Фирдоуси и «Дохунда» С. Айни. Из каждого произведения извлекались по 9 случайных выборок текстовых фрагментов, размеры которых в словах и символах показаны в таблице 2.16.

Таблица 2.16. – Информация о размерах фрагментов в словах и символах

Номер фрагмента	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Число слов	7000	3500	1700	900	450	250	130	60	20
Число символов	40000	20000	10000	5000	2500	1200	600	300	100

Как видно из таблицы, размеры фрагментов уменьшаются от номера к номеру.

Замечание. Формальное деление числа символов на соответствующее ему число слов не будет означать среднюю длину слова, исчисляемую количеством букв, поскольку к символам помимо букв относятся также знаки препинания и арифметических операций, цифры, обозначения типа «№», «@», «$» и т.п.

Этап 2. Для каждого фрагмента выбранных произведений строится ЦП, который определяется распределением частотности буквенных триграмм, содержащихся в рассматриваемом фрагменте.

ЦП представляется в табличном виде:

$$\begin{array}{lllll} N: & 1 & 2 & \ldots & 42875 \\ P: & p_1 & p_2 & \ldots & p_{42875}, \end{array}$$

в котором первая строка – номера триграмм, расположенных в алфавитном порядке, а вторая – относительные частоты встречаемости буквенных триграмм в тексте $\boldsymbol{T}$, причём $\sum_{i=1}^{42875} p_i = 1$.

Этап 3. Вычисление расстояний $\rho(T_1, T_2)$ между ЦП 9 фрагментов $\boldsymbol{T_1}$ и 12 произведениями $\boldsymbol{T_2}$ рассматриваемой коллекции текстов. Соответствующие вычисления производились по формуле

$$\rho(T_1, T_2) = \sqrt{\frac{42875}{2}} \max_s \left| \sum_{i=1}^{s} p_i^{(1)} - p_i^{(2)} \right| \qquad (2.12)$$

где $p_i^{(1)}$ и $p_i^{(2)}$ – частоты встречаемости в фрагментах $\boldsymbol{T_1}$ и в произведениях $\boldsymbol{T_2}$ буквенных триграмм i (i = 1, … , 42875) и (s=1, … , 42875).

Этап 4. Определение автора текстового фрагмента производится методом ближайшего соседа [248, 249]. Сущность метода заключается в том, что классифицируемый фрагмент $\boldsymbol{T_1}$ объявляется принадлежащим тому автору, чьё произведение $\boldsymbol{T_2}$ в сравнении с другими произведениями оказывается наиболее «сходным» с фрагментом. Иными словами, рассматриваемые объекты являются ближайшими соседями, и расстояния между их ЦП минимальны в сравнении с прочими расстояниями.

Полученные результаты сведены в две группы таблиц соответственно с номерами 2.17.1, 2.17.2, 2.17.3 и 2.18.1, 2.18.2, 2.18.3. В ячейках таблиц представлены расстояния от 12 произведений модельной коллекции текстов до 9 фрагментов из романа С. Айни «Дохунда» (в 1-й группе) и до 9 фрагментов из поэмы А. Фирдоуси «Рустам ва Сӯҳроб» (во 2-й группе). В этих таблицах используются те же сокращения для имен авторов и названий произведений, что и в § 2.1.1.

В последующих таблицах первые 2 колонки указывают авторов и их произведения, а ячейки 9-и других колонок – значения расстояний фрагментов различных длин до соответствующих произведений, вычисленные по формуле (2.12). Отметим также, что в названиях таблиц используются фразы о фрагментах, извлеченных из ***«начала»***, ***«середины»*** и ***«конца»*** произведений, тем самым в определенном смысле подчеркивается случайный характер выбора фрагментов из текстов произведений.

Таблица 2.17.1. – Расстояния между ЦПП из коллекции текстов и фрагментами, извлеченными из ***«начала»*** романа С. Айни «Дохунда»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**P&C**	2.5436	3.2746	4.0778	4.9858	7.0728	8.3370	13.0604	13.0436	17.4607
	Б&М	2.8389	3.3457	4.2794	5.2013	6.7834	8.1658	12.9767	13.0843	16.4193
ЧР	**ММ1**	4.5573	4.9898	4.8864	6.6990	7.8859	10.0514	15.4750	14.9192	19.8855
	ММ2	4.9369	5.2784	5.2621	7.1493	8.5230	10.6333	15.9215	15.6541	20.3319
АС	**Д1**	3.7778	4.0955	5.7570	5.4593	9.5500	11.9414	17.3650	16.8092	21.7755
	Д2	3.2920	3.6975	5.4515	5.1465	9.1779	11.4169	16.8406	16.2847	21.2510
СТ	**Н**	4.6043	5.0820	6.7323	6.4761	10.5185	12.6640	17.6498	17.5569	22.0602
	ПКР	4.5403	4.5474	6.1939	6.0378	9.8875	12.0331	17.3618	17.1073	21.7723
СА	**О**	**2.4537**	**2.6193**	**2.7422**	**2.8953**	**6.2598**	**8.4703**	**15.0159**	**13.3952**	**19.1626**
	АД	**2.5855**	**3.3419**	**4.5192**	**5.1568**	**7.3261**	**9.9000**	**16.5161**	**14.4657**	**20.3867**
	Д	**1.3933**	**1.5331**	**2.4368**	**3.1220**	**6.0826**	**8.2455**	**14.6358**	**13.3522**	**18.5771**
	МС	**2.5920**	**2.6079**	**3.6626**	**4.1679**	**6.4691**	**8.7516**	**14.3550**	**13.6448**	**18.7595**

Закрашенные ячейки этой таблицы показывают, что из 5 фрагментов, взятых из «начала» романа С. Айни «Дохунда», 4 фрагмента оказались ближайшими соседями для самого произведения, 1 фрагмент – ближайшим соседом для «Одина». Кроме того, ещё 4 фрагмента (размерами в 1200, 600, 300 и 100) оказались ближайшими соседями с поэмами А. Фирдоуси «Рустам ва Сӯҳроб» и «Бежан бо Манижа».

Таблица 2.17.2. – Расстояния между ЦПП из коллекции текстов и фрагментами, извлеченными из ***«середины»*** романа С. Айни «Дохунда»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**P&C**	3.4190	4.3473	5.3576	7.6926	8.3520	8.6127	9.7670	11.4969	10.5352
	Б&М	3.2696	4.3573	5.3702	8.0313	8.4579	8.4677	9.2603	11.5266	10.4632
ЧР	**ММ1**	4.4524	6.5993	7.6724	9.2869	9.6367	10.9922	11.2117	14.0512	12.9939
	ММ2	4.4962	6.6099	7.7163	9.5829	9.6805	11.0361	11.6471	14.0950	13.0662
АС	**Д1**	4.0043	5.1870	5.9765	7.5675	7.9172	9.2728	9.3339	12.3317	11.3541
	Д2	3.8169	5.2556	6.1263	7.6704	8.0232	9.3757	9.4368	12.4347	11.4985
СТ	**Н**	5.4066	5.7630	4.6882	5.7247	6.1149	7.2221	7.0992	10.0105	10.5001
	ПКР	4.9815	5.3854	4.3806	5.9040	6.2538	7.6093	7.6704	10.6683	10.1116
СА	**О**	**2.6887**	**2.5860**	**3.5989**	**5.1946**	**6.4493**	**6.8998**	**8.2669**	**9.9588**	**8.9281**
	АД	**3.7605**	**3.8733**	**3.8071**	**5.2930**	**5.6427**	**7.0019**	**7.3483**	**10.0572**	**9.2996**
	Д	**1.3223**	**3.2110**	**4.1526**	**5.9626**	**6.6500**	**7.4723**	**8.3303**	**10.5313**	**9.5461**
	МС	**1.8747**	**4.0502**	**5.0631**	**6.5988**	**7.4730**	**8.3041**	**8.3679**	**11.3631**	**10.2414**

Как явствует из этой таблицы, для 8 фрагментов, взятых из «середины» романа «Дохунда», 1 фрагмент оказался ближайшим соседом для самого произведения, 1 фрагмент – ближайшим соседом для «Аҳмади Девбанд» и 6 фрагментов – ближайшими соседями для «Одина». Кроме того, всего лишь 1 фрагмент (размером в 600) оказался ближайшим соседом произведения С. Турсуна «Нисфирӯзӣ».

Таблица 2.17.3. – Расстояния между ЦПП из коллекции текстов и фрагментами, извлеченными из ***«конца»*** романа С. Айни «Дохунда»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**Р&С**	3.8841	3.9251	4.1163	3.9445	3.9522	3.9639	7.4304	14.6443	21.3940
	Б&М	3.6828	3.4150	3.6290	3.4911	3.8723	4.0097	7.3355	14.6856	21.8645
ЧР	**ММ1**	4.3022	4.1543	4.6408	4.3220	4.2693	5.6790	6.0951	14.2847	18.8217
	ММ2	4.8349	4.6017	4.8931	4.9006	4.6349	6.0447	6.7588	14.1887	18.7897
АС	**Д1**	5.7539	5.5672	5.7929	5.1681	5.3109	6.6321	7.9289	15.4841	21.3662
	Д2	4.9912	4.7692	4.9948	4.7723	4.7961	6.0891	7.8009	15.0878	21.0327
СТ	**Н**	6.9169	6.3962	7.0006	5.8497	6.3689	6.2776	9.9668	16.1272	22.7488
	ПКР	6.5221	6.0474	6.6827	6.0261	6.0511	5.9598	9.6490	16.4912	23.0388
СА	**О**	**2.2807**	**2.2140**	**2.5305**	**2.8207**	**3.5478**	**5.3861**	**7.3665**	**16.7962**	**23.7372**
	АД	**4.2205**	**3.6596**	**3.9958**	**3.8645**	**4.7742**	**6.2306**	**7.5450**	**17.9199**	**24.2198**
	Д	**2.1227**	**1.7827**	**2.1394**	**1.9463**	**2.9653**	**4.5498**	**6.0168**	**15.9467**	**22.7146**
	МС	**2.3353**	**2.3652**	**2.5097**	**2.3758**	**3.0477**	**4.8559**	**6.5671**	**16.3018**	**22.7403**

Для фрагментов из «конца» романа «Дохунда» (размерами не менее 2500 и 600 символов) основной результат – тот же, что и в 2-х предыдущих случаях: ближайшими для них соседями служат только произведения С. Айни. Кроме того, всего лишь 1 фрагмент (размером в 1200) оказался ближайшим соседом поэмы А. Фирдоуси «Рустам ва Сӯҳроб». Интересно, что 2 других фрагмента (размерами в 300 и 100) оказались ближайшими соседями с поэмой Дж. Руми «Маснавии Маънавӣ, Дафтари 2».

Таблица 2.18.1. – Расстояния между ЦПП из коллекции текстов и фрагментами, извлеченными из ***«начала»*** поэмы А. Фирдоуси «Рустам ва Сӯҳроб»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**Р&С**	**1.2145**	**1.9171**	**2.3541**	**3.1874**	**4.6208**	**7.1645**	**10.3483**	**12.8943**	**30.1004**
	Б&М	**1.2658**	**1.9665**	**2.1085**	**2.4471**	**4.7228**	**6.5503**	**9.6747**	**12.6821**	**30.6926**
ЧР	**ММ1**	3.7744	3.9194	4.8797	5.5956	6.9713	8.9408	12.2689	14.8149	28.3623
	ММ2	4.6926	4.8139	5.7761	6.3669	7.2809	8.9293	12.3052	14.8512	28.3735
АС	**Д1**	5.5402	5.6765	6.4536	7.3554	8.7512	10.6483	14.0031	16.5491	30.2733
	Д2	5.0168	5.1531	5.9928	6.7576	8.1467	10.0106	13.3693	15.9153	30.1712
СТ	**Н**	6.0932	6.3242	7.1044	7.9390	9.2566	10.9242	14.3001	16.8460	30.7027
	ПКР	5.7606	6.0879	6.7812	7.5424	8.8653	10.5016	13.8775	16.4235	31.1204
СА	**О**	3.7822	4.0725	4.5350	4.9071	7.3910	8.5622	11.9381	14.4841	32.6792
	АД	4.8325	5.2168	5.3870	6.2469	8.3657	10.1112	13.3101	15.6656	32.6272
	Д	2.7935	2.9915	3.4553	4.2986	6.4322	7.9085	11.2843	13.8303	31.3211
	МС	3.2544	3.1030	3.6611	4.3830	6.6346	7.9413	11.1920	13.7380	31.4827

В этой (табл. 2.18.1) и двух следующих таблицах (табл. 2.18.2, 2.18.3) 9

фрагментов выбираются из поэмы А. Фирдоуси «Рустам ва Сӯҳроб». Закрашенные ячейки этой таблицы показывают, что из 8 фрагментов, взятых из «начала» поэмы А. Фирдоуси «Рустам ва Сӯҳроб», 3 оказались ближайшими соседями для самой поэмы, а 5 – ближайшими соседями для «Бежан бо Манижа». Кроме того, всего лишь 1 фрагмент (размером в 100) оказался ближайшим соседом поэмы Дж. Руми «Маснавии Маънавӣ, Дафтари 1».

Таблица 2.18.2. – Расстояния между ЦПП из коллекции текстов и фрагментами, извлеченными из *«середины»* поэмы А. Фирдоуси «Рустам ва Сӯҳроб»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**Р&С**	**0.4846**	**1.3781**	**2.1632**	**5.4587**	**5.1008**	**4.8602**	**6.4552**	**13.7455**	**8.8868**
	Б&М	**1.4404**	**1.3466**	**2.1985**	**5.7512**	**4.8642**	**4.5393**	**6.1343**	**13.3244**	**9.0973**
ЧР	**ММ1**	3.2171	3.6585	4.6331	7.5103	7.7921	7.3386	8.9336	15.0123	8.2637
	ММ2	3.7341	4.1424	4.9218	7.9591	8.2353	7.7045	9.2995	15.1765	8.5181
АС	**Д1**	4.4891	4.8707	4.6132	5.2929	5.5714	5.7995	6.7155	16.6201	7.9693
	Д2	4.0016	4.4098	4.2239	5.4881	5.7656	5.3905	7.0004	16.0956	7.4449
СТ	**Н**	5.0472	5.4121	5.1365	4.3385	4.6293	6.0845	8.6883	16.9048	9.0352
	ПКР	4.7104	5.1177	4.8828	4.5224	4.8984	6.6804	8.8903	17.2545	9.7001
СА	**О**	3.8836	3.3571	4.3277	5.0310	5.2625	4.0714	5.1297	16.5983	7.7348
	АД	4.5072	3.9214	3.5965	3.2339	3.5559	5.3045	7.9951	17.9817	9.2228
	Д	2.5160	2.2287	3.5185	4.5523	4.8965	4.0003	5.5953	15.9816	7.9074
	МС	4.0026	3.9737	5.2536	6.2873	6.6419	5.3630	5.8594	16.1091	8.4788

Закрашенные ячейки этой таблицы показывают, что из 4 фрагментов, взятых из «середины» поэмы А. Фирдоуси «Рустам ва Сӯҳроб», 2 оказались ближайшими соседями для самой поэмы, а 2 – ближайшими соседями для «Бежан бо Манижа». Кроме того, всего лишь 1 фрагмент (размером в 100) оказался ближайшим соседом поэмы А. Суруш «Дафтари 2». Интересно, что 4 других фрагмента (размерами в 5000, 2500, 1200 и 600) оказались ближайшими соседями произведений С. Айни «Аҳмади Девбанд», «Дохунда» и «Одина».

Таблица 2.18.3. – Расстояния между ЦПП из коллекции текстов и фрагментами, извлеченными из ***«конца»*** поэмы А. Фирдоуси «Рустам ва Сӯҳроб»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**Р&С**	**1.3895**	**1.5794**	**2.6426**	**2.6984**	**3.2431**	**3.9399**	**5.5064**	**5.6156**	**39.1433**
	Б&М	**2.7287**	**2.6193**	**3.1115**	**2.5858**	**3.4727**	**4.3456**	**6.2583**	**5.0843**	**39.0127**
ЧР	**ММ1**	3.5675	3.5368	5.1599	4.3623	4.7371	5.9534	6.1731	7.6321	39.8656
	ММ2	3.3789	3.8119	5.4557	5.2581	4.9102	6.0461	5.9077	7.8169	39.9809
АС	**Д1**	4.0025	4.1121	4.8656	5.9392	5.4671	5.6012	6.1705	9.7727	41.6974
	Д2	4.1156	3.8787	4.6548	5.4783	5.2563	5.3899	6.4522	8.9747	41.3468
СТ	**Н**	4.5439	4.6536	5.0784	6.4915	5.0801	5.1285	6.3449	10.0200	41.3853
	ПКР	4.2481	4.3555	4.8028	6.2114	5.1282	5.2898	6.0085	9.6504	41.4478
СА	**О**	5.0091	5.0026	5.3300	4.8818	5.5306	6.5280	8.3643	7.1241	40.8085
	АД	3.9205	3.3670	3.2374	5.7966	5.3760	5.2459	6.4543	7.3038	42.5603
	Д	3.9141	3.7115	4.1312	3.8565	5.1128	6.2047	7.5880	6.3478	40.5254
	МС	4.6003	4.1281	4.3335	3.8912	4.5663	5.5200	7.0682	6.2548	40.1321

Как явствует из этой таблицы, для 9 фрагментов, взятых из «конца» поэмы А. Фирдоуси «Рустам ва Сӯҳроб», 6 оказались ближайшими соседями для самой поэмы, а 3 – ближайшими соседями для «Бежан бо Манижа».

Заключение. Таким образом, результаты, представленные в таблицах, показывают, что ближайшими соседями по отношению к выбранным фрагментам являются в основном произведения именно того автора, из произведения которого извлекались сами фрагменты. В иной интерпретации это значит, что методом ближайшего соседа путем вычисления расстояний по формуле (2.12) представляется возможным установить авторство достаточно малого кусочка литературного произведения, причём для поэтических произведений (в сравнении с прозаическими) более успешно**.**

Для художественных текстов можно предложить оценку эффективности применяемого метода, опираясь на вполне естественную гипотезу, согласно которой фрагмент, извлекаемый из какого-либо произведения, должен быть «однородным» с любыми произведениями одного и того же автора. На языке «расстояний» этому соответствует утверждение о том, что ближайшими соседями искомого фрагмента являются, прежде всего, произведения его автора.

Для прозаического произведения, по данным таблицы 2.17.1, метод ближайшего соседа безошибочно определяет автора фрагментов, состоящих из не менее 2500 символов.

По данным таблицы 2.17.2, метод безошибочно определяет автора 8 фрагментов из 9 и для 1 фрагмента размером 600 символов допускает ошибку.

По данным таблицы 2.17.3, метод ближайшего соседа безошибочно определяет автора 6 фрагментов из 9 и для 3-х фрагментов размерами 1200, 300 и 100 символов допускает ошибку.

Для поэтического произведения, по данным таблицы 2.18.1, метод ближайшего соседа безошибочно определяет автора 8 фрагментов из 9 и для 1 фрагмента размером 100 символов допускает ошибку.

По данным таблицы 2.18.2, метод безошибочно определяет автора 4 фрагментов из 9 и для 5-и фрагментов размерами 5000, 2500, 1200, 600 и 100 символов допускает ошибку.

По данным таблицы 2.18.3, метод ближайшего соседа безошибочно определяет автора фрагментов, состоящих из не менее 100 символов.

Результаты данного параграфа опубликованы в [29-А].

§ 2.2. Применение частотности слогов

§ 2.2.1. Об определение автора текста на основе частотности слогов

Решается задача распознавания авторов произведений по отдельности для классической и современной поэзий, а также современной прозы. Произведениям

сопоставляется ЦП, характеризуемый распределением в них частотности слогов. Устанавливается эффективность применения γ-классификатора для идентификации авторов произведений. Сконструированы ЦП и метрическое пространство произведений. В предположении уникальности авторского творчества устанавливаются пороговые значения метрики, на основе которых определяются классы «однородных» произведений.

В этом пункте мы продолжаем тестирование количественных описаний текстов, начатое в работах [1-А-10-А], на предмет их пригодности для идентификации авторов произведений. В качестве таковых в § 2.1.1 рассматривались частотности букв таджикского алфавита (униграммы), в §§ 2.1.3-2.1.6 – буквенных биграмм и триграмм, в [283] – набора из пяти натуральных единиц измерения текста, в [256, 257] – частотности длин слов и знаков препинаний, в [14-А] – частотности слогов, в [258] – частотности длин предложений. Существенным моментом в сравнении с нашим предыдущим исследованием [14-А] является изучение вопроса о распознавании авторов текстов, относящихся к произведениям классической и современной поэзии, а также к современной прозе. Следуя [14-А], будем называть *цифровым портретом текста* распределение в нём частотности слогов. В настоящем пункте изучается вопрос об эффективности применения такого показателя для распознавания авторов поэтических и прозаических произведений.

1. Обработка статистического материала включала в себя 4 этапа.

Этап 1. Создание компьютерной программы и вычисление с её помощью ЦП произведений – распределений частотности слогов по отдельности для всех текстов, упомянутых в § 2.1.1.

Этап 2. Создание компьютерной программы и вычисление с её помощью парных расстояний между ЦП произведений по формуле, предложенной в § 1.3.

Этап 3. Настройка γ-классификатора. Существо настройки заключается в том, чтобы определить такое значение вещественного параметра γ, при котором достигается максимальное значение критерия «γ-однородности» произведений, см. § 1.4.

Этап 4. Установление эффективности применения настроенного γ-классификатора для распознавания авторов произведений.

На этапе 1 цифровые портреты произведений представляются в табличном виде:

$$\begin{array}{lccccc} \bar{N}: & 1 & 2 & . & . & . \quad m \\ P: & p_1 & p_2 & . & . & . \quad p_m, \end{array}$$

где первая строка – список слогов; m – общее число слогов; вторая строка – частоты p_i встречаемости в пределах произведений слогов $i(i = 1,2, \dots, m)$,

причём

$$\sum_{i=1}^{m} p_k = 1.$$

На этапе 2 вычисления расстояний $\rho(T_1, T_2)$ между текстами T_1 и T_2 производились по формуле T_1 и T_2

$$\rho(T_1, T_2) = \sqrt{\frac{m}{2}} \max_s \left| \sum_{k=1}^{s} (p_k^{(1)} - p_k^{(2)}) \right|,$$

в которой m (= 4699) – количество слогов; $p_k^{(1)}$ и $p_k^{(2)}$ – частоты встречаемости в текстах T_1 и T_2 суммарные количества слогов k, k = 1, …, m и $(\mathrm{s} = 1, \cdots, \mathrm{m})$.

Результаты вычислений показаны в таблицах 2.19-2.21.

На этапе 3 качество классификатора при фиксированном γ оценивается величиной π, вычисляемой по формуле

$$\pi = 1\text{-}\ \tau/\mathrm{L}\,, \tag{2.13}$$

где L (= 45) – суммарное число взаимных расстояний между 10 текстами исходной коллекции; $\tau = \tau(\gamma)$ – число нарушений неравенств

$$\rho(T_1, T_2) \le \gamma, \tag{2.14}$$

$$\rho(T_1, T_2) > \gamma. \tag{2.15}$$

Первое проверяется на 5 парах текстов одних и тех же авторов, второе – на 40 парах текстов различных авторов.

На этапе 4 производится настройка γ-классификатора на основе вполне естественной гипотезы о том, что произведения одного автора «однородны», а разных авторов «неоднородны». На языке ЦП, характеризующих распределение частотности слогов 10 пар произведений, определение γ сводится к отысканию такого его значения, при котором общее число τ нарушений неравенств (2.14), (2.15) по отдельности на текстах 3-х модельных коллекций становится минимальным. Для нахождения таких γ используется алгоритм, предложенный в § 1.4.

2. Результаты вычислений расстояний между 10 произведениями классической поэзии представлены в табл. 2.19.

Таблица 2.19. – Расстояния между произведениями ***классической поэзии***

Автор (Произ.)		**Число слов**	**АР**		**АФ**		**СШ**		**ХШ**		**ЧР**	
			АП	**К**	**Р&С**	**Б&М**	**Ғ1**	**Ғ2**	**Ғ1**	**Ғ2**	**ММ1**	**ММ2**
			2248	**5054**	**16355**	**14799**	**16261**	**13001**	**33724**	**28923**	**48713**	**41661**
АР	**АП**	**2248**										
	К	**5054**	**0.4775**									
АФ	**Р&С**	**16355**	2.3558	2.2488								

Автор (Произ.)		Число слов	АР		АФ		СШ		ҲШ		ЧР	
			АП	Қ	Р&С	Б&М	Ғ1	Ғ2	Ғ1	Ғ2	ММ1	ММ2
			2248	5054	16355	14799	16261	13001	33724	28923	48713	41661
	Б&М	14799	2.2524	2.2457	0.3677							
СШ	Ғ1	16261	2.0731	2.1144	3.3916	3.4490						
	Ғ2	13001	1.6014	1.6323	3.0318	3.1053	1.1241					
ҲШ	Ғ1	33724	1.8949	1.8559	3.4425	3.5437	1.5606	0.7579				
	Ғ2	28923	1.3305	1.3044	2.9259	2.9793	1.8785	0.8233	0.6425			
ЧР	ММ1	48713	2.3194	2.2617	1.4781	1.3371	3.6806	3.4156	3.8837	3.3723		
	ММ2	41661	2.5024	2.4182	1.5273	1.4660	3.7688	3.5658	4.0338	3.5225	0.3800	

Для классической поэзии оптимальное значение γ оказалось следующим

$$\gamma^{opt} \in [0.6426;\ 0.7578).$$

Применять этот факт для выяснения метрической близости пары произведений T_1 и T_2 необходимо следующим образом, см. [16-А, 44-А]:

– если $\rho(T_1,\ T_2) \leq 0.6426$, то T_1 и T_2 однородны;

– если $\rho(T_1,\ T_2) > 0.7578$, то T_1 и T_2 неоднородны;

– если $0.6426 < \gamma \leq 0.7578$, то ситуация – неопределенная.

Из данных таблицы следует, что только одно расстояние, именно 1.1241 соответственно между ЦП двух произведений С. Шерозӣ «Ғазалиёт қисми 1» и «Ғазалиёт қисми 2» нарушает сформулированную гипотезу. Эти пары согласно (2.15) утверждают неоднородность указанных двух произведений С. Шерозӣ, хотя принадлежат одному автору.

Желтым цветом в таблице 2.19 отмечен 1 случай нарушения гипотезы однородности.

3. Результаты вычислений расстояний между 10 произведениями современной поэзии представлены в табл. 2.20.

Таблица 2.20. – Расстояния между произведениями в ***современной поэзии***

Автор (Произ.)		Число слов	АС		АШ		ГС		ИФ		МТ	
			Д1	Д2	БТ	ШР	О	Ш	101Г	МГМ	ҚҲ	ҲА
			7890	9322	32036	12810	12103	51434	9841	41217	8463	6118
АС	Д1	7890										
	Д2	9322	0.5670									
АШ	БТ	32036	3.4208	3.4509								
	ШР	12810	3.1261	3.1562	0.6288							
ГС	О	12103	1.7043	1.7261	2.1103	1.8441						
	Ш	51434	1.6682	1.5333	2.0954	1.8256	0.4486					
ИФ	101Г	9841	2.2011	2.3188	1.6492	1.8497	1.7560	1.8058				
	МГМ	41217	1.5945	1.6246	1.8598	1.5649	1.3480	1.3661	0.9455			
МТ	ҚҲ	8463	1.7543	1.7910	1.9522	1.6708	1.3577	1.4517	1.2415	0.8950		
	ҲА	6118	1.3349	1.3651	2.4054	2.1747	1.8997	1.6859	1.7645	1.0887	1.2949	

Для современной поэзии оптимальное значение γ оказалось следующим

$$\gamma^{opt} \in [0.9456;\ 1.0886).$$

Применять этот факт для выяснения метрической близости пары

произведений T_1 и T_2 необходимо следующим образом, см. [16-А, 44-А]:

– если $\rho(T_1, T_2) \leq 0.9456$, то T_1 и T_2 однородны;

– если $\rho(T_1, T_2) > 1.0886$, то T_1 и T_2 неоднородны;

– если $0.9456 < \gamma \leq 1.0886$, то ситуация – неопределенная.

И здесь в табл. 2.20 закрашенные жёлтым цветом ячейки (в данном случае их – 2) показывают нарушение сформулированной гипотезы для соответствующих пар произведений.

4. Результаты вычислений расстояний между 10 произведениями современной прозы представлены в табл. 2.21.

Таблица 2.21. – Расстояния между произведениями в ***современной прозе***

Автор (Произ.)		Число слов	АЗ		ГМ		МШ		СТ		СА	
			Б	З	БМ	СМ	СБ	Х	Н	ПКР	Д	МС
			70804	79431	46608	50368	113592	91202	9936	4041	71134	48801
АЗ	Б	70804										
	З	79431	**0.4795**									
ГМ	БМ	46608	0.9196	0.8942								
	СМ	50368	0.8686	0.8078	**0.2167**							
МШ	СБ	113592	1.9568	1.8291	1.1456	1.2570						
	Х	91202	2.3137	1.9916	1.8129	1.9727	**0.9183**					
СТ	Н	9936	1.7596	1.9018	2.3949	2.2858	3.4761	3.5162				
	ПКР	4041	0.9151	1.1549	1.6573	1.6286	2.5006	2.6143	**1.4232**			
СА	Д	71134	0.9905	0.6859	0.7337	0.5378	1.7286	1.9442	2.2282	1.3286		
	МС	48801	2.0143	1.6973	1.1844	1.2539	1.6432	1.5542	3.1832	2.3349	**1.2557**	

Для современной прозы оптимальное значение γ оказалось следующим

$$\gamma^{opt} \in [0.4796; 0.5377).$$

Применять этот факт для выяснения метрической близости пары произведений T_1 и T_2 современной прозы необходимо следующим образом, см. [16-А, 44-А]:

– если $\rho(T_1, T_2) \leq 0.4796$, то T_1 и T_2 однородны;

– если $\rho(T_1, T_2) > 0.5377$, то T_1 и T_2 неоднородны;

– если $0.4796 < \gamma \leq 0.5377$, то ситуация – неопределенная.

И здесь закрашенные в табл. 2.21 жёлтым цветом ячейки (в данном случае их – 3) показывают нарушение сформулированной гипотезы для соответствующих пар произведений.

5. Вычисления по формуле (2.13) коэффициента эффективности π:

– для классической поэзии выдает значение π = 98%,

– для современной поэзии выдаёт значение π = 96%,

– для современной прозы выдаёт значение π = 93%.

распознавания автора по цифровому портрету его произведений.

Полученные значения показывают, что распознавание автора текста по цифровому портрету (распределению частотности слогов) для поэтических

произведений (в сравнении с прозаическими) более успешно.

Результаты данного параграфа опубликованы в [20-А, 54-А].

§ 2.2.2. О распознавании автора текстового фрагмента на основе частотности слогов

На примере модельной коллекции таджикских литературных произведений изучается задача о возможности определения авторства фрагмента текста минимального размера, извлеченного из коллекции. Рассматривается модельная коллекция текстов таджикского языка, составленная из произведений классической поэзии и современной прозы на кириллической графике. Каждому произведению сопоставлен ЦП – распределение частотностей слогов.

В настоящем пункте, используя γ-классификатор §§ 1.3-1.4 и цифровой текстовый портрет § 2.2.1, характеризующий распределение частотности слогов, приводится описание процесса идентификации авторов произведений таджикского текста. Отметим, что ранее аналогичный вопрос изучался для символьных (буквенных) униграмм, биграмм и триграмм с учетом пробела [10-А]. Существенным моментом в сравнении с нашим предыдущим исследованием § 2.2.1 является изучение вопроса о минимально допустимом размере текстового фрагмента, для которого удается получить удовлетворительный результат решения рассматриваемой задачи. Отметим также, что в сравнении с исследованием [10-А] выбора фрагментов из текстов, извлеченных из *«начала», «середины»* и *«конца»* произведений, решается задача по отдельности для поэзий и прозы.

Модельная коллекция текстов, на которой разворачивается наше исследование, та же самая, что и в § 2.1.1.

Обработка коллекционного материала происходила в 4 этапа.

Этап 1 состоял из выбора двух произведений различных авторов, каковыми оказались «Рустам ва Сӯҳроб» А. Фирдоуси и «Дохунда» С. Айни. Из каждого произведения извлекались по 9 случайных выборок текстовых фрагментов, размеры которых в словах и символах показаны в таблице 2.22.

Таблица 2.22. – Информация о размерах фрагментов в словах и символах

Номер фрагмента	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Число слов	7000	3500	1700	900	450	250	130	60	20
Число символов	40000	20000	10000	5000	2500	1200	600	300	100

Как видно из таблицы, размеры фрагментов уменьшаются от номера к номеру.

Замечание. Формальное деление числа символов на соответствующее ему число слов не будет означать среднюю длину слова, исчисляемую количеством

букв, поскольку к символам помимо букв относятся также знаки препинания и арифметических операций, цифры, обозначения типа «№», «@», «$» и т.п.

Этап 2. Для каждого фрагмента выбранных произведений строится ЦП, который определяется распределением частотности слогов, содержащихся в рассматриваемом фрагменте.

ЦП представляется в табличном виде:

$$\begin{array}{ll} N: & 1 \quad 2 \quad \ldots \quad 4699 \\ P: & p_1 \quad p_2 \quad \ldots \quad p_{4699}, \end{array}$$

в котором первая строка – номера слогов, расположенных в алфавитном порядке, а вторая – относительные частоты встречаемости слогов в тексте $\boldsymbol{T}$, причём $\sum_{i=1}^{4699} p_i = 1$.

Этап 3. Вычисление расстояний $\rho(T_1, T_2)$ между ЦП 9 фрагментов $\boldsymbol{T_1}$ и 12 произведениями $\boldsymbol{T_2}$ рассматриваемой коллекции текстов. Соответствующие вычисления производились по формуле

$$\rho(T_1, T_2) = \sqrt{\frac{4699}{2}} \max_s \left| \sum_{i=1}^{s} p_i^{(1)} - p_i^{(2)} \right| \qquad (2.16)$$

где $p_i^{(1)}$ и $p_i^{(2)}$ – частоты встречаемости в фрагментах $\boldsymbol{T_1}$ и в произведениях $\boldsymbol{T_2}$ слоги i (i = 1, … , 4699) и (s=1, … , 4699).

Этап 4. Определение автора текстового фрагмента производится методом ближайшего соседа [248, 249]. Сущность метода заключается в том, что классифицируемый фрагмент $\boldsymbol{T_1}$ объявляется принадлежащим тому автору, чьё произведение $\boldsymbol{T_2}$ в сравнении с другими произведениями оказывается наиболее «сходным» с фрагментом. Иными словами, рассматриваемые объекты являются ближайшими соседями, и расстояния между их ЦП минимальны в сравнении с прочими расстояниями.

Полученные результаты сведены в две группы таблиц соответственно с номерами 2.23.1, 2.23.2, 2.23.3 и 2.24.1, 2.24.2, 2.24.3. В ячейках таблиц представлены расстояния от 12 произведений модельной коллекции текстов до 9 фрагментов из романа С. Айни «Дохунда» (в 1-й группе) и до 9 фрагментов из поэмы А. Фирдоуси «Рустам ва Сӯҳроб» (во 2-й группе). В этих таблицах используются те же сокращения для имен авторов и названий произведений, что и в § 2.1.1.

В последующих таблицах первые 2 колонки указывают авторов и их

произведения, а ячейки 9-и других колонок – значения расстояний фрагментов различных длин до соответствующих произведений, вычисленные по формуле (2.16). Отметим также, что в названиях таблиц используются фразы о фрагментах, извлеченных из ***«начала»***, ***«середины»*** и ***«конца»*** произведений, тем самым в определенном смысле подчеркивается случайный характер выбора фрагментов из текстов произведений.

Таблица 2.23.1. – Расстояния между ЦП произведений из коллекции текстов и фрагментами, извлеченными из ***«начала»*** романа С. Айни «Дохунда»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**Р&С**	6.3695	7.1232	7.8197	8.5387	8.9456	9.5991	10.9118	12.3623	11.5601
	Б&М	6.4239	7.2247	7.9387	8.5930	9.0186	9.6535	10.9661	12.4167	11.6927
ЧР	**ММ1**	6.4452	7.3103	7.8154	8.6790	9.2783	9.7618	10.9391	12.5296	11.5694
	ММ2	6.5154	7.3806	7.9142	8.7825	9.3406	9.8093	10.9867	12.5772	11.6682
АС	**Д1**	2.3202	2.9658	3.4091	4.0021	4.3421	5.1231	6.3582	7.8087	8.4455
	Д2	1.9626	2.4899	2.9332	3.6074	4.1482	4.7111	5.7935	7.2440	8.1312
СТ	**Н**	2.9121	3.4157	4.1643	5.2362	5.7770	6.0344	7.1939	8.7168	8.0384
	ПКР	2.0974	2.7899	3.0875	4.1579	4.5044	4.8475	6.0184	7.4689	7.7576
СА	**О**	**1.1152**	**0.7910**	**1.2667**	**1.9913**	**2.5321**	**3.1241**	**4.2753**	**5.6096**	**6.7269**
	АД	**0.8216**	**1.3748**	**2.0268**	**2.9149**	**3.5208**	**3.9658**	**5.0950**	**6.6855**	**6.9733**
	Д	**0.8846**	**1.7220**	**2.3934**	**3.4019**	**3.9427**	**4.2456**	**5.3807**	**6.9712**	**6.9269**
	МС	**0.9365**	**1.0393**	**1.7243**	**2.7843**	**3.3251**	**3.7088**	**4.8193**	**6.4097**	**6.1256**

Из таблицы следует, что все фрагменты из «Дохунда» размерами от 40000 и вплоть до 100 символов являются ближайшими соседями произведений С. Айни и никого другого, см. закрашенные ячейки.

Таблица 2.23.2. Расстояния между ЦП произведений из коллекции текстов и фрагментами, извлеченными из ***«середины»*** романа С. Айни «Дохунда»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**Р&С**	5.4151	5.2286	4.9224	4.6715	3.8927	4.7503	5.7426	8.2287	8.4377
	Б&М	5.4251	5.3087	4.9888	4.6816	3.9027	4.8237	5.5900	8.1834	8.2912
ЧР	**ММ1**	5.6409	5.5062	5.1923	4.8974	4.3965	5.4529	5.7981	8.8929	9.1020
	ММ2	5.6885	5.5471	5.2387	4.9449	4.4019	5.4583	6.0765	8.9456	9.1546
АС	**Д1**	1.9196	1.7086	2.1147	2.1103	2.8131	2.1499	2.9670	5.4137	5.6227
	Д2	1.5420	1.4204	1.7071	1.9796	2.5039	2.0440	2.6350	5.1337	5.3427
СТ	**Н**	2.0588	1.8502	1.3530	1.2512	1.1586	1.6728	3.1173	4.3233	4.5323
	ПКР	1.4037	1.1499	0.8151	1.3859	2.3372	1.6841	2.1231	3.7533	3.9623
СА	**О**	**1.9574**	**2.1686**	**2.8577**	**3.4253**	**4.7296**	**3.8661**	**2.9399**	**3.0876**	**5.9069**
	АД	**1.5625**	**1.6925**	**2.2959**	**2.7754**	**3.9701**	**3.3468**	**2.5537**	**3.1997**	**5.2632**
	Д	**0.7984**	**0.9815**	**1.4150**	**1.9568**	**3.2494**	**2.4757**	**2.4331**	**3.7542**	**4.4790**
	МС	**1.6037**	**1.6340**	**2.5528**	**3.0610**	**4.2653**	**3.6160**	**3.3452**	**2.8264**	**5.1587**

Закрашенные ячейки этой таблицы показывают, что из 3 фрагментов, взятых из «середины» романа С. Айни «Дохунда», 2 фрагмента оказались ближайшими соседями для самого произведения, 1 фрагмент – ближайшим соседом для «Марги судхӯр». Кроме того, ещё 6 фрагментов (размерами в 10000, 5000, 2500, 1200, 600

и 100) оказались ближайшими соседями с произведениями С. Турсуна «Повести Камони Рустам» и «Нисфирӯзӣ».

Таблица 2.23.3. Расстояния между ЦП произведений из коллекции текстов и фрагментами, извлеченными из ***«конца»*** романа С. Айни «Дохунда»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**P&C**	6.1437	5.8309	6.2934	6.5214	6.9118	6.5426	7.0409	8.6174	11.4686
	Б&М	6.2052	5.8853	6.3477	6.5757	7.0421	6.6779	7.0953	8.6908	11.5421
ЧР	**ММ1**	6.2693	5.9986	6.5843	6.9070	7.3396	7.2473	7.4445	8.9098	11.7611
	ММ2	6.3545	6.0450	6.6248	6.9474	7.3800	7.2878	7.4849	8.9563	11.8075
АС	**Д1**	2.2821	2.0076	2.5787	2.7677	3.2732	3.1395	3.2794	5.5803	6.4744
	Д2	2.0410	1.7555	2.2108	2.4654	2.9952	2.9282	2.9422	5.2830	6.2646
СТ	**Н**	3.1529	2.7242	3.1965	3.2472	3.2939	2.6328	3.4611	5.4268	7.9179
	ПКР	2.0621	1.8191	2.0092	2.1619	2.6360	2.2688	2.1885	4.4050	6.6501
СА	**О**	**0.9796**	**1.2635**	**1.0988**	**1.1737**	**1.4296**	**2.1007**	**3.0433**	**3.0145**	**4.7209**
	АД	**0.6132**	**0.8842**	**0.9544**	**1.1947**	**1.8241**	**1.7571**	**2.3806**	**3.0897**	**5.8838**
	Д	**1.0959**	**0.5663**	**1.0056**	**1.2866**	**1.9260**	**1.8295**	**1.7516**	**3.8533**	**6.1713**
	МС	**0.7632**	**1.0035**	**0.9358**	**1.0870**	**0.9119**	**1.9101**	**2.8528**	**3.5125**	**5.6282**

Как явствует из этой таблицы, для 9 фрагментов, взятых из «конца» романа «Дохунда», 2 фрагмент оказались ближайшими соседями для самого произведения, а 7 других фрагментов оказались ближайшими соседями произведений С. Айни «Аҳмади Девбанд», «Марги судхӯр» и «Одина».

Таблица 2.24.1. Расстояния между ЦП произведений из коллекции текстов и фрагментами, извлеченными из ***«начала»*** поэмы А. Фирдоуси «Рустам ва Сӯҳроб»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**P&C**	**0.2568**	**0.5873**	**0.6320**	**1.2767**	**2.3756**	**1.8599**	**2.7681**	**4.6335**	**4.6573**
	Б&М	**0.5129**	**0.6123**	**0.7846**	**1.2638**	**2.4455**	**1.8413**	**2.7664**	**4.6536**	**4.6949**
ЧР	**ММ1**	1.3615	1.4439	1.0930	1.2149	2.3840	1.8076	2.0700	3.8537	5.1531
	ММ2	1.4433	1.5256	1.1617	1.0831	2.1067	1.8860	2.3424	3.8164	5.3754
АС	**Д1**	5.6685	5.7833	5.5388	6.0878	6.7886	6.1538	6.2823	7.8924	6.5124
	Д2	5.6986	5.8134	5.5689	6.2470	7.0975	6.3851	6.6478	8.4593	7.0296
СТ	**Н**	4.2159	4.4756	4.5697	5.1446	5.9619	5.8866	6.7059	8.4870	7.5157
	ПКР	4.9935	5.2866	5.2504	6.0221	6.8787	6.6642	7.3866	9.1168	8.1266
СА	**О**	6.8716	7.1266	6.9886	7.8813	8.6764	7.7740	8.4522	10.1271	8.9246
	АД	6.0721	6.3687	6.2307	7.1235	7.8511	7.4173	8.1372	9.9055	8.8698
	Д	5.6660	5.9626	5.8246	6.7173	7.4558	6.9419	7.5274	9.2456	8.2316
	МС	6.3488	6.6454	6.5074	7.4002	8.1571	7.8785	8.5027	10.2120	9.1457

В этой (табл. 2.24.1) и двух следующих таблицах (табл. 2.24.2, 2.24.3) 9 фрагментов выбираются из поэмы А. Фирдоуси «Рустам ва Сӯҳроб». Закрашенные ячейки этой таблицы показывают, что из 4 фрагментов, взятых из «начала» поэмы А. Фирдоуси «Рустам ва Сӯҳроб», все 4 оказались ближайшими соседями для самой поэмы. Кроме того, 5 других фрагментов оказались ближайшими соседями с поэмами Дж. Руми «Маснавии Маънавӣ, Дафтари 1» и «Маснавии Маънавӣ, Дафтари 2».

Таблица 2.24.2. Расстояния между ЦП произведений из коллекции текстов и фрагментами, извлеченными из *«**середины**»* поэмы А. Фирдоуси «Рустам ва Сӯҳроб»

Авторы (произв.)		Длины фрагментов (в символах)								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**P&C**	**0.2128**	**0.5043**	**0.8562**	**1.2535**	**1.0103**	**1.1794**	**1.8677**	**5.3677**	**6.4868**
	Б&М	**0.4290**	**0.6642**	**1.0680**	**1.5127**	**1.1845**	**1.0040**	**1.7125**	**5.5454**	**6.6645**
ЧР	**ММ1**	1.5424	1.3574	1.5186	1.8470	1.5543	1.5629	1.4353	5.2518	6.3340
	ММ2	1.5946	1.4196	1.5604	2.0053	1.6695	1.7777	1.5134	4.9373	5.9742
АС	**Д1**	5.5358	5.7445	5.4985	5.1773	5.9143	5.7733	6.2468	9.7909	10.9100
	Д2	5.5659	5.7746	5.5286	5.2074	5.9444	5.8192	6.2221	9.6078	10.7269
СТ	**Н**	4.0253	4.2250	4.2165	4.5179	4.5931	4.5705	4.6292	8.2851	9.4042
	ПКР	4.8467	4.9772	4.9941	4.8772	4.9525	5.3231	5.4019	9.2876	10.2957
СА	**О**	6.8068	6.9005	6.6560	6.1729	6.7050	6.7987	7.7212	11.6373	12.5162
	АД	5.9219	6.0397	5.7862	5.5527	5.9693	6.1542	7.0386	10.9657	11.9651
	Д	5.5157	5.6336	5.3583	5.0918	5.3410	5.6008	6.3516	10.2729	11.2646
	МС	6.1986	6.3164	6.2084	6.0284	6.2933	6.5374	7.3759	11.3074	12.4265

Закрашенные ячейки этой таблицы показывают, что из 6 фрагментов, взятых из «середины» поэмы А. Фирдоуси «Рустам ва Сӯҳроб», 5 оказались ближайшими соседями для самой поэмы, а 1 фрагмент – ближайшим соседом для «Бежан бо Манижа». Интересно, что три самых маленьких фрагмента (размерами в 600, 300 и 100) оказались ближайшими соседями с поэмами Дж. Руми «Маснавии Маънавӣ, Дафтари 1» и «Маснавии Маънавӣ, Дафтари 2».

Таблица 2.24.3. Расстояния между ЦП произведений из коллекции текстов и фрагментами, извлеченными из *«**конца**»* поэмы А. Фирдоуси «Рустам ва Сӯҳроб»

Авторы (произв.)		Длины фрагментов (в символах)								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**P&C**	**0.2970**	**0.6326**	**0.7517**	**1.2058**	**2.2150**	**1.8116**	**2.5614**	**3.6217**	**7.0009**
	Б&М	**0.4054**	**0.7524**	**0.7564**	**1.0816**	**2.0840**	**1.6999**	**2.6416**	**3.7151**	**6.9180**
ЧР	**ММ1**	1.4939	1.4527	1.5796	1.6461	1.6562	1.6767	2.8390	3.8865	6.7477
	ММ2	1.5431	1.5019	1.6288	1.7278	1.6483	1.7987	2.8799	4.1090	6.6354
АС	**Д1**	5.4491	5.4744	5.4089	5.4385	5.2945	5.0635	5.2184	4.4615	8.1946
	Д2	5.4792	5.5045	5.4390	5.4944	5.8420	5.5493	5.0882	4.7393	8.6450
СТ	**Н**	3.9949	3.9149	4.2568	4.3021	4.8534	4.8405	4.0109	4.6699	8.7819
	ПКР	4.7605	4.7826	5.0344	5.2695	5.6467	5.6338	5.1509	5.6445	9.9003
СА	**О**	6.7077	6.6907	6.8336	7.1288	7.4567	7.3970	7.3895	6.9663	10.7857
	АД	5.8371	5.8840	6.0757	6.3709	6.7542	6.6538	6.6401	6.8622	10.6726
	Д	5.4310	5.4778	5.6696	5.9648	6.3263	6.2517	6.1265	6.1368	9.9652
	МС	6.1138	6.1607	6.3524	6.6476	6.9356	6.9227	6.9227	6.4659	10.5801

Как явствует из этой таблицы, для 6 фрагментов, взятых из «конца» поэмы А. Фирдоуси «Рустам ва Сӯҳроб», 5 оказались ближайшими соседями для самой поэмы, а 1 фрагмент – ближайшим соседом для «Бежан бо Манижа». Кроме того, ещё 3 фрагмента (размерами в 2500, 1200 и 100) оказались ближайшими соседями с поэмами Дж. Руми «Маснавии Маънавӣ, Дафтари 1» и «Маснавии Маънавӣ, Дафтари 2».

Заключение. Таким образом, результаты, представленные в таблицах,

показывают, что ближайшими соседями по отношению к выбранным фрагментам являются в основном произведения именно того автора, из произведения которого извлекались сами фрагменты. В иной интерпретации это значит, что методом ближайшего соседа путем вычисления расстояний по формуле (2.16) представляется возможным установить авторство достаточно малого кусочка литературного произведения, причём для прозаических произведений (в сравнении с поэтическими) более успешно.

Для художественных текстов можно предложить оценку эффективности применяемого метода, опираясь на вполне естественную гипотезу, согласно которой фрагмент, извлекаемый из какого-либо произведения, должен быть «однородным» с любыми произведениями одного и того же автора. На языке «расстояний» этому соответствует утверждение о том, что ближайшими соседями искомого фрагмента являются, прежде всего, произведения его автора.

Для прозаического произведения, по данным таблиц 2.23.1 и 2.23.3, метод ближайшего соседа безошибочно определяет автора фрагментов размерами не менее 100 символов, а по данным таблицы 2.23.2, – вплоть до 20000 символов.

Для поэтического произведения, по данным таблицы 2.24.1, метод ближайшего соседа безошибочно определяет автора 4 фрагментов из 9 и для 5-и фрагментов размерами 5000, 2500, 1200, 600 и 300 символов допускает ошибку.

По данным таблицы 2.24.2, метод безошибочно определяет автора 6 фрагментов из 9 и для 3-х фрагментов размерами 600, 300 и 100 символов допускает ошибку.

По данным таблицы 2.24.3, метод ближайшего соседа безошибочно определяет автора 6 фрагментов из 9 и для 3-х фрагментов размерами 2500, 1200 и 100 допускает ошибку.

Результаты данного параграфа опубликованы в [27-А].

§ 2.3. Применение частотности длин предложений (в словах)

§ 2.3.1. Применение специфичного ЦП для идентификации авторов произведений

Решается задача распознавания авторов произведений по отдельности для классической и современной поэзий, а также современной прозы. Произведениям сопоставляется ЦП, характеризуемый распределением в них частотности длин предложений. Устанавливается эффективность применения γ-классификатора для идентификации авторов произведений.

В этом пункте мы продолжаем тестирование количественных описаний текстов, начатое в [1-А-10-А], на предмет их пригодности для идентификации авторов произведений. В качестве таковых в [255, 6-А] рассматривались частотности букв таджикского алфавита (униграммы), в [7-А, 8-А] – буквенных

биграмм и триграмм, в [283] – набора из пяти натуральных единиц измерения текста, в [256, 257] – частотности длин слов и знаков препинаний, в [14-А] – частотности слогов. Теперь мы вновь обращаемся к количественному показателю, использованному нами в предыдущих работах [13-А, 258] – *распределению частот встречаемости в текстах длин предложений,* понимая под этим число слов, входящих в состав предложений. Существенным моментом в сравнении с нашим предыдущим исследованием [13-А] является изучение вопроса о распознавании авторов текстов, относящихся к произведениям классической и современной поэзии, а также к современной прозе. Следуя [258], будем называть *цифровым портретом текста* распределение в нём частотности длин предложений. В данном пункте изучается вопрос об эффективности применения такого показателя для распознавания авторов поэтических и прозаических произведений.

1. Обработка статистического материала включала в себя 4 этапа.

Этап 1. Создание компьютерной программы и вычисление с её помощью ЦП произведений – распределений частотности длин предложений по отдельности для всех текстов, упомянутых в § 2.1.1.

Этап 2. Создание компьютерной программы и вычисление с её помощью парных расстояний между ЦПП по формуле, предложенной в § 1.3.

Этап 3. Настройка γ-классификатора. Существо настройки заключается в том, чтобы определить такое значение вещественного параметра γ, при котором достигается максимальное значение критерия «γ-однородности» произведений, см. § 1.4.

Этап 4. Установление эффективности применения настроенного γ-классификатора для распознавания авторов произведений.

На этапе 1 цифровые портреты произведений представляются в табличном виде:

$$\begin{array}{lll} \overline{N} & : & 1 \quad 2 \quad . \; . \; . \quad m \\ P & : & p_1 \quad p_2 \quad . \; . \; . \quad p_m, \end{array}$$

где первая строка – список длин предложений, исчисляемых количеством слов; m – максимальная длина предложения среди произведений модельной коллекции (как отмечено в [13-А] такая длина встретилась в романе «Дохунда» и оказалась равной 178 словам, следовательно, здесь и во всем дальнейшем m = 178); вторая строка – частоты p_i встречаемости в пределах произведений предложений длины $i (i = 1,2, \dots, m)$, причём

$$\sum_{i=1}^{m} p_k = 1.$$

На этапе 2 вычисления расстояний $\rho(T_1, T_2)$ между текстами T_1 и T_2 производились по формуле T_1 и T_2

$$\rho(T_1, T_2) = \sqrt{\frac{m}{2}} \max_s \left| \sum_{k=1}^{s} (p_k^{(1)} - p_k^{(2)}) \right|,$$

в которой m (= 178) – максимальная длина предложения среди произведений модельной коллекции; $p_k^{(1)}$ и $p_k^{(2)}$ – частоты встречаемости в текстах T_1 и T_2 предложений длиной k, k = 1, … , m, и $(s = 1, \cdots, m)$.

Результаты вычислений показаны в таблицах 2.25-2.27.

На этапе 3 качество классификатора при фиксированном γ оценивается величиной π, вычисляемой по формуле

$$\pi = 1 - \tau/L\,, \tag{2.17}$$

где L (= 45) – суммарное число взаимных расстояний между 10 текстами исходной коллекции; $\tau = \tau(\gamma)$ – число нарушений неравенств

$$\rho(T_1, T_2) \le \gamma, \tag{2.18}$$

$$\rho(T_1, T_2) > \gamma. \tag{2.19}$$

Первое проверяется на 5 парах текстов одних и тех же авторов, второе – на 40 парах текстов различных авторов.

На этапе 4 производится настройка γ-классификатора на основе вполне естественной гипотезы о том, что произведения одного автора «однородны», а разных авторов «неоднородны». На языке ЦП, характеризующих распределение частотности длин 10 пар произведений, определение γ сводится к отысканию такого его значения, при котором общее число τ нарушений неравенств (2.18), (2.19) по отдельности на текстах 3-х модельных коллекций становится минимальным. Для нахождения таких γ используется алгоритм, предложенный в § 1.4.

2. Результаты вычислений расстояний между 10 произведениями классической поэзии представлены в табл. 2.25.

Таблица 2.25. – Расстояния между произведениями ***классической поэзии***

Автор (Произ.)		**Число слов**	**АР**		**АФ**		**СШ**		**ХШ**		**ЧР**	
			АП	**К**	**Р&С**	**Б&М**	**Ғ1**	**Ғ2**	**Ғ1**	**Ғ2**	**ММ1**	**ММ2**
			2248	**5054**	**16355**	**14799**	**16261**	**13001**	**33724**	**28923**	**48713**	**41661**
АР	**АП**	**2248**										
	К	**5054**	**1.0121**									
АФ	**Р&С**	**16355**	1.8897	2.3301								
	Б&М	**14799**	1.7038	2.1441	**0.2554**							

Автор (Произ.)		Число слов	АР		АФ		СШ		ҲШ		ЧР	
			АП	Қ	Р&С	Б&М	Ғ1	Ғ2	Ғ1	Ғ2	ММ1	ММ2
			2248	5054	16355	14799	16261	13001	33724	28923	48713	41661
СШ	Ғ1	16261	2.4522	2.0118	4.3419	4.1559						
	Ғ2	13001	2.4973	2.0569	4.3871	4.2011	0.2813					
ҲШ	Ғ1	33724	2.6134	2.1731	4.5031	4.3171	0.2255	0.2986				
	Ғ2	28923	2.1466	1.7062	4.0363	3.8503	0.4141	0.4291	0.5012			
ЧР	ММ1	48713	1.8013	2.2376	0.9915	1.0861	4.1895	4.2718	4.3507	3.8839		
	ММ2	41661	1.9974	2.4337	1.1187	1.2482	4.3185	4.4681	4.4796	4.0391	0.1961	

Для классической поэзии оптимальное значение γ оказалось следующим

$$\gamma^{opt} \in [0.2814;\ 0.2985).$$

Применять этот факт для выяснения метрической близости пары произведений T_1 и T_2 необходимо следующим образом, см. [16-А, 44-А]:

– если $\rho(T_1,\ T_2) \leq 0.2814$, то T_1 и T_2 однородны;

– если $\rho(T_1,\ T_2) > 0.2985$, то T_1 и T_2 неоднородны;

– если $0.2814 < \gamma \leq 0.2985$, то ситуация – неопределенная.

Из данных таблицы следует, что только три расстояния, именно 1.0121; 0.5012 и 0.2255 соответственно между ЦП двух произведений А. Рудаки «Абёти пароканда» и «Қасоид», двух произведений Ҳ. Шерозӣ «Ғазалиёт қисми 1» и «Ғазалиёт қисми 2», а также произведениями С. Шерозӣ «Ғазалиёт қисми 1» и Ҳ. Шерозӣ «Ғазалиёт қисми 1» нарушают сформулированную гипотезу. Первые две пары согласно (2.19) утверждают неоднородность указанных двух произведений А. Рудаки и двух произведений Ҳ. Шерозӣ, а третья пара оказывается «однородной», хотя принадлежат различным авторам.

Желтым цветом в таблице 2.25 отмечены 3 случая нарушения гипотезы однородности.

3. Результаты вычислений расстояний между 10 произведениями современной поэзии представлены в табл. 2.26.

Таблица 2.26. – Расстояния между произведениями в ***современной поэзии***

Автор (Произ.)		Число слов	АС		АШ		ГС		ИФ		МТ	
			Д1	Д2	БТ	ШР	О	Ш	101Г	МГМ	ҚҲ	ҲА
			7890	9322	32036	12810	12103	51434	9841	41217	8463	6118
АС	Д1	7890										
	Д2	9322	0.4853									
АШ	БТ	32036	4.0554	4.1164								
	ШР	12810	3.8219	3.8789	1.7116							
ГС	О	12103	3.9469	4.0039	1.6512	2.7648						
	Ш	51434	2.8258	2.9791	1.5264	1.7675	1.1211					
ИФ	101Г	9841	4.2367	4.3176	1.9836	3.0271	0.7992	1.8808				
	МГМ	41217	2.7667	2.9612	1.8794	2.9261	1.6357	1.1801	2.4349			
МТ	ҚҲ	8463	3.4053	3.5324	1.6655	2.4381	1.7751	1.0201	2.4925	1.4633		
	ҲА	6118	3.8181	3.8751	2.5354	1.7953	3.7936	2.7784	3.8431	3.9585	2.9082	

Для современной поэзии оптимальное значение γ оказалось следующим

$$\gamma^{opt} \in [1.1212;\ 1.1800).$$

Применять этот факт для выяснения метрической близости пары произведений T_1 и T_2 необходимо следующим образом, см. [16-А, 44-А]:

– если $\rho(T_1, T_2) \leq 1.1212$, то T_1 и T_2 однородны;

– если $\rho(T_1, T_2) > 1.1800$, то T_1 и T_2 неоднородны;

– если $1.1212 < \gamma \leq 1.1800$, то ситуация – неопределенная.

И здесь в табл. 2.26 закрашенные жёлтым цветом ячейки (в данном случае их – 5) показывают нарушение сформулированной гипотезы для соответствующих пар произведений.

4. Результаты вычислений расстояний между 10 произведениями современной прозы представлены в табл. 2.27.

Таблица 2.27. – Расстояния между произведениями в ***современной прозе***

Автор (Произ.)		**Число слов**	**АЗ**		**ГМ**		**МШ**		**СТ**		**СА**	
			Б	**З**	**БМ**	**СМ**	**СБ**	**Х**	**Н**	**ПКР**	**Д**	**МС**
			70804	**79431**	**46608**	**50368**	**113592**	**91202**	**9936**	**4041**	**71134**	**48801**
АЗ	**Б**	**70804**										
	З	**79431**	**1.0761**									
ГМ	**БМ**	**46608**	2.4853	1.4347								
	СМ	**50368**	2.1938	1.1649	**0.3889**							
МШ	**СБ**	**113592**	3.0251	2.3241	1.3479	1.6108						
	Х	**91202**	4.0775	3.4589	2.4316	2.7229	**1.1348**					
СТ	**Н**	**9936**	0.6124	0.9986	1.8573	1.5771	2.1502	3.1674				
	ПКР	**4041**	0.7608	1.3605	1.7805	1.5385	1.8218	2.8962	**0.3619**			
СА	**Д**	**71134**	4.0806	3.5927	2.8006	2.9851	1.5412	0.6741	3.1141	2.9803		
	МС	**48801**	4.0644	3.4351	2.4538	2.7268	1.1536	0.6827	3.1784	2.8831	**0.4171**	

Для современной прозы оптимальное значение γ оказалось следующим

$$\gamma^{opt} \in [0.4172;\ 0.6123).$$

Применять этот факт для выяснения метрической близости пары произведений T_1 и T_2 современной прозы необходимо следующим образом, см. [16-А, 44-А]:

– если $\rho(T_1, T_2) \leq 0.4172$, то T_1 и T_2 однородны;

– если $\rho(T_1, T_2) > 0.6123$, то T_1 и T_2 неоднородны;

– если $0.4172 < \gamma \leq 0.6123$, то ситуация – неопределенная.

И здесь закрашенные в табл. 2.27 жёлтым цветом ячейки (в данном случае их – 2) показывают нарушение сформулированной гипотезы для соответствующих пар произведений.

5. Вычисления по формуле (2.17) коэффициента эффективности π

– для классической поэзии выдает значение π = 93%,

– для современной поэзии выдаёт значение π = 89%,

– для современной прозы выдаёт значение π = 96%.

распознавания автора по цифровому портрету его произведений.

Полученные значения показывают, что распознавание автора текста по цифровому портрету (распределению частотности длин предложений) для прозаических произведений (в сравнении с поэтическими) более успешно.

Результаты данного параграфа опубликованы в [17-А].

§ 2.3.2. О распознавании автора текстового фрагмента на основе частотности длин предложений (в словах)

На примере модельной коллекции таджикских литературных произведений изучается задача о возможности определения авторства фрагмента текста минимального размера, извлеченного из коллекции.

В данном параграфе, используя γ-классификатор §§ 1.3-1.4 и цифровой текстовый портрет, предложенный в [258] и характеризующий распределение частотности длин предложений, измеряемых количеством содержащихся в них слов, мы занимаемся идентификацией авторов произведений. Существенным моментом в сравнении с нашим предыдущим исследованием [13-А] является изучение вопроса о минимально допустимом размере текстового фрагмента, для которого ещё удается получить удовлетворительный результат решения рассматриваемой задачи. Отметим, что ранее аналогичный вопрос изучался для других ЦП, именно для символьных (буквенных) униграмм, биграмм и триграмм с учетом пробела, [10-А].

Модельная коллекция текстов, на которой разворачивается наше исследование, та же самая, что и в § 2.1.1.

Обработка коллекционного материала происходила в 4 этапа.

Этап 1 состоял из выбора двух произведений различных авторов, каковыми оказались «Рустам ва Сӯҳроб» А. Фирдоуси и «Дохунда» С. Айни. Из каждого произведения извлекались по 9 случайных выборок текстовых фрагментов, размеры которых в словах и символах показаны в таблице 2.28.

Таблица 2.28. – Информация о размерах фрагментов в словах и символах

Номер фрагмента	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Число слов	7000	3500	1700	900	450	250	130	60	20
Число символов	40000	20000	10000	5000	2500	1200	600	300	100

Как видно из таблицы, размеры фрагментов уменьшаются от номера к номеру.

Замечание. Формальное деление числа символов на соответствующее ему число слов не будет означать среднюю длину слова, исчисляемую количеством букв, поскольку к символам помимо букв относятся также знаки препинания и арифметических операций, цифры, обозначения типа «№», «@», «$» и т.п.

Этап 2. Для каждого фрагмента выбранных произведений строится ЦП, который определяется распределением частотности длин предложений, содержащихся в рассматриваемом фрагменте.

ЦП представляется в табличном виде:

$$\begin{array}{lll} \overline{N} & : & 1 \quad 2 \quad . \; . \; . \quad m \\ P & : & p_1 \quad p_2 \quad . \; . \; . \quad p_m, \end{array}$$

где первая строка – список длин предложений, исчисляемых количеством слов; m – максимальная длина предложения среди произведений модельной коллекции (как отмечено в [13-А] такая длина встретилась в романе «Дохунда» и оказалась равной 178 словам, следовательно, здесь и во всем дальнейшем m = 178); вторая строка – частоты p_i встречаемости в пределах фрагмента предложений длины i $(i = 1, 2, \dots, m)$, причём

$$\sum_{i=1}^{m} p_i = 1.$$

Этап 3. Вычисления расстояний $\rho(v_1, v_2)$ между ЦП 9 фрагментов v_1 и 12 произведениями v_2 рассматриваемой коллекции текстов. Соответствующие вычисления производились по формуле

$$\rho(v_1, v_2) = \sqrt{m/2}\, max_s \left|\sum_{i=1}^{s} \left(p_i^{(1)} - p_i^{(2)}\right)\right|, \qquad (2.20)$$

где $p_i^{(1)}$ и $\mathrm{p_i^{(2)}}$ – частоты встречаемости в фрагментах v_1 и в произведениях v_2 предложений длиной i (i = 1, … , m) и $(\mathrm{s} = 1, \cdots, \mathrm{m})$.

Этап 4. Определение автора текстового фрагмента производится методом ближайшего соседа, см., например, [248]. Существо метода заключается в том, что классифицируемый фрагмент $\mathrm{v_1}$ объявляется принадлежащим тому автору, чьё произведение $\mathrm{v_2}$ в сравнении с другими произведениями оказывается наиболее «сходным» с фрагментом. Иными словами, рассматриваемые объекты являются ближайшими соседями, и расстояния между их ЦП минимальны в сравнении с прочими расстояниями.

Полученные результаты сведены в две группы таблиц соответственно с номерами 2.29.1, 2.29.2, 2.29.3 и 2.30.1, 2.30.2, 2.30.3. В ячейках таблиц представлены расстояния от 12 произведений модельной коллекции текстов до 9 фрагментов из романа С. Айни «Дохунда» (в 1-й группе) и до 9 фрагментов из поэмы А. Фирдоуси «Рустам ва Сӯҳроб» (во 2-й группе). В этих таблицах используются те же сокращения для имен авторов и названий произведений, что и в § 2.1.1.

В последующих таблицах первые 2 колонки указывают авторов и их произведения, а ячейки 9-и других колонок – значения расстояний фрагментов различных длин до соответствующих произведений, вычисленные по формуле (2.20). Отметим также, что в названиях таблиц используются фразы о фрагментах, извлеченных из *«начала»*, *«середины»* и *«конца»* произведений, тем самым в определенном смысле подчеркивается случайный характер выбора фрагментов из текстов произведений.

Таблица 2.29.1. – Расстояния между ЦПП из коллекции текстов и фрагментами, извлеченными из ***«начала»*** романа С. Айни «Дохунда»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**Р&С**	3.3783	4.9715	6.7872	6.9954	7.1174	7.1174	7.0964	7.0964	9.0921
	Б&М	3.4286	5.0218	6.8375	7.0457	7.1677	7.1677	6.9105	6.9105	9.0923
ЧР	**ММ1**	3.7737	5.3669	7.1825	7.3907	7.5127	7.5127	7.2551	6.9441	9.2609
	ММ2	3.9322	5.5254	7.3411	7.5493	7.6713	7.6713	7.3256	7.0731	9.3478
АС	**Д1**	1.6159	3.2091	5.0495	5.6676	5.4033	5.9929	5.9929	5.9929	8.0034
	Д2	1.6925	3.2655	5.1699	5.7881	5.5237	6.1133	6.1133	6.1133	8.0427
СТ	**Н**	2.2554	3.8487	5.7194	6.3374	6.0731	6.6627	6.6627	6.6627	8.6085
	ПКР	2.1139	3.6101	5.5036	6.1217	5.8573	6.4471	6.4471	6.4471	8.5379
СА	**О**	**1.8822**	**0.9487**	**1.8043**	**2.4224**	**2.1581**	**2.7477**	**2.9876**	**3.3602**	**5.6145**
	АД	**1.0943**	**2.0211**	**3.1963**	**3.5291**	**3.5265**	**3.8711**	**4.3428**	**3.8543**	**6.9466**
	Д	**1.0951**	**0.9983**	**2.6248**	**3.2429**	**2.9785**	**3.5682**	**3.5682**	**3.5682**	**5.9816**
	МС	**1.1901**	**1.3745**	**2.6615**	**3.2386**	**2.9917**	**3.5639**	**3.8691**	**3.5639**	**6.3826**

Из таблицы следует, что все фрагменты из «Дохунда» размерами от 20000 и вплоть до 100 символов имеют своим ближайшим соседом роман «Одина» (соответствующие такой ситуации ячейки закрашены) и лишь самый большой фрагмент (в 40000 символов) подтверждает, что он заимствован из самого «Дохунда». Между прочим, все романы С. Айни также являются ближайшими соседями к рассматриваемым фрагментам.

Таблица 2.29.2. – Расстояния между ЦПП из коллекции текстов и фрагментами, извлеченными из ***«середины»*** романа С. Айни «Дохунда»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**Р&С**	4.0421	3.9059	4.8551	5.2593	5.3341	4.8101	4.4169	7.9547	7.9547
	Б&М	4.0924	3.9563	4.8591	5.2633	5.3382	4.8141	4.4211	7.9587	7.9587
ЧР	**ММ1**	4.4374	4.3013	5.0532	5.4575	5.5473	4.9459	4.5529	8.0906	8.0906
	ММ2	4.5961	4.4599	5.2117	5.6161	5.7059	5.1421	4.7491	8.2867	8.2867
АС	**Д1**	2.4126	2.2341	3.2282	3.6326	3.7074	3.1833	2.7902	6.3281	6.3281
	Д2	2.5229	2.3443	3.3385	3.7428	3.8177	3.2936	2.9005	6.4383	6.4383
СТ	**Н**	2.9931	2.8144	3.8087	4.2131	4.2879	3.7638	3.3707	6.9084	6.9084
	ПКР	2.8855	2.5889	3.4821	3.8511	3.9261	3.4019	3.0088	6.5465	6.5465
СА	**О**	**1.0901**	**1.3311**	**0.8312**	**1.2726**	**1.1678**	**2.5293**	**3.3155**	**3.8195**	**3.8195**
	АД	**1.0011**	**0.9379**	**1.4863**	**2.3843**	**2.2794**	**1.2381**	**2.4173**	**4.2576**	**4.2576**
	Д	**0.2542**	**0.5841**	**0.7267**	**1.3751**	**1.2702**	**1.7772**	**2.5633**	**3.8264**	**3.8264**
	МС	**0.4261**	**0.6716**	**0.8598**	**1.7709**	**1.6661**	**1.5033**	**2.2895**	**3.8991**	**3.8991**

И в этом случае (фрагменты взяты из *середины*) основной результат аналогичен предыдущему: все фрагменты из «Дохунда» являются ближайшими соседями произведений С. Айни и никого другого, см. закрашенные ячейки.

Таблица 2.29.3. – Расстояния между ЦПП из коллекции текстов и фрагментами из ***«конца»*** романа С. Айни «Дохунда»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**Р&С**	4.7589	4.5362	4.7926	4.5627	4.4791	4.1811	4.3751	5.5892	7.9477
	Б&М	4.8092	4.6254	4.8789	4.6932	4.5689	4.1851	4.3753	5.5071	7.8655
ЧР	**ММ1**	5.1811	5.0052	5.2587	5.0729	5.0141	4.3171	4.5439	4.7694	7.1279
	ММ2	5.3187	5.1428	5.3963	5.2106	5.1147	4.5131	4.6308	4.7171	7.0755
АС	**Д1**	3.0187	2.8223	3.0769	2.9433	3.0137	2.5544	3.2864	2.0492	4.4077
	Д2	3.1077	2.9318	3.1853	3.0041	3.1341	2.6647	3.3258	2.2036	4.2854
СТ	**Н**	3.6932	3.5173	3.7708	3.6222	3.6836	3.1348	3.8915	1.5331	3.9898
	ПКР	3.5259	3.3573	3.6278	3.4129	3.5202	2.9746	3.8209	1.7486	4.1071
СА	**О**	**0.5874**	**1.0162**	**1.0039**	**0.7948**	**0.9653**	**1.7235**	**2.6669**	**5.0254**	**7.3839**
	АД	**1.6258**	**1.4884**	**1.5808**	**1.3033**	**1.9484**	**1.7994**	**2.2296**	**4.2521**	**6.6105**
	Д	**0.7024**	**0.5651**	**0.6566**	**0.7606**	**0.9338**	**0.9594**	**1.9028**	**4.2613**	**6.6198**
	МС	**1.1033**	**0.9658**	**1.0388**	**0.7064**	**1.3155**	**1.1665**	**2.0086**	**4.3671**	**6.7256**

Для фрагментов из *«конца»* романа «Дохунда» (размерами не менее 600 символов) основной результат – тот же, что и в 2-х предыдущих случаях: ближайшими для них соседями служат только произведения С. Айни. Интересно, что 2 самых маленьких фрагмента (в 300 и 100 символов) оказались ближайшими соседями произведения С. Турсуна «Нисфирӯзӣ», см. соответствующие ячейки.

Таблица 2.30.1. – Расстояния между ЦПП из коллекции текстов и фрагментами, извлеченными из ***«начала»*** поэмы А. Фирдоуси «Рустам ва Сӯҳроб»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**Р&С**	**0.2446**	**0.2731**	**0.3562**	**0.6964**	**1.0645**	**1.5623**	**1.6398**	**2.5832**	**2.3376**
	Б&М	**0.4305**	**0.3892**	**0.4319**	**0.6153**	**0.8806**	**1.7033**	**1.4559**	**2.3993**	**2.5235**
ЧР	**ММ1**	0.9901	1.0436	0.8228	0.7992	0.8037	1.7378	1.1974	2.1408	2.4899
	ММ2	1.1172	1.1709	0.9501	0.9614	0.9311	1.6598	1.2309	2.1128	2.3611
АС	**Д1**	4.0087	4.0623	3.8414	3.8825	3.8224	4.2462	4.1241	4.1241	5.0263
	Д2	4.0895	4.1431	3.9223	3.9395	3.9032	4.3271	4.2854	4.2854	5.1791
СТ	**Н**	4.5136	4.5672	4.3463	4.1863	4.3273	4.7511	4.3043	3.9898	5.6702
	ПКР	4.2691	4.3228	4.1019	3.9419	4.0829	4.5067	4.0598	3.9578	5.3268
СА	**О**	5.2231	5.3725	5.4883	5.7978	6.0394	5.7729	6.2695	6.2695	6.6863
	АД	3.8487	3.9701	3.9471	4.2565	4.4981	4.2317	4.7282	4.7282	5.5797
	Д	4.3117	4.4612	4.5771	4.8865	5.1281	4.8617	5.3582	5.3582	5.8658
	МС	4.2165	4.3661	4.4818	4.7913	5.0329	4.7665	5.2631	5.2631	5.8701

В этой и двух следующих таблицах 9 фрагментов выбираются из поэмы А. Фирдоуси «Рустам ва Сӯҳроб». Закрашенные ячейки показывают, что для 5 фрагментов из 6 ближайшим соседом является именно «Рустам ва Сӯҳроб» и лишь для одного – «Бежан бо Манижа». Кроме того, три других фрагмента (размерами в 2500, 600 и 300) оказались ближайшими соседями с поэмами Дж.

Руми «Маснавии Маънавӣ, Дафтари 1» и «Маснавии Маънавӣ, Дафтари 2».

Таблица 2.30.2. – Расстояния между ЦПП из коллекции текстов и фрагментами, извлеченными из ***«середины»*** поэмы А. Фирдоуси «Рустам ва Сӯҳроб»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**P&C**	**0.3255**	**0.5782**	**0.7907**	**0.6964**	**1.1081**	**1.8881**	**1.2893**	**2.3376**	**3.0772**
	Б&М	**0.3047**	**0.5357**	**0.6607**	**0.8725**	**1.0259**	**1.7041**	**1.4753**	**2.5235**	**3.2611**
ЧР	**ММ1**	0.9739	0.8091	0.7627	0.8391	0.6793	1.4456	1.7241	2.4899	3.5196
	ММ2	1.1011	0.9363	0.8899	0.7901	0.7263	1.4177	1.8513	2.3611	3.5476
АС	**Д1**	3.9925	3.8278	3.7813	3.6814	3.7762	3.7762	4.7428	4.7428	5.4129
	Д2	4.0734	3.9086	3.8622	3.8527	3.9706	3.9706	4.8236	4.8236	5.5345
СТ	**Н**	4.4974	4.3327	4.2862	4.1863	3.8326	3.7581	5.2477	5.2477	6.0731
	ПКР	4.2531	4.0882	4.0418	3.9419	3.5881	3.5137	5.0032	5.0032	5.7501
СА	**О**	4.9925	5.1048	5.3831	5.4634	5.5813	5.1967	5.6381	6.6863	6.6863
	АД	3.6281	3.7565	4.0347	4.1151	4.2331	4.0901	4.5315	5.5797	5.5797
	Д	4.0812	4.1877	4.4659	4.5462	4.6642	4.3762	4.8176	5.8658	5.8658
	МС	3.9866	4.1152	4.3933	4.4737	4.5916	4.3805	4.8219	5.8701	5.8701

Как явствует из этой таблицы, для 7 фрагментов, взятых из «середины» поэмы А. Фирдоуси «Рустам ва Сӯҳроб», 4 оказались ближайшими соседями для самой поэмы, а 3 – ближайшими соседями для «Бежан бо Манижа». Кроме того, ещё 2 фрагмента (размерами в 2500 и 1200) оказались ближайшими соседями с поэмами Дж. Руми «Маснавии Маънавӣ, Дафтари 1» и «Маснавии Маънавӣ, Дафтари 2».

Таблица 2.30.3. – Расстояния между ЦПП из коллекции текстов и фрагментами, извлеченными из ***«конца»*** поэмы А. Фирдоуси «Рустам ва Сӯҳроб»

Авторы (произв.)		**Длины фрагментов (в символах)**								
		40000	**20000**	**10000**	**5000**	**2500**	**1200**	**600**	**300**	**100**
АФ	**P&C**	**0.1697**	**0.3001**	**0.6383**	**0.6929**	**0.7327**	**2.5832**	**0.8071**	**2.6795**	**8.6943**
	Б&М	**0.3536**	**0.4833**	**0.8242**	**0.6367**	**0.8725**	**2.3993**	**0.6911**	**2.8205**	**8.7274**
ЧР	**ММ1**	1.0127	0.7868	0.7906	1.1345	1.7241	2.1408	1.7241	2.8551	9.1419
	ММ2	1.1399	0.9141	0.8552	1.2617	1.8513	2.1128	1.8513	2.7771	9.2124
АС	**Д1**	4.0314	3.8054	3.7466	4.1531	4.7428	4.7428	4.7428	4.7428	7.6941
	Д2	4.1122	3.8863	3.8275	4.2341	4.8236	4.8236	4.8236	4.8236	7.6671
СТ	**Н**	4.5362	4.3103	4.2515	4.6581	5.2477	5.2477	5.2477	5.2477	8.1859
	ПКР	4.2918	4.0659	4.0071	4.4136	5.0032	5.0032	5.0032	5.0032	8.2392
СА	**О**	5.2333	5.3321	5.5663	5.6861	5.5681	5.7978	5.2213	4.3129	4.8744
	АД	3.7071	3.8157	4.1216	4.1727	4.0548	4.2565	3.6801	2.9355	6.2296
	Д	4.3221	4.4208	4.6551	4.7847	4.6668	4.8865	4.3101	3.8711	5.3493
	МС	4.2268	4.3257	4.5598	4.6412	4.5555	4.7913	4.2148	3.4692	5.7558

Закрашенные ячейки этой таблицы показывают, что из 7 фрагментов, взятых из «конца» поэмы А. Фирдоуси «Рустам ва Сӯҳроб», 5 оказались ближайшими соседями для самой поэмы, а 2 – ближайшими соседями для «Бежан бо Манижа». Кроме того, всего лишь 1 фрагмент (размером в 1200) оказался ближайшим соседом поэмы Дж. Руми «Маснавии Маънавӣ, Дафтари 2». Особый интерес представляет собой «выброс», который указывает на то, что фрагмент размером в

100 символов из *«конца»* поэмы «Рустам ва Сӯҳроб» выступает в качестве ближайшего соседа романа С. Айни «Одина».

Заключение. Итак, результаты, представленные в таблицах, показывают, что ближайшими соседями по отношению к выбранным фрагментам являются в основном произведения именно того автора, из произведения которого извлекались сами фрагменты. В иной интерпретации это значит, что методом ближайшего соседа путем вычисления расстояний по формуле (2.20) представляется возможным установить авторство достаточно малого кусочка литературного произведения, причём для прозаических произведений (в сравнении с поэтическими) более успешно.

Для художественных текстов можно предложить оценку эффективности применяемого метода, опираясь на вполне естественную гипотезу, согласно которой фрагмент, извлекаемый из какого-либо произведения, должен быть «*однородным*» с любыми произведениями одного и того же автора. На языке «расстояний» этому соответствует утверждение о том, что ближайшими соседями искомого фрагмента являются, прежде всего, произведения его автора.

Для прозаического произведения, по данным таблиц 2.29.1 и 2.29.2, метод ближайшего соседа безошибочно определяет автора фрагментов размерами не менее 100 символов, а по данным таблицы 2.29.3, – вплоть до 600 символов.

Для поэтического произведения, по данным таблицы 2.30.1, метод ближайшего соседа безошибочно определяет автора 6 фрагментов из 9 и для 3-х фрагментов размерами 2500, 600 и 300 допускает ошибку.

По данным таблицы 2.30.2, метод безошибочно определяет автора 7 фрагментов из 9 и для 2-х фрагментов размерами 2500 и 1200 допускает ошибку.

По данным таблицы 2.30.3, метод ближайшего соседа безошибочно определяет автора 6 фрагментов из 9 и для 2-х фрагментов размерами 1200 и 100 допускает ошибку.

Результаты данного параграфа опубликованы в [15-А].

§ 2.4. Исследование эффективности распознавания автора текстов на узбекском языке

§ 2.4.1. О распознавании автора текста на узбекском языке с помощью символьных униграмм

Рассматривается модельная коллекция текстов узбекского языка, составленная из произведений классической поэзии и современной прозы на кириллической графике. Каждому произведению сопоставлен ЦП – распределение частотностей символьных униграмм.

Необходимость определения автора текста является актуальной проблемой в сфере литературоведения. В настоящем пункте в качестве исследовательского инструмента тестируется классификатор З.Д. Усманова, описанной в §§ 1.3 и 1.4.

1. Модельная коллекция текстов, извлеченная из [268] и предназначенная для исследовательских целей, сформирована из 8 текстов (по 2 текста 4 авторов) и представляется поэмами А. Навои «Лайли ва Мажнун» (АН, Л&М, 162 *Кб*) и «Фарҳод ва Ширин» (АН, Ф&Ш, 340 *Кб*); двумя произведениями З.М. Бобура «Ғазалиёт» (БЗМ, Ғ, 35*Кб*) и «Маҳрами асрор топмадим» (БЗМ, МТ, 120*Кб*); текстами А. Кодирий «Меҳробдан чаён» (АК, М, 801*Кб*) и «Ўткан кунлар» (АК, Ў, 1215*Кб*); прозою С. Айни «Судхўрнинг ўлими» (СА, С, 406*Кб*) и «Эсдаликлар» (СА, Э, 1131*Кб*).

Для авторов и их произведений приняты обозначения, указываемые в скобках: первые две буквы – это инициалы авторов, вторые – сокращенные шифры текстов, третьи – информация о объёмах произведений в килобайтах.

2. Результаты вычислений расстояний между 8 произведениями показаны в таблицах 2.31 и 2.32.

Таблица 2.31. – Расстояния между произведениями без учёта пробела

Авторы и произведения		**Число слов**	**АН**		**БЗМ**		**АК**		**СА**	
			Л&М	Ф&Ш	Ғ	МТ	М	Ў	С	Э
			14455	**29403**	**2609**	**9016**	**59426**	**89319**	**30530**	**81863**
АН	Л&М	**14455**								
	Ф&Ш	**29403**	**0.0377**							
БЗМ	Ғ	**2609**	0.0827	0.0861						
	МТ	**9016**	0.0899	0.0795	**0.0373**					
АК	М	**59426**	0.2014	0.2047	0.1868	0.1906				
	Ў	**89319**	0.2449	0.2482	0.2478	0.2421	**0.0611**			
СА	С	**30530**	0.2611	0.2643	0.2459	0.2467	0.0957	0.1089		
	Э	**81863**	0.2536	0.2571	0.2444	0.2441	0.0861	0.1219	**0.0248**	

Таблицы 2.31 и 2.32 – симметричные. В них числовыми значениями заполнены ячейки, расположенные ниже главной диагонали. Различаются они тем, что число униграмм в таблице 2.31 равно 35, а в таблице 2.32 – 36 (на единицу больше за счёт добавления к буквам символа пробела).

Таблица 2.32. – Расстояния между произведениями с учетом пробела

Авторы и произведения		**Число слов**	**АН**		**БЗМ**		**АК**		**СА**	
			Л&М	Ф&Ш	Ғ	МТ	М	Ў	С	Э
			14455	**29403**	**2609**	**9016**	**59426**	**89319**	**30530**	**81863**
АН	Л&М	**14455**								
	Ф&Ш	**29403**	**0.0507**							
БЗМ	Ғ	**2609**	0.0714	0.0644						
	МТ	**9016**	0.0747	0.0702	**0.0276**					
АК	М	**59426**	0.2202	0.2074	0.1978	0.1947				
	Ў	**89319**	0.2593	0.2464	0.2484	0.2411	**0.0538**			
СА	С	**30530**	0.2711	0.2582	0.2451	0.2388	0.0819	0.0932		
	Э	**81863**	0.2662	0.2533	0.2487	0.2455	0.0752	0.1049	**0.0217**	

В § 1.4 предложен алгоритм для вычисления оптимального значения γ^{opt}, при котором достигается максимальная эффективность π для коллекции $T = \{T^{(j)}\}$. В основу алгоритма положена вполне естественная гипотеза о том, что *произведения одного автора являются однородными, а разных авторов – неоднородными*. Алгоритм, реализованный в виде компьютерной программы, применён к коллекции текстов п. 1, для которых в таблицах 2.31 и 2.32 подсчитаны расстояния между 8 текстами.

Для символьных (буквенных) униграмм без учета пробела оптимальное значение γ оказалось следующим

$$\gamma^{opt} \in [0.0612;\ 0.0794),$$

а для униграмм с учетом пробела –

$$\gamma^{opt} \in [0.0539;\ 0.0643).$$

Полученные результаты необходимо применять следующим образом. В роли оптимального значения γ выступают не одно, а два числа: нижняя и верхняя границы полуинтервала возможных значений γ. Применять этот факт в первом случае (без учёта пробела) для выяснения метрической близости пары произведений T_1 и T_2 необходимо следующим образом, см. например § 2.1.3:

– если $\rho(T_1, T_2) \leq 0.0612$, то T_1 и T_2 однородны;

– если $\rho(T_1, T_2) > 0.0794$, то T_1 и T_2 неоднородны;

– если $0.0612 < \gamma \leq 0.0794$, то ситуация – неопределенная.

Аналогично следует понимать второй случай с учетом пробела.

С учетом сказанного, вычисления по формуле (1.7) коэффициента эффективности π для обоих случаев выдают значение 100%-ной эффективности распознавания автора по цифровому портрету его произведений.

Интересно отметить, что в работах [6-А, 9-А] для определения автора текста на таджикско-персидском языке в кириллической графике аналогичные интервалы значений оказались следующими: для униграмм без пробела – [0.0536; 0.0650), а для униграмм с пробелом – [0.0423; 0.0625). Таким образом, можно говорить о близости интервалов значений γ.

3.Выводы:

– символьные униграммы являются вполне приемлемыми количественными характеристиками для решения проблемы идентификации авторов текстов;

– учёт пробелов в униграммах повышает точность классификации;

– γ-классификатор из §§ 1.3-1.4 позволяет по частотности элементов алфавита буквенных униграмм (с пробелами и без них) с достаточно высокой степенью эффективности идентифицировать произведения поэтов классической узбекской литературы, а также различных авторов современной узбекской прозы.

Высказанное утверждение опирается на результаты обработки ограниченного по объёму материала, который, тем не менее, как по составу авторов, так и по списку использованных произведений представляет собой *представительную выборку* из генеральной совокупности изучаемой предметной области.

Сделанный вывод согласуется с аналогичными результатами для русского и таджикского языков, [227, 283].

§ 2.4.2. О распознавании автора текста на узбекском языке с помощью символьных биграмм

Рассматривается модельная коллекция текстов узбекского языка, составленная из произведений классической поэзии и современной прозы на кириллической графике. Каждому произведению сопоставлен ЦП – распределение частотностей буквенных биграмм.

Необходимость определения автора текста является актуальной проблемой в сфере литературоведения. В настоящем пункте в качестве исследовательского инструмента тестируется классификатор З.Д. Усманова, описанный в §§ 1.3 и 1.4.

1. Результаты вычислений расстояний между 8 произведениями показаны в таблицах 2.33 и 2.34.

Таблица 2.33. – Расстояния между произведениями биграмм без учёта пробела

Авторы и произведения		**Число слов**	**АН**		**БЗМ**		**АК**		**СА**	
			Л&М	Ф&Ш	Ғ	МТ	М	Ў	С	Э
			14455	**29403**	**2609**	**9016**	**59426**	**89319**	**30530**	**81863**
АН	Л&М	**14455**								
	Ф&Ш	**29403**	**0.2393**							
БЗМ	Ғ	**2609**	0.5461	0.5631						
	МТ	**9016**	0.5398	0.4712	**0.2485**					
АК	М	**59426**	1.1986	1.2182	1.3658	1.3341				
	Ў	**89319**	1.4731	1.4701	1.6881	1.6783	**0.3607**			
СА	С	**30530**	1.5713	1.5991	1.7523	1.7079	0.6121	0.6579		
	Э	**81863**	1.5332	1.5398	1.7469	1.7351	0.5459	0.7686	**0.1478**	

Таблицы 2.33 и 2.34 – симметричные. В них числовыми значениями заполнены ячейки, расположенные ниже главной диагонали. Различаются они тем, что число биграмм в таблице 2.33 равно 35^2=1225, а в таблице 2.34 – 36^2=1296 (на 36^2 за счёт добавления к буквам символа пробела).

Таблица 2.34. – Расстояния между произведениями биграмм с учетом пробела

Авторы и произведения		**Число слов**	**АН**		**БЗМ**		**АК**		**СА**	
			Л&М	Ф&Ш	Ғ	МТ	М	Ў	С	Э
			14455	**29403**	**2609**	**9016**	**59426**	**89319**	**30530**	**81863**
АН	Л&М	**14455**								
	Ф&Ш	**29403**	**0.3042**							
БЗМ	Ғ	**2609**	0.4755	0.3921						

Авторы и произведения		**Число слов**	**АН**		**БЗМ**		**АК**		**СА**	
			Л&М	Ф&Ш	Ғ	МТ	М	Ў	С	Э
			14455	**29403**	**2609**	**9016**	**59426**	**89319**	**30530**	**81863**
	МТ	**9016**	0.4543	0.4212	**0.2001**					
АК	М	**59426**	1.3222	1.2522	1.4627	1.4417				
	Ў	**89319**	1.5618	1.4851	1.7531	1.7321	**0.3226**			
СА	С	**30530**	1.6841	1.5941	1.7701	1.7491	0.5883	0.5607		
	Э	**81863**	1.6151	1.5491	1.8121	1.7911	0.5122	0.6671	**0.1542**	

В § 1.4 предложен алгоритм для вычисления оптимального значения $\gamma^{\text{опт}}$, при котором достигается максимальная эффективность π для коллекции $T = \{T^{(j)}\}$. В основу алгоритма положена вполне естественная гипотеза о том, что *произведения одного автора являются однородными, а разных авторов – неоднородными*. Алгоритм, реализованный в виде компьютерной программы, применён к коллекции текстов § 2.4.1, для которых в таблицах 2.33 и 2.34 подсчитаны расстояния между 8 текстами.

Для символьных (буквенных) биграмм без учета пробела оптимальное значение γ оказалось следующим

$$\gamma^{\text{опт}} \in [0.3608;\ 0.4711),$$

а для биграмм с учетом пробела –

$$\gamma^{\text{опт}} \in [0.3227;\ 0.3920).$$

Полученные результаты необходимо применять следующим образом. В роли оптимального значения γ выступают не одно, а два числа: нижняя и верхняя границы полуинтервала возможных значений γ. Применять этот факт в первом случае (без учёта пробела) для выяснения метрической близости пары произведений T_1 и T_2 необходимо следующим образом, см. например § 2.1.5:

– если $\rho(T_1, T_2) \leq 0.3608$, то T_1 и T_2 однородны;

– если $\rho(T_1, T_2) > 0.4711$, то T_1 и T_2 неоднородны;

– если $0.3608 < \gamma \leq 0.4711$, то ситуация – неопределенная.

Аналогично следует понимать второй случай с учетом пробела.

С учетом сказанного, вычисления по формуле (1.7) коэффициента эффективности π для обоих случаев выдают значение 100%-ной эффективности распознавания автора по цифровому портрету его произведений.

Интересно отметить, что в работах [7-А, 9-А] для определения автора текста на таджикско-персидском языке в кириллической графике аналогичные интервалы значений оказались следующими; для биграмм без пробела – [0.4250; 0.4420), а для биграмм с пробелом – [0.3610; 0.3911). Таким образом, можно говорить о близости интервалов значений γ.

2.Выводы:

– символьные биграммы являются вполне приемлемыми количественными характеристиками для решения проблемы идентификации авторов текстов;

– учёт пробелов в биграммах повышает точность классификации;

– γ-классификатор из §§ 1.3 и 1.4 позволяет по частотности элементов алфавита буквенных биграмм (с пробелами и без них) с достаточно высокой степенью эффективности идентифицировать произведения поэтов классической узбекской литературы, а также различных авторов современной узбекской прозы.

Высказанное утверждение опирается на результаты обработки ограниченного по объёму материала, который, тем не менее, как по составу авторов, так и по списку использованных произведений представляет собой *представительную выборку* из генеральной совокупности изучаемой предметной области.

Сделанный вывод согласуется с аналогичными результатами для русского и таджикского языков, [227, 283].

§ 2.4.3. О распознавании автора текста на узбекском языке с помощью символьных триграмм

Рассматривается модельная коллекция текстов узбекского языка, составленная из произведений классической поэзии и современной прозы на кириллической графике. Каждому произведению сопоставлен ЦП – распределение частотностей буквенных триграмм.

Необходимость определения автора текста является актуальной проблемой в сфере литературоведения. В настоящем пункте в качестве исследовательского инструмента тестируется классификатор З.Д. Усманова, описанный в §§ 1.3-1.4.

1. Результаты вычислений расстояний между 8 произведениями показаны в таблицах 2.35 и 2.36.

Таблица 2.35. – Расстояния между произведениями триграмм без учёта пробела

Авторы и произведения		**Число слов**	**АН**		**БЗМ**		**АК**		**СА**	
			Л&М	Ф&Ш	Ғ	МТ	М	Ў	С	Э
			14455	**29403**	**2609**	**9016**	**59426**	**89319**	**30530**	**81863**
АН	Л&М	**14455**								
	Ф&Ш	**29403**	**1.4205**							
БЗМ	Ғ	**2609**	3.2627	3.3418						
	МТ	**9016**	3.2384	2.8312	**1.5639**					
АК	М	**59426**	7.1277	7.2213	8.1767	7.9136				
	Ў	**89319**	8.7316	8.7436	10.1162	9.9477	**2.1351**			
СА	С	**30530**	9.3662	9.4648	10.4313	10.1933	3.6204	3.9013		
	Э	**81863**	9.0693	9.1311	10.3851	10.2732	3.2381	4.5689	**0.8751**	

Таблицы 2.35 и 2.36 – симметричные. В них числовыми значениями заполнены ячейки, расположенные ниже главной диагонали. Различаются они тем, что число триграмм в таблице 2.35 равно 35^3=42875, а в таблице 2.36 – 36^3=46656 (на 36^3 за счёт добавления к буквам символа пробела).

Таблица 2.36. – Расстояния между произведениями триграмм с учетом пробела

Авторы и произведения		**Число слов**	**АН**		**БЗМ**		**АҚ**		**СА**	
			Л&М	Ф&Ш	Ғ	МТ	М	Ў	С	Э
			14455	**29403**	**2609**	**9016**	**59426**	**89319**	**30530**	**81863**
АН	Л&М	**14455**								
	Ф&Ш	**29403**	**1.9061**							
БЗМ	Ғ	**2609**	2.8507	2.3491						
	МТ	**9016**	2.7533	2.5571	**1.2108**					
АҚ	М	**59426**	7.9438	7.5328	8.7813	8.6487				
	Ў	**89319**	9.4701	8.9117	10.5526	10.4011	**1.9435**			
СА	С	**30530**	10.1755	9.5718	10.6429	10.5081	3.5299	3.3644		
	Э	**81863**	9.7266	9.3109	10.8771	10.7516	3.0876	4.0025	**0.9664**	

В § 1.4 предложен алгоритм для вычисления оптимального значения $\gamma^{\text{опт}}$, при котором достигается максимальная эффективность π для коллекции $T = \{T^{(j)}\}$. В основу алгоритма положена вполне естественная гипотеза о том, что *произведения одного автора являются однородными, а разных авторов – неоднородными*. Алгоритм, реализованный в виде компьютерной программы, применён к коллекции текстов § 2.4.1, для которых в таблицах 2.35 и 2.36 подсчитаны расстояния между 8 текстами.

Для символьных (буквенных) триграмм без учета пробела оптимальное значение γ оказалось следующим

$$\gamma^{\text{опт}} \in [2.1352;\ 2.8311),$$

а для триграмм с учетом пробела –

$$\gamma^{\text{опт}} \in [1.9436;\ 2.3490).$$

Полученные результаты необходимо применять следующим образом. В роли оптимального значения γ выступают не одно, а два числа: нижняя и верхняя границы полуинтервала возможных значений γ. Применять этот факт в первом случае (без учёта пробела) для выяснения метрической близости пары произведений T_1 и T_2 необходимо следующим образом, см. например § 2.3.1:

– если $\rho(T_1, T_2) \leq 2.1352$, то T_1 и T_2 однородны;

– если $\rho(T_1, T_2) > 2.8311$, то T_1 и T_2 неоднородны;

– если $2.1352 < \gamma \leq 2.8311$, то ситуация – неопределенная.

Аналогично следует понимать второй случай с учетом пробела.

С учетом сказанного, вычисления по формуле (1.7) коэффициента эффективности π для обоих случаев выдают значение 100%-ной эффективности распознавания автора по цифровому портрету его произведений.

Интересно отметить, что в работах [8-А, 9-А] для определения автора текста на таджикско-персидском языке в кириллической графике аналогичные интервалы значений оказались следующими: для триграмм без пробела – [2.5243; 2.6433), а для триграмм с пробелом – [2.1759; 2.3457). Таким образом, можно говорить о близости интервалов значений γ.

2.Выводы:

– символьные триграммы являются вполне приемлемыми количественными характеристиками для решения проблемы идентификации авторов текстов;

– учёт пробелов в триграммах повышает точность классификации;

– γ-классификатор из §§ 1.3 и 1.4 позволяет по частотности элементов алфавита буквенных триграмм (с пробелами и без них) с достаточно высокой степенью эффективности идентифицировать произведения поэтов классической узбекской литературы, а также различных авторов современной узбекской прозы.

Высказанное утверждение опирается на результаты обработки ограниченного по объёму материала, который, тем не менее, как по составу авторов, так и по списку использованных произведений представляет собой *представительную выборку* из генеральной совокупности изучаемой предметной области. Сделанный вывод согласуется с аналогичными результатами для русского и таджикского языков, [227, 283]. Результаты §§ 2.4.1. – 2.4.3. опубликованы в [48-А, 51-А, 55-А].

§ 2.5. Выводы по результатам главы 2

В настоящей главе мы обращались к модельной коллекции, составленной из трёх частей: произведений классиков таджикско-персидской литературы, произведений современных поэтов и произведений современных прозаиков. Каждая часть коллекции состоит из 10 произведений, по два произведения пяти авторов. Тестированы количественные признаки высокого уровня на предмет возможности их использования в качестве информативных признаков для распознавания автора на примере модельных коллекций художественных произведений таджикского языка, а также узбекского языка и в роли исследовательского аппарата применялись γ-классификатор З.Д. Усманова и метод ближайшего соседа. Наша цель заключалась не только в том, чтобы выявить различия в размерах и расположениях оптимальных полуинтервалов γ, но также и в определении числа нарушений гипотезы однородности, вычислении коэффициента эффективности распознавания авторов по их произведениям в целом и возможно минимальным фрагментам. Фрагменты извлекались из *«начала»*, *«середины»* и *«конца»* произведения, «в пределах» которых бессистемно и случайным образом выбирались кусочки текста различных размеров.

Путем применения метрического классификатора и метода ближайшего (по расстоянию) соседа удалось идентифицировать авторов убывающих по размерам последовательности текстовых фрагментов от величины в 7000 слов (40000 символов) вплоть до 20 слов (100 символов). Описаны результаты экспериментов с минимальным объёмом выборки слов (символов) для распознавания автора текста.

ГЛАВА 3. РАСПОЗНАВАНИЕ ПРИЗНАКОВ ОДНОРОДНОСТИ

Итак, в предыдущей главе на примере модельной коллекции текстов путём подбора оптимального значения γ удалось обучить γ-классификатор относительно успешному распознаванию авторов произведений на таджикском и узбекском языках. В этой главе мы переходим к изучению следующего довольно естественного вопроса: возможно ли идентифицировать другие признаки однородности, такие как тематики текста, язык, оригинал и его перевод, стиль произведений, шифры научных работ и т.д. на основе γ-классификатора. Очевидно, что решение такой задачи имеет чрезвычайно важное практическое значение.

§ 3.1. Распознавание автора и тематики текста

§ 3.1.1. О метризации произведений художественной литературы

В этом параграфе сконструированы ЦП и метрическое пространство произведений. В предположении уникальности авторского творчества устанавливаются пороговые значения метрики, на основе которых определяются классы «однородных» произведений.

1. Пусть $\{\boldsymbol{T}\}$ – конечное множество текстов, написанных на естественном языке, символьный алфавит которого содержит $\boldsymbol{\alpha}$ букв и символ «пробел». Будем характеризовать каждый текст $\boldsymbol{T}$ упорядоченным набором из $\boldsymbol{m}$ символьных $\boldsymbol{N}$-грамм ($\boldsymbol{N} = \boldsymbol{1}, \boldsymbol{2}, \dots$), обозначаемых $\boldsymbol{N_1}, \dots, \boldsymbol{N_m}$, где $\boldsymbol{m} \leq (\boldsymbol{\alpha}+\boldsymbol{1})^{\boldsymbol{N}}$. Тексту $\boldsymbol{T}$ поставим в соответствие точку $(\boldsymbol{\lambda_1}, \dots, \boldsymbol{\lambda_m})$ в-мерном декартовом пространстве, координаты которой $\boldsymbol{\lambda_k}$являются относительными частотами встречаемости $\boldsymbol{N}$-граммы $\boldsymbol{N_k}$, $\boldsymbol{k} = \boldsymbol{1}, \dots, \boldsymbol{m}$, в тексте $\boldsymbol{T}$. Отметим, что $\sum_{\boldsymbol{k=1}}^{\boldsymbol{m}} \boldsymbol{\lambda_k} = \boldsymbol{1}$.

Свяжем с текстом T положительную дискретную функцию

$$F(s) = \sum_{k=1}^{s} \lambda_k \qquad (s = 1, \cdots, m). \tag{3.1}$$

Пусть T_1, T_2 – произвольная пара текстов из множества $\{T\}$ и

$$F_\beta(s) = \sum_{k=1}^{s} \lambda_k^{(\beta)} -$$

соответствующие им дискретные функции, $\beta = 1, 2$ и $s = 1, \cdots, m$. Назовём *расстоянием между текстами* T_1 *и* T_2 *положительное число* $\rho(T_1, T_2)$, *вычисляемое по формуле*

$$\rho(T_1, T_2) = \sqrt{\frac{m}{2}}\, max_s |F_1(s) - F_2(s)|. \tag{3.2}$$

Введенное таким образом расстояние между любыми двумя элементами из множества $M = \{T\}$ превращает последнее в метрическое пространство[1].

[1] Используемое расстояние удовлетворяет трём аксиомам метрического пространства.

2. Метрика в коллекции произведений. В этом пункте рассматривается метрическое пространство 13 произведений художественной литературы советского периода. Необходимые сведения о составе коллекции мы сопровождаем сокращенным обозначением фамилии автора и названий его трудов:

М.А. Шолохов (Ш) «Тихий дон», т.1 (тд1), 92953 *слова*; «Тихий дон», т.2 (тд2), 94471 *слово*; «Тихий дон», т.3 (тд3), 107849 *слов*; «Тихий дон», т.4 (тд4), 126891 *слово*; «Поднятая целина» (пц), 204938 *слов*; «Судьба человека» (сч), 10891 *слово*.

А.С. Серафимович (С) «Железный поток» (жп), 41247 *слов*; «Скитания» (с), 39828 *слов*; «Сопка с крестами» (ск), 4990 *слов*;

Ф.Д. Крюков (К) «На тихом Дону» (нтд), 30037 *слов*; «В глубине» (вг), 27357 *слов*; «К источнику исцелений» (ки), 16625 *слов*; «Казачка» (к), 12162 *слова*.

Приведенные данные показывают, что творчество М.А. Шолохова представлено шестью текстами (четыре тома «Тихого Дона» рассматриваются как отдельные произведения), в то время как А.С. Серафимовича и Ф.Д. Крюкова – тремя и четырьмя текстами, соответственно.

Для каждого текста в качестве информативного признака использовалась частотность встречающихся в нём символьных 3-грамм, формируемых из символа пробела и 33 букв русского алфавита. Общее число таковых 3-грамм не превосходит $39304 = (33+1)^3$.

На основе распределения частотностей 3-грамм каждому тексту поставлена в соответствие дискретная функция $F(s)$, $s = 1, \cdots, 39304$, определяемая соотношением (3.1). Расстояния между текстами подсчитываются по формуле (3.2). Результаты вычислений приведены в таблице 3.1.

Таблица 3.1. – Метрическое пространство коллекции текстов

А/П	А/П	Ш						К				С		
		тд1	тд2	тд3	тд4	пц	сч	нтд	вг	ки	к	жп	с	ск
Ш	тд1													
	тд2	1.4248												
	тд3	1.1711	0.8267											
	тд4	1.1081	1.7896	1.3667										
	пц	1.9571	3.0267	2.7731	1.5735									
	сч	2.6845	3.7542	3.4994	2.3341	1.8442								
К	нтд	1.5616	1.3411	0.9599	1.3959	2.7444	3.4979							
	вг	1.0771	1.2597	0.9769	1.2271	2.0559	2.8909	0.8948						
	ки	1.4773	1.4074	1.2295	1.3871	2.4221	3.1399	1.4266	0.8999					
	к	1.4163	1.5291	1.2015	1.2681	2.4907	3.1923	1.3073	1.1474	1.0611				
С	жп	1.5317	2.4402	2.1538	1.2384	1.3491	2.6578	1.8949	1.6184	1.9145	1.8337			
	с	1.9049	2.8802	2.6183	1.4356	1.0608	1.6386	2.6039	1.9951	2.1984	2.2832	1.0499		
	ск	2.2561	2.4434	2.0518	1.8429	2.3132	2.8209	1.9295	1.5486	1.9341	1.8766	1.3281	1.4238	

Отметим, что в таблице использованы сокращения, принятые ранее, и, в

связи с симметричностью метрики, заполнены ячейки только ниже главной диагонали[2].

3. Пороговое значение метрики. Метрическое пространство, представленное в таблице 3.1, содержит информацию о взаимоотношениях между элементами тестируемой коллекции. То или иное свойство, которое нам хотелось бы приписать какой-то паре произведений на основе количественных показателей, зависит существенно и от математической модели описания объектов, и от математических методов исследования проблемы. В нашей ситуации помимо количественного портрета произведения художественной литературы в виде дискретной функции $\boldsymbol{F(s)}$ на множестве символьных 3-грамм ответственным моментом являются *выбор порогового значения для расстояний между парами элементов коллекции и выводы, которые пытаются привязать к получаемому результату.* Даже вполне естественные требования, которых мы придерживаемся для определения порогового значения $\boldsymbol{\gamma}$ величины метрики, не могут гарантировать безоговорочную объективность выводимых нами заключений.

Итак, для вычисления γ воспользуемся, прежде всего, следующим определением: *пару текстов* T_1 *и* T_2 *назовём* γ*- однородными* $(\gamma > 0)$*, если*

$$\rho\,(T_1\,, T_2) \leq\ \gamma \tag{3.3}$$

и γ*- неоднородными, если*

$$\rho\,(T_1\,, T_2) >\ \gamma. \tag{3.4}$$

Теперь по отношению к коллекции текстов, в которой представлены произведения нескольких авторов, возникает необходимость установления, по возможности, единого значения γ для всех текстовых пар. С этой целью вводится рабочая гипотеза, которую, однако, не следует воспринимать как единственно верное отражение реальной ситуации: *любые два произведения одного автора однородны, а двух разных авторов неоднородны*[3].

С позиции методов классификации образов высказанная гипотеза означает, что 13 интересующих нас текстов должны разделяться на 3 класса (по числу авторов), и каждый класс должен содержать все произведения одного автора. Столь идеальная ситуация не всегда реализуется на практике. Нарушения происходят за счёт случаев «родства» произведений различных авторов. По этой причине мы используем математическую модель, отражающую рабочую гипотезу не абсолютно точно, а лишь «приближенно».

[2] Смысл разноцветных ячеек объясняется в п. 4.

[3] Более определенным было бы высказывание типа: произведения одного автора одинаковы по стилю, а разных авторов не одинаковы. В этом случае, однако, потребовалось описать ЦП авторского стиля.

Предположим, что коллекция текстов $\{T\} = \{T^{(k)}\}$ разделена на непересекающиеся подколлекции $T^{(k)}$, каждая из которых содержит произведения только одного, k-го автора $(k = 1, \cdots, n)$. Для фиксированного значения γ подсчитаем общее число $\aleph^0$ однородных пар произведений всех авторов и общее число $\aleph^н$ неоднородных пар произведений, принадлежащих различным авторам. Отношение

$$\eta = \frac{\aleph^0 + \aleph^н}{\mathbb{N}}, \tag{3.5}$$

в котором $\mathbb{N}$ – общее число пар текстов в коллекции T, характеризует для заданного γ эффективность представления рабочей гипотезы посредством математической модели (3.1) – (3.4). В случае, когда все пары собственных произведений авторов оказываются однородными, то есть удовлетворяют неравенству (3.3), и все пары произведений различных авторов оказываются неоднородными, то есть подчиняются неравенству (3.4), тогда $\eta = 1$. Математическая модель точно отражает идеальную ситуацию. Другой крайности, именно $\eta = 0$, отвечает полная непригодность математической модели.

В общем случае имеем $0 < \eta < 1$, и выглядит вполне естественным осуществить подбор такого значения γ, при котором достигается максимальное значение коэффициента η (3.5), уточняя тем самым математический образ существа рабочей гипотезы.

Отметим, что представленная математическая модель совместно с описанием алгоритма для нахождения оптимального γ предложена в § 1.4. Именно такое γ предлагается использовать в качестве *порогового значения метрики* для принятия решений в рамках конкретных коллекций.

4. Метрическая близость литературных произведений. Для изучения данных, содержащихся в таблице 3.1, подсчитывается пороговое значение $\boldsymbol{\gamma}$ для метрики пространства 13 произведений советской художественной литературы. Вычисления, выполненные с помощью упомянутого ранее алгоритма, привели к следующему результату:

$$\gamma \in [1.4266; 1.5486).$$

Смысл этого соотношения заключается в том, что в роли порогового значения метрики выступают не одно, а два числа: нижняя и верхняя границы полуинтервала возможных значений γ. Применять этот факт для выяснения метрической близости пары текстов T_1 и T_2 необходимо следующим образом:

– если $\rho(T_1, T_2) \leq \gamma_0 = 1.4266$, то тексты T_1 и T_2 однородны;

– если $\rho(T_1, T_2) > \gamma^0 = 1.5486$, то тексты T_1 и T_2 неоднородны;

– и, наконец, если $1.4266 < \gamma < 1.5486$, то ситуация – неопределенная.

Воспользуемся этим правилом, прежде всего, для исследования отношений

между собственными произведениями трёх авторов. Для А.С. Серафимовича такая информация отображается в таблице 3.1 в трёх серых клетках, стоящих на пересечении строк и столбцов с индексами **ЖП**, **С** и **СК**. Поскольку числа в этих клетках строго меньше $\gamma_0 = 1.4266$, то следует заключить, что произведения А.С. Серафимовича однородны (схожи, близки, родственны и т.д.) между собой.

Аналогичное положение имеет место с четырьмя текстами Ф.Д. Крюкова (соответствующие данные показаны в шести ячейках серого цвета). Его произведения также оказываются однородными.

Пестрая картина обнаруживается в творчестве М.А. Шолохова. В таблице 3.1 на пересечении соответствующих строк и столбцов присутствуют ячейки и серого и желтого цветов. Серый, как и в предшествующих случаях, обозначает однородность соответствующей пары собственных произведений. Таких ячеек – 5. Жёлтых – больше, их – 10, и они характеризуют неоднородность соответствующих собственных текстов (значения расстояний в них больше γ^0).

Таким образом, на основе данных таблицы 3.1, заключаем, что

– «Тихий Дон»- т. 1 и «Тихий Дон»- т. 3 однородны с другими томами;

«Тихий Дон»- т. 2 и «Тихий Дон»- т. 4 неоднородны только между собой;

– все четыре тома «Тихого Дона» неоднородны с «Поднятой целиной» и «Судьбой человека»;

– «Поднятая целина» и «Судьба человека» неоднородны между собой.

В таблице 3.1 показаны также данные об отношениях произведений разных авторов. Белый цвет закреплен за теми ячейками, в которых расстояния между соответствующими текстами больше γ^0, что означает неоднородность рассматриваемых элементов. Такое положение отмечается между текстами Ф.Д. Крюкова и А.С. Серафимовича.

Красным цветом окрашены клетки, в которых расстояния между текстами М.А. Шолохова, с одной стороны, и текстами Ф.Д. Крюкова и А.С. Серафимовича, с другой стороны, оказываются меньше, чем γ_0. Такая ситуация указывает на однородность соответственных объектов. В этой связи интересно обратить внимание на то, что тексты Ф.Д. Крюкова, проявляя однородность с текстами «Тихий Дон»- т. 1 и «Тихий Дон»- т. 4, безусловно неоднородны с «Поднятой целиной» и «Судьбой человека», то есть с более поздними трудами М.А. Шолохова.

«Железный поток» А.С. Серафимовича однороден с 4-м томом «Тихого Дона» и «Поднятой целиной» М.А. Шолохова, а «Скитания» – только с «Поднятой целиной».

Ещё один цвет, светло-коричневый, использован для трех ячеек: соответственные им тексты не удаётся классифицировать вполне определённым образом.

Заключение. Метрическая близость литературных произведений,

безусловно, отражает какую-то общность сравниваемых объектов, но не более того. Интерпретация её как «схожесть», «родство», «единообразие» или, как это принято в настоящем параграфе, «однородность» произведений также не вносит ясность в существо вопроса. Отождествление метрической близости с понятием «совпадение стилей», хотя и представляется весьма привлекательным предложением для исследователей, всё же требует серьезного обоснования. Сказанное в равной мере относится и ко всем другим поспешным и даже безответственным выводам, которые хотелось бы приписать двум метрически близким произведениям.

§ 3.1.2. О применимости γ-классификатора к определению тематики и авторства художественных произведений

γ-классификатор дискретных случайных величин, подтвердивший высокую эффективность при идентификации авторства текстовых фрагментов в §§ 2.1-2.5, тестируется на предмет приспособляемости к распознаванию тематики художественных произведений.

1. Изучение статистических закономерностей художественных произведений основывается, с одной стороны, на математических моделях (цифровых портретах) печатного текста и, с другой стороны, на математических методах обработки данных. На сегодняшний день и тех и других – огромное количество. Не обсуждая вопрос, какими соображениями следует руководствоваться для формирования особо продуктивной пары «портрет-метод», отметим, что в нашей работе цифровым портретом текста служит распределение частот буквенных 3-грамм, а инструментом исследования – классификатор дискретных случайных величин, предложенный в §§ 1.3-1.4. Этот классификатор содержит вещественный параметр γ, со значениями которого связываются понятия -однородных и γ-неоднородных текстов. Настройка классификатора (определение оптимального значения параметра) производится на обучающих выборках с использованием гипотез, учитывающих специфику решаемых проблем.

В данном параграфе нас интересует способность классификатора настраиваться на определение авторства и тематики произведений. В качестве рабочей гипотезы в первом случае будет приниматься утверждение об однородности произведений одного автора и неоднородности произведений различных авторов; во втором случае – однородность произведений по одной тематике и неоднородность по различным тематикам.

2. Коллекция текстов. В качестве обучающей выборки используется небольшая модельная коллекция текстов, в состав которой включены двадцать произведений шести авторов. Соответствующие данные (с заключёнными в скобках принятой нами аббревиатурой и сведениями о размерах произведений в словоупотреблениях) приводятся далее:

М. Шолохов (МШ) «Они сражались за Родину» (ОР: 47251), «Судьба человека» (СЧ: 10891) и «Тихий Дон, т.т. 1-4» (ТД1-4: 92953, 94471, 107849, 126891);

Н. Островский (НО) «Как закалялась сталь» (КЗ: 97625) и «Рождённые бурей» (РБ: 52010);

Б. Полевой (БП) «Повесть о настоящем человеке, части 1-4» (ПЧ1-4: 23022, 27663, 22199, 15438) и «Полководец» (П: 24918);

К. Симонов (КС) «Живые и мертвые, книги 1,2» (ЖМ1,2: 147249, 218811);

А. Фадеев (АФ) «Молодая гвардия, части 1,2» (МГ1,2: 112299, 87511) и «Разгром» (Р: 44387);

Д. Фурманов (ДФ) «Чапаев» (Ч: 81769) и «Мятеж» (М: 101133).

Как видно, произведения Н. Островского и Д. Фурманова посвящены Гражданской войне, Б. Полевого и К. Симонова – Великой Отечественной, а М. Шолохова и А. Фадеева – обеим войнам.

Задача А. *Определить эффективность применения γ-классификатора для распознавания авторов и тематик произведений модельной коллекции.*

3. ЦП текстов. Каждому произведению модельной коллекции ставится в соответствие закон распределения частот его символьных 3-грамм, формируемых из 33 букв русского алфавита и символа пробела. Число таких элементов будет не более 39304 $(=(33+1)^3)$. Располагая эти элементы в алфавитном порядке и замещая их порядковыми номерами, получим в табличном виде специфическое распределение частот встречаемости 3-грамм:

$$\begin{array}{lccccc} \overline{N}: & 1 & 2 & . & . & . \quad m \\ P: & \lambda_1 & \lambda_2 & . & . & . \quad \lambda_m, \end{array} \tag{3.6}$$

где в первой строке числами натурального ряда обозначаются символьные триграммы, $\overline{N}$- их множество, $m \le 39304$, λ_k $(k = 1,\ldots,m,)$ – их относительные частотности в пределах конкретного произведения, причём $\lambda_k \ge 0$ и $\sum_1^m \lambda_k = 1$.

На основе (3.6) строится дискретный аналог функции распределения следующим образом:

$$F(s) = \sum_{k=1}^{s} \lambda_k, \qquad s = 1,\ldots,m.$$

Эту функцию, равно как и распределение (3.6), будем называть *цифровым портретом* произведений, что и в § 1.3.

4. Расстояния между произведениями. Пусть W_1 и W_2 – любые два произведения из нашей коллекции. Соответствующие им дискретные функции запишем в виде

$$F^{(p)}(s) = \sum_{k=1}^{s} \lambda_s^{(p)},$$

где $p = 1,2$ и $s = 1,\ldots,m.$

Определение 1. *Расстоянием между* W_1 *и* W_2 *назовем вещественное число* $\rho(W_1, W_2)$, *определяемое по формуле*

$$\rho(W_1, W_2) = \sqrt{m/2} \max_s \left| \sum_{k=1}^{s} (\lambda_s^{(1)} - \lambda_s^{(2)}) \right|, \qquad (3.7)$$

то есть расстояние между текстами вычисляется как минимальное расстояние между их дискретными функциями, помноженное на весовой коэффициент $\sqrt{m/2}$.

Результаты вычислений по формуле (3.7) представлены в таблицах 3.2 и 3.3.

Таблица 3.2. – Расстояния между парами произведений (для распознавания авторства произведения)

МШ						НО		БП					КС		АФ			ДФ	
ОР	СЧ	ТД1	ТД2	ТД3	ТД4	КЗ	РБ	ПЧ1	ПЧ2	ПЧ3	ПЧ4	П	Ж&М1	Ж&М2	МГ1	МГ2	Р	Ч	М
2.18																			
2.31	2.68																		
2.66	3.75	1.42																	
2.27	3.51	1.17	0.83																
1.48	2.33	1.11	1.79	1.37															
2.52	3.84	1.93	1.23	1.45	1.79														
2.15	2.65	1.13	1.71	1.35	1.31	1.54													
1.82	3.25	1.79	1.62	1.58	1.34	2.16	1.78												
1.62	3.46	1.73	1.18	0.87	1.61	1.57	1.19	0.99											
1.77	3.54	1.63	1.07	0.96	1.65	1.58	1.31	0.97	0.54										
2.11	3.92	1.98	1.11	1.01	2.04	1.62	1.83	0.88	0.82	0.55									
3.73	5.66	3.56	2.19	2.44	3.58	2.21	3.43	2.71	2.68	2.49	2.21								
1.51	2.95	1.55	1.85	1.46	1.02	1.64	1.29	1.48	1.21	1.56	1.97	2.82							
1.87	2.01	2.34	2.99	2.58	1.61	2.89	2.21	2.67	2.41	2.74	3.13	4.23	1.56						
2.13	2.69	1.51	1.96	1.55	1.54	1.83	1.36	1.44	1.46	1.51	1.91	3.18	1.38	2.04					
2.63	3.38	1.92	1.49	1.08	1.99	1.22	1.82	1.37	1.31	1.12	1.51	2.39	1.76	2.51	0.87				
1.04	2.31	1.87	2.24	1.92	1.04	2.06	1.65	1.77	1.61	1.85	2.22	3.52	1.05	1.20	0.78	2.15			
1.98	2.93	1.21	1.98	1.34	1.08	1.45	0.91	1.42	1.01	1.39	1.53	3.56	1.22	2.38	1.68	2.15	1.66		
2.11	3.13	1.72	1.94	1.42	1.49	1.69	1.01	1.41	1.11	1.34	1.70	3.40	1.21	2.30	1.60	2.01	1.68	0.76	

Таблица 3.3. – Расстояния между парами произведений (для распознавания тематики произведений)

Книги о Великой Отечественной войне											Книги о Гражданской войне								
ОР	СЧ	ПЧ1	ПЧ2	ПЧ3	ПЧ4	П	Ж&М1	Ж&М2	МГ1	МГ2	ТД1	ТД2	ТД3	ТД4	КЗ	РБ	Р	Ч	М
2.18																			
1.82	3.25																		
1.62	3.46	0.99																	
1.77	3.54	0.97	0.54																
2.10	3.92	0.88	0.82	0.55															
3.73	5.66	2.71	2.68	2.49	2.21														
1.51	2.95	1.48	1.21	1.56	1.97	2.81													
1.87	2.01	2.67	2.40	2.73	3.13	4.23	1.56												
2.13	2.68	1.44	1.46	1.51	1.91	3.18	1.38	2.04											
2.63	3.38	1.37	1.31	1.12	1.51	2.39	1.76	2.51	0.87										
2.31	2.68	1.77	1.73	1.64	1.99	3.56	1.55	2.34	1.51	1.92									
2.66	3.75	1.62	1.18	1.07	1.11	2.19	1.85	2.99	1.96	1.49	1.42								
2.27	3.51	1.58	0.87	0.96	1.01	2.44	1.46	2.58	1.55	1.08	1.17	0.83							
1.48	2.33	1.34	1.61	1.65	2.04	3.58	1.02	1.61	1.54	1.99	1.11	1.79	1.37						
2.52	3.84	2.16	1.56	1.58	1.62	2.21	1.64	2.89	1.83	1.22	1.93	1.23	1.45	1.79					
2.15	2.65	1.78	1.19	1.31	1.83	3.43	1.29	2.21	1.36	1.82	1.13	1.70	1.35	1.31	1.54				
1.04	2.31	1.76	1.61	1.85	2.22	3.52	1.04	1.21	0.79	2.15	1.87	2.24	1.92	1.04	2.06	1.65			
1.98	2.93	1.42	1.01	1.39	1.53	3.56	1.22	2.38	1.68	2.16	1.20	1.98	1.34	1.08	1.45	0.91	1.66		
2.11	3.13	1.40	1.11	1.33	1.71	3.41	1.21	2.31	1.61	2.01	1.72	1.94	1.42	1.49	1.68	1.01	1.68	0.76	

В этих таблицах в ячейках на пересечении строк и столбцов приведены значения расстояний между соответствующими произведениями. Отметим, что в таблице использованы сокращения, принятые ранее, и, в связи с симметричностью расстояния относительно пары произведений, заполнены ячейки только ниже главной диагонали. Различия таблиц – в раскраске ячеек, смысл которых объясняется в п. 6.

5. γ – однородные и γ – неоднородные произведения

Пусть γ – некоторое положительное число.

Определение 2. *Пару произведений* W_1 *и* W_2 *назовём* γ *– однородными, если*

$$\rho(W_1, W_2) \le \gamma, \tag{3.8}$$

и γ *– неоднородными, если*

$$\rho(W_1, W_2) > \gamma, \tag{3.9}$$

Из определения следует, что в зависимости от значения γ произведения W_1 и W_2 могут оказаться как γ-однородными, так и γ-неоднородными. Выбор походящих значений γ естественно связывать с теми закономерностями, которые мы собираемся выявить среди элементов коллекции. В нашем случае таковых будет две:

H1: *произведения одного автора однородны, а разных авторов не-однородны;*

H2: *произведения по одной тематике однородны, а по разным тематикам неоднородны*.

6. Эффективность γ-классификатора. Для того чтобы оценить, насколько успешно γ-классификатор может быть приспособлен к подтверждению упомянутых закономерностей, воспользуемся количественным показателем π, вычисляемым по формуле:

$$\pi = 1 - \tau / N. \tag{3.10}$$

В этой формуле N – число всевозможных пар произведений (для модельной коллекции таких пар 190, поскольку различные тома, книги и части произведений учитываются как отдельные произведения, таковых – 20) и τ – суммарное число нарушений неравенства (3.8) для произведений, принадлежащих одному и тому же автору, и неравенства (3.9) для произведений, принадлежащих различным авторам. Для случая применения γ-классификатора к тематикам произведений τ – суммарное число нарушений неравенства (3.8) для произведений, принадлежащих одной тематике (или Гражданской, или Великой Отечественной войне), и неравенства (3.9) для произведений, описывающих различные войны.

Из формулы (3.10) следует, что π может принимать значения из отрезка [0, 1], причём $\pi = 0$, если $\tau = N$, и $\pi = 1$, если $\tau = 0$. В первом случае математическую модель классификатора следует признать непригодной, а во втором – полностью согласованной с обучающей выборкой. Поскольку эффективность классификатора зависит от параметра γ, то следует определить такое значение γ, при котором π достигает максимальное значение.

6.1. Соответствующий алгоритм предложен в § 1.4. Применяя его для случая H1, устанавливаем, что $\gamma \in [1.1081; 1.1142)$. Последнее означает, что в роли порогового значения γ выступают не одно, а два числа: *нижняя и верхняя границы* полуинтервала возможных значений γ. Применять этот факт для выяснения

метрической близости пары произведения W_1 и W_2 необходимо следующим образом:

– если $\rho(W_1, W_2) \le 1.1081$, то W_1 и W_2 однородны;

– если $\rho(W_1, W_2) > 1.1142$, то W_1 и W_2 неоднородны;

– если $1.1081 < \gamma \le 1.1142$, то ситуация – неопределенная.

Именно с помощью этого правила для удобства читателя в таблице используются три цвета:

– жёлтый, показывающий нарушение неравенства в двух случаях: когда расстояние между двумя произведениями одного автора вместо того, чтобы быть не больше, оказываются строго больше 1.1081, и когда между двумя произведениями разных авторов вместо того, чтобы быть строго больше оказываются не больше 1.1142;

– серый, показывающий выполнение необходимого неравенства для двух произведений одного автора;

– белый, показывающий выполнение необходимого неравенства для двух произведений различных авторов.

Непосредственный подсчет ячеек желтого цвета даёт $\tau = 33$ и потому с учётом того, что $N = 190$, получим

$$\pi = 0.83.$$

Интересно отметить, что если из состава модельной коллекции извлечь произведения М.Шолохова, то для оставшейся части получится довольно компактная картина, см. таблицу 3.4.

Таблица 3.4. – Расстояния между элементами усеченной модельной коллекции (для распознавания авторства произведения)

НО		БП					КС		АФ			ДФ	
КЗ	РБ	ПЧ1	ПЧ2	ПЧ3	ПЧ4	П	Ж&М1	Ж&М2	МГ1	МГ2	Р	Ч	М
1.54													
2.16	1.78												
1.57	1.19	**0.99**											
1.58	1.31	**0.97**	**0.54**										
1.62	1.83	**0.88**	**0.82**	**0.55**									
2.21	3.43	**2.71**	**2.68**	**2.49**	**2.21**								
1.64	1.29	1.48	1.21	1.56	1.97	2.82							
2.89	2.21	2.67	2.41	2.74	3.13	4.23	**1.56**						
1.83	1.36	1.44	1.46	1.51	1.91	3.18	1.38	2.04					
1.22	1.82	1.37	1.31	1.12	1.51	2.39	1.76	2.51	**0.87**				
2.06	1.65	1.77	1.61	1.85	2.22	3.52	1.05	1.20	**0.78**	**2.15**			
1.45	0.91	1.42	1.01	1.39	1.53	3.56	1.22	2.38	1.68	2.15	1.66		
1.69	1.01	1.41	1.11	1.34	1.70	3.40	1.21	2.30	1.60	2.01	1.68	**0.76**	

Вычисления, аналогичные предыдущим, приводят к следующим результатам: $\gamma \in [0.9929; 1.0055)$, $\tau = 8$ и с учётом того, что $N = 91$, получаем

$$\pi = 0.91,$$

то есть для усеченной коллекции текстов (без произведений М. Шолохова) γ-классификатор показывает более высокую эффективность в распознавании

авторства. Таблица 3.4 подтверждает это:

– из восьми случаев нарушения неравенств семь относятся к неравенству (3.8), означающему неоднородность авторских произведений, и лишь одно – к неравенству (3.9), означающему однородность произведений двух авторов («Чапаева» Д. Фурманова и «Рожденные бурей» Н. Островского).

– все другие пары произведений двух различных авторов (таковых – 74) оказались неоднородными, как это и должно быть, см. **H1.**

Заметное улучшение результатов для усеченной коллекции наводит на мысль, что творчество М. Шолохова выходит за рамки общей закономерности, сформулированной в гипотезе **H1.** В определенном смысле этот факт находит поддержку и в результатах § 3.1.1, в котором также, как и в данном случае, ряд собственных произведений М. Шолохова оказываются неоднородными между собой и однородными с некоторыми произведениями А.С. Серафимовича и Ф.Д. Крюкова.

6.2. Выполняя необходимые вычисления для случая **H2**, когда речь идёт уже о γ – однородных и γ – неоднородных тематиках произведений, устанавливаем $\gamma \in [2.0059; 2.0085)$,с помощью чего осуществляем раскраску таблицы 3.2 по тем же правилам применения цветов, что и для ячеек таблицы 3.3. Опять таки непосредственный подсчёт числа жёлтых ячеек даёт $\tau = 91$ и с учётом того, что по-прежнему $N = 190$, получаем

$$\pi = 0.52.$$

Таким образом, тестирование γ-классификатора на предмет распознавания как авторства, так и тематики произведений показало относительно приемлемую эффективность (83%) в первом случае и не вполне удовлетворительную (52%) – во втором случае. Вероятная причина заключалась в том, что в обоих случаях использовался однотипный ЦП – распределение частот встречаемости символьных 3-грамм, который оказался неприемлемым для распознавания тематик произведений. Из этого следует, что не только метод обработки данных, но также и количественный образ объектов исследований важны для достижения желаемых результатов.

§ 3.1.3. К вопросу о распознавании однородных пар произведений художественной литературы

В этом параграфе на примере небольшой коллекции $\boldsymbol{C}$ произведений художественной литературы советского периода изучается совместное влияние ЦП, метрического пространства и классификатора текстов на принятие решения об «однородности» и «неоднородности» произведений. С помощью γ-классификатора на предмет возможной «однородности» изучаются пары основных произведений М.А. Шолохова, А.С. Серафимовича и Ф.Д. Крюкова,

представляемые девятью различными ЦП. Необходимые сведения об использованной нами коллекции, заимствованной из § 3.1.1, сопровождаются сокращенным обозначением фамилий авторов и названий их трудов:

М.А. Шолохов (Ш) «Тихий дон», т. 1 (тд1), 92953 *слова*; «Тихий дон», т. 2 (тд2), 94471 *слово*; «Тихий дон», т. 3 (тд3), 107849 *слов*; «Тихий дон», т. 4 (тд4), 126891 *слово*; «Поднятая целина» (пц), 204938 *слов*; «Судьба человека» (сч), 10891 *слово*.

А.С. Серафимович (С) «Железный поток» (жп), 41247 *слов*; «Скитания» (с), 39828 *слов*; «Сопка с крестами» (ск), 4990 *слов*;

Ф.Д. Крюков (К) «На тихом Дону» (нтд), 30037 *слов*; «В глубине» (вг), 27357 *слов*; «К источнику исцелений» (ки), 16625 *слов*; «Казачка» (к), 12162 *слова*.

Таким образом, творчество М.А. Шолохова представлено шестью текстами (четыре тома «Тихого Дона» рассматриваются как отдельные произведения), а А.С. Серафимовича и Ф.Д. Крюкова – тремя и четырьмя текстами, соответственно.

В §§ 3.1.1 и 3.1.2 для каждого текста в качестве ЦП использовалось распределение частотности встречающихся в нём символьных 3-грамм, формируемых из символа пробела и 33 букв русского алфавита. В нашем случае (в дополнение к предыдущему) текстам сопоставлены ещё 8 различных ЦП.

Приведём определения понятий, используемых в дальнейшем.

1. ЦП текстов. Для формирования количественного образа текста мы выбрали такие его элементы, которые встречаются во многих исследованиях, см. табл. 3.5.

Таблица 3.5. – Список элементов для формирования ЦП текстов

№	Элементы текста	Число элементов (m)
1	униграммы	33
2	биграммы	1089
3	униграммы с учётом пробела	34
4	биграммы с учётом пробела	1156
5	триграммы с учётом пробела	39304
6	словоформы	106482
7	знаки пунктуаций	15
8	униграммы + знаки пунктуаций	48
9	длины предложений (в словах)	120

Из таблицы видно, что мы воспользовались 9 различными типами элементов. Уточним, что числа буквенных биграмм и триграмм (с пробелами) не превосходят, соответственно, значений $(33 + 1)^2 = 1156$ и $(33 + 1)^3 = 39304$. Отметим также, что указанные в таблице количества словоформ и различных длин предложений предварительно подсчитаны для всей коллекции текстов $\boldsymbol{C}$.

Определение 1. ***Алфавитом*** *называется совокупность элементов текста,*

упорядоченных каким-либо образом.

Очевидно, что упорядочение элементов производится для того, чтобы осуществлять необходимые операции с текстами. Для унограмм применяется привычная сортировка по алфавиту в словаре (символ пробела размещается в конце), для биграмм и триграмм – лексикографическая сортировка, для знаков пунктуаций – общепринятый порядок, для словоформ – в порядке убывания их частотности во всей коллекции текстов и для длин предложений – от 1 до 120 (максимальная длина предложения в коллекции $\boldsymbol{C}$).

Определение 2. ***ЦП*** *текста называется распределение в тексте частот встречаемости элементов алфавита.*

Следовательно, ЦПТ – это пара, составленная, с одной стороны, из упорядоченных элементов текста и, с другой стороны, из информации об относительной частоте встречаемости в тексте самих элементов.

ЦП текста T записывается в табличном виде:

$$\begin{array}{lccccc} N: & 1 & 2 & \ldots & m \\ P: & p_1 & p_2 & \ldots & p_m, \end{array} \qquad (3.11)$$

в котором первая строка – порядковые номера (индексы) алфавитных элементов (m – число элементов), а вторая – относительные частоты встречаемости в T алфавитных элементов, причём

$$\sum_{k=1}^{m} p_k = 1.$$

Цифровым портретом текста будем называть также и дискретную функцию $F(s)$, определяемую равенством

$$F(s) = \sum_{k=1}^{s} p_k \qquad (s = 1, \cdots, m). \qquad (3.12)$$

2. Расстояния между ЦП текстов. Пусть T_1, T_2 – произвольная пара текстов, характеризуемых на основе какого-либо единого алфавита**,** и

$$F^{(\alpha)}(s) = \sum_{k=1}^{s} p_k^{(\alpha)} - \qquad (3.13)$$

соответствующие им ЦП, представленные дискретными функциями, $\alpha = 1, 2,$ и $s = 1, \cdots, m.$

Определение 3. *Расстоянием между текстами* T_1 *и* T_2 *называется положительное число* $\rho\,(T_1, T_2)$*, определяемое формулой*

$$\rho\,(T_1, T_2) = \sqrt{m/2}\ max_s \left| F^{(1)}(s) - F^{(2)}(s) \right|, \qquad (3.14)$$

то есть расстояние между двумя текстами вычисляется как максимальное

расстояние по оси ординат между их дискретными функциями $F^{(1)}(s)$ и $F^{(2)}(s)$, помноженное на весовой коэффициент $\sqrt{m/2}$. Отметим также, что равенство $\rho\,(T_1, T_2) = 0$ означает совпадение ЦП T_1 и T_2, но не самих текстов.

3. Гипотеза $\mathbb{H}$ «однородности» авторских произведений. В предположении уникальности авторского творчества вполне естественной представляется:

ГИПОТЕЗА $\mathbb{H}$. *Авторские произведения «однородны», а разных авторов «неоднородны».*

Обнаруживаемые в творчестве авторов «однородности» тех или иных особенностей стилей проявляются в их произведениях, словоупотреблениях, синтаксисе, композиции, интонациях, ритмах и многом другом. Не уточняя этого понятия, ограничимся тем, что сопоставим ему синонимы «похожий», «одинаковый», «сходный», «однотипный», «родственный» и т.п. Все они привязываются к понятию авторского стиля, который индивидуализирует творчество автора на фоне его коллег из писательского сообщества.

В литературе можно указать много примеров нарушения этой гипотезы, однако мы принимаем её к исполнению, как первое приближение к реальной ситуации, позволяющей различать авторов произведений, а нам преобразовать гипотезу в математическую модель.

4. Математическая модель $\mathbb{H}$-гипотезы. Пусть γ – некоторое положительное число.

Определение 4. *Тексты T_1, T_2 называются* ***γ- однородными*** **(***принадлежат одному и тому же автору***),** *если*

$$\rho\,(T_1\,, T_2) \leq \gamma, \tag{3.15}$$

и ***γ- неоднородными*** **(***принадлежат разным авторам***),** *если*

$$\rho\,(T_1\,, T_2) > \gamma. \tag{3.16}$$

Неравенства (3.15) и (3.16) являются математической интерпретацией (моделью) гипотезы $\mathbb{H}$.

Определение 5. *γ-классификатор – зависящий от одного вещественного параметра γ алгоритм принятия решения об отнесении пары текстов T_1 и T_2 к одному или двум разным авторам.*

Очевидно, что от значения γ зависит однородность или неоднородность любой пары текстов и, следовательно, степень выполнимости гипотезы. Если рассматривать предельные значения, именно $\gamma = 0$ и $\gamma = \infty$, то в первом случае все тексты оказываются неоднородными, а во втором – напротив, однородными. Принадлежность двух текстов одному автору в рамках математической модели означает справедливость неравенства (3.15), а двум разным авторам – справедливость неравенства (3.16). Гипотеза $\mathbb{H}$ может нарушаться для каких-то

пар текстов одного и того же автора в случае, когда вместо неравенства (3.15) имеет место неравенство (3.16), а также в случае, когда какие-то два текста разных авторов удовлетворяют неравенству (3.15) вместо того, чтобы выполнялось неравенство (3.16).

Пусть $\tau = \tau(\gamma)$ – суммарное количество нарушений гипотезы $\mathbb{H}$ одновременно в двух случаях: невыполнения неравенства «однородности» в случае двух текстов, принадлежащих одному автору, и невыполнения неравенства «неоднородности» в случае двух текстов, принадлежащих разным авторам. Тогда для фиксированного γ *показатель выполнения гипотезы будет определяться величиной* π, задаваемой формулой

$$\pi = 1 - \tau(\gamma)/L\,, \qquad (3.17)$$

где L – число взаимных расстояний между всеми парами текстов из коллекции $\boldsymbol{C}$. В нашем случае $L = C_{13}^2 = 78$, кроме того, из этого количества 24 пары – расстояния между авторскими тестами и 54 пары – между текстами разных авторов. Из этой формулы следует, что π может принимать значения из отрезка [0, 1], причём $\pi = 0$, если $\tau = L\,(= 78)$, и $\pi = 1$, если $\tau = 0$. В первом случае гипотезу $\mathbb{H}$ следует признать непригодной, а во втором – полностью согласованной с обучающей выборкой.

В связи с тем, что эффективность γ-классификатора зависит от значения параметра γ, представляет интерес найти такое его значение, при котором π принимает максимальное значение. *Именно в этом и заключается суть* ***настройки γ-классификатора*** *на данных обучающей выборки*. Если такая настройка будет приемлемой, то можно говорить о решении задачи обучения -классификатора и его предрасположенности к распознаванию авторов произведений коллекции $\boldsymbol{C}$.

5. Настройка γ-классификатора на текстах коллекции $\boldsymbol{C}$. Соответствующий алгоритм подробно описан в § 1.4. В нашей работе его использование осуществлялось путём последовательного выполнения следующих операций:

– вычисления (в соответствии со списком элементов таблицы 3.5) 9 различных ЦП для каждого из 13 текстов коллекции $\boldsymbol{C}$;

– вычисления по формулам (3.11) – (3.14) для всех ЦП 78 парных расстояний $\rho\,(T_1, T_2)$ между произведениями коллекции $\boldsymbol{C}$;

– вычисление с помощью алгоритма настройки γ-классификатора (по отдельности для каждого из 9 типов ЦП и для всех 13 текстов коллекции $\boldsymbol{C}$) оптимального полуинтервала $\gamma^{\text{опт}}$, для всех значений γ которого величина $\tau = \tau(\gamma)$ суммарного числа случаев нарушения гипотезы $\mathbb{H}$ достигает минимального значения $\boldsymbol{\tau^{min}}$ и, следовательно, величина π показателя

выполнения гипотезы $\mathbb{H}$ принимает максимальное значение $\boldsymbol{\pi^{max}}$ (3.17); результаты вычислений представлены в таблице 3.6.

Таблица 3.6. – Границы оптимального полуинтервала $\gamma^{\text{опт}}$ и значения $\boldsymbol{\tau^{min}}$ и $\boldsymbol{\pi^{max}}$ для различных ЦП

Список ЦП на основе	$\boldsymbol{\gamma^{\text{опт}}}$	$\boldsymbol{\tau^{min}}$	$\boldsymbol{\pi^{max}}$
униграммы	[0.0268; 0.0270)	19	0.756
биграммы	[0.1858; 0.1953)	20	0.744
униграмм с учётом пробела	[0.0310; 0.0311)	23	0.705
биграмм с учётом пробела	[0.1834; 0.1846)	21	0.731
триграмм с учётом пробела	[1.1712; 1.2014)	20	0.744
словоформы	[4.8559; 5.0641)	29	0.628
знаков пунктуаций	[0.0299; 0.0401)	23	0.705
униграмм + знаков пунктуаций	[0.0365; 0.0390)	22	0.718
длин предложений (в словах)	[0.3478; 0.3582)	30	0.615

В этой таблице во 2-м столбце приведены оптимальные значения $\gamma^{\text{опт}}$**,** при которых ошибки τ нарушений гипотезы $\mathbb{H}$ оказываются минимальными. Отметим, что в роли $\gamma^{\text{опт}}$ выступают не отдельные значения, а полуинтервалы значений. Причина в том, что $\tau = \tau(\gamma)$ для всех рассматриваемых ЦП является целочисленной кусочно-гладкой функцией, которая во всех точках указанных полуинтервалов принимает постоянное значение.

В 3-м столбце показаны значения τ^{min} в соответствующих полуинтервалах. При сравнении этих данных (в зависимости от описания текстов теми или иными ЦП) предпочтение следует отдать ЦП на основе униграмм, биграмм и триграмм с пробелом: для них τ^{min} принимает значения 19, 20 и 20 меньшие, чем для других ЦП. Указанным трём числам соответствуют в столбце 4 максимальные значения коэффициентов эффективности π, именно 0.756, 0.744 и 0.744 . В свою очередь, не только эти, но также и все другие числа столбца 4 могут показаться недостаточно высокими, чтобы выводить из них какие-либо заключения.

Однако теперь следует обратить внимание на особенности коллекции $\boldsymbol{C}$. В ней к 4-х томному роману-эпопее «Тихий Дон» наряду с другими сочинениями М.А. Шолохова присоединены также произведения А.С. Серафимовича и Ф.Д. Крюкова. По мнению текстологов, некоторые из их трудов проявляют определенное нами в гипотезе $\mathbb{H}$ свойство «однородности» с главным творением М.А. Шолохова, см. например, [250, 251], что заранее предполагает нарушение сформулированной гипотезы и, следовательно, невозможность достижения γ-классификатором эффективности $\pi^{max} = 1$.

6. Результаты классификации произведений. В девяти последующих таблицах (каждая для соответствующего ЦП) на основе данных о 78 расстояниях между всеми элементами коллекции $\boldsymbol{C}$ приводятся полученные с помощью γ-классификатора распределения материала на однородные и неоднородные пары произведений. Распределение осуществлялось согласно определения 4 при

значениях $\gamma = \gamma^{\text{опт}}$**:** пара текстов объявлялась *однородной*, если расстояние между их ЦП не превосходило значения $\gamma^{\text{опт}}$, и *неоднородной*, если расстояние оказывалось строго больше $\gamma^{\text{опт}}$. В предлагаемых далее таблицах используются принятые в вводной части статьи сокращения имен авторов и названий произведений. В ячейках, расположенных на пересечении столбцов и строк, выдаётся информация о соотношениях между соответствующими текстами: серый цвет ячейки применяется для обозначения однородности соответствующих текстов, бесцветные ячейки – для неоднородных текстов.

Во избежание громоздкости таблиц в ячейках не выписываются надлежащие значения расстояний.

Таблица 3.7. – Однородность текстов. ЦП на основе униграмм

А/П	А/П	Ш						К				С		
		тд1	тд2	тд3	тд4	пц	сч	нтд	вг	ки	к	жп	с	ск
Ш	тд1													
	тд2													
	тд3													
	тд4													
	пц													
	сч													
К	нтд													
	вг													
	ки													
	к													
С	жп													
	с													
	ск													

Данные таблицы 3.7 показывают, что при описании коллекционного материала с помощью ЦП на основе униграмм всего 7 ячеек серого цвета.

Ячейка (тд3, тд2) – однородная, что указывает на однородность 2 и 3-го томов «Тихого Дона» М.А. Шолохова.

Ячейки (тд4, тд1), (тд4, тд2) (тд4, тд1) – однородные, указывают на однородность 4-го тома с 1-м, 2-м и 3-м томами «Тихого Дона» М.А. Шолохова.

Ячейка (вг, тд4) – однородная, указывает на однородность произведения Д.Ф. Крюкова и 4-го тома «Тихого Дона» М.А. Шолохова.

Ячейка (вг, нтд) – однородная, указывает на однородность произведений Д.Ф. Крюкова «В глубине» и «На тихом Дону».

Ячейка (с, жп) – однородная, указывает на однородность произведений «Скитания» и «Железный поток» А.С. Серафимовича.

Остальные ячейки – бесцветные. Соответствующие им пары произведений – неоднородные.

Таблица 3.8. – Однородность текстов. ЦП на основе биграмм

А/П	А/П	Ш						К				С		
		тд1	тд2	тд3	тд4	пц	сч	нтд	вг	ки	к	жп	с	ск
Ш	тд1													
	тд2													
	тд3													
	тд4													
	пц													
	сч													
К	нтд													
	вг													
	ки													
	к													
С	жп													
	с													
	ск													

В этой таблице – 10 ячеек серого цвета. В сравнении с таблицей 3.7, у М.А. Шолохова проявляется ещё одна однородная пара (тд3, тд1), а у Д.Ф. Крюкова – две дополнительные однородные ячейки (вг, тд2), (вг, тд3), указывающие на однородность произведения «В глубине» с 2-м и 3-м томами «Тихого Дона» М.А. Шолохова. С переходом на новый ЦП у А.С. Серафимовича нет изменений.

Таблица 3.9. – Однородность текстов. ЦП на основе униграмм с учётом пробела

А/П	А/П	Ш						К				С		
		тд1	тд2	тд3	тд4	пц	сч	нтд	вг	ки	к	жп	с	ск
Ш	тд1													
	тд2													
	тд3													
	тд4													
	пц													
	сч													
К	нтд													
	вг													
	ки													
	к													
С	жп													
	с													
	ск													

В новой таблице – 11 ячеек серого цвета. У М.А. Шолохова однородными парами становятся (тд4, тд1), (тд3, тд2). У Д.Ф. Крюкова проявляется однородность пар (вг, нтд), (кн, вг) и (к, кн), то есть его собственных произведений. У А.С. Серафимовича обнаруживается связь с М.А. Шолоховым в виде появления однородных пар (жп, тд4) и (с, пц).

Таблица 3.10. – Однородность текстов. ЦП на основе биграмм с учётом пробела

А/П	А/П	Ш						К				С		
		тд1	тд2	тд3	тд4	пц	сч	нтд	вг	ки	к	жп	с	ск
Ш	тд1													
	тд2													
	тд3													
	тд4													
	пц													
	сч													
К	нтд													
	вг													
	ки													
	к													
С	жп													
	с													
	ск													

У М.А. Шолохова однородные пары – те же, что и в таблице 3.9, а вот у Д.Ф. Крюкова исчезает связь (вг, тд4), но остаются связи между всеми собственными произведениями. У А.С. Серафимовича остаются связи с «Поднятой целиной» М.А. Шолохова (с, пц) и собственными трудами (с, жп).

Таблица 3.11. – Однородность текстов. ЦП на основе триграмм с учётом пробела

А/П	А/П	Ш						К				С		
		тд1	тд2	тд3	тд4	пц	сч	нтд	вг	ки	к	жп	с	ск
Ш	тд1													
	тд2													
	тд3													
	тд4													
	пц													
	сч													
К	нтд													
	вг													
	ки													
	к													
С	жп													
	с													
	ск													

Поскольку содержательный анализ таблиц не представлял особых трудностей, начиная с таблицы 3.11 мы ограничимся простой констатацией представленных данных.

М.А. Шолохов: из 15 пар собственных произведений однородными оказались только 3 пары – (тд4, тд1), (тд3, тд1) (тд3, тд2).

Д.Ф. Крюков: из 6 пар собственных произведений однородными оказались 4 пары – (вг, нтд), (кн, вг), (к, вг), (к, кн).

А.С. Серафимович: из 3 пар собственных произведений однородной оказалась 1 пара – (с, жп).

Между М.А. Шолоховым и Д.Ф. Крюковым выявились 3 однородные пары –

(тд1, вг), (тд3, ндт) и (тд3, вг).

Между М.А. Шолоховым и А.С. Серафимовичем – 1 однородная пара (пц, с).

Таблица 3.12. – Однородность текстов. ЦП на основе словоформ

А/П \ А/П		Ш						К				С		
		тд1	тд2	тд3	тд4	пц	сч	нтд	вг	ки	к	жп	с	ск
Ш	тд1	■												
	тд2	▒	■											
	тд3	▒	▒	■										
	тд4				■									
	пц				▒	■								
	сч						■							
К	нтд			▒				■						
	вг								■					
	ки			▒				▒		■				
	к				▒	▒					■			
С	жп	▒	▒									■		
	с			▒				▒		▒			■	
	ск			▒				▒					▒	■

М.А. Шолохов: из 15 пар собственных произведений однородными оказались только 4 пары – (тд3, тд1), (тд2, тд1), (тд3, тд2) и (пц, тд4).

Д.Ф. Крюков: из 6 пар собственных произведений однородной оказалась только 1 пара – (кн, нтд).

А.С. Серафимович: из 3 пар собственных произведений однородной оказалась 1 пара – (ск, с).

Между М.А. Шолоховым и Д.Ф. Крюковым выявились 4 однородные пары – (тд3, кн), (тд3, ндт), (тд4, к) и (пц, к).

Между М.А. Шолоховым и А.С. Серафимовичем – 4 однородные пары – (тд1, жп), (тд2, жп), (тд3, ск) и (тд3, с).

Между Д.Ф. Крюковым и А.С. Серафимовичем – 3 однородные пары – (нтд, ск), (нтд, с) и (кн, с).

Таблица 3.13. – Однородность текстов. ЦП на основе знаков пунктуации

А/П \ А/П		Ш						К				С		
		тд1	тд2	тд3	тд4	пц	сч	нтд	вг	ки	К	жп	с	ск
Ш	тд1	■												
	тд2	▒	■											
	тд3			■										
	тд4			▒	■									
	пц			▒	▒	■								
	сч						■							
К	нтд			▒	▒			■						
	вг								■					
	ки								▒	■				
	к	▒	▒								■			
С	жп											■		
	с												■	
	ск													■

М.А. Шолохов: из 15 пар собственных произведений однородными оказались только 4 пары – (тд2, тд1), (пц, тд3), (тд4, тд3) и (пц, тд4).

Д.Ф. Крюков: из 6 пар собственных произведений однородной оказалась только 1 пара – (кн, вг).

А.С. Серафимович: его произведения оказались неоднородными и между собой и со всеми другими произведениями.

Между М.А. Шолоховым и Д.Ф. Крюковым выявились 4 однородные пары – (к, тд1), (к, тд2), (нтд, тд3) и (нтд, тд4).

Таблица 3.14. – Однородность текстов. ЦП на основе униграмм и знаков пунктуации

А/П	А/П	Ш						К				С		
		тд1	тд2	тд3	тд4	пц	сч	нтд	вг	ки	к	жп	с	ск
Ш	тд1	■												
	тд2		■											
	тд3		▒	■										
	тд4	▒			■									
	пц					■								
	сч						■							
К	нтд			▒				■						
	вг				▒				■					
	ки								▒	■				
	к				▒				▒	▒	■			
С	жп											■		
	с												■	
	ск													■

М.А. Шолохов: из 15 пар собственных произведений однородными оказались 2 пары – (тд4, тд1) и (тд3, тд2).

Д.Ф. Крюков: из 6 пар собственных произведений однородными оказались 3 пары – (к, вг), (кн, вг) и (к, кн).

А.С. Серафимович: его произведения оказались неоднородными и между собой и со всеми другими произведениями.

Между М.А. Шолоховым и Д.Ф. Крюковым выявились 3 однородные пары – (нтд, тд3) и (к, тд4), (вг, тд4).

Таблица 3.15. – Однородность текстов. ЦП на основе длин предложений

А/П	А/П	Ш						К				С		
		тд1	тд2	тд3	тд4	пц	сч	нтд	вг	ки	к	жп	с	ск
Ш	тд1	■												
	тд2	▒	■											
	тд3		▒	■										
	тд4			▒	■									
	пц				▒	■								
	сч						■							
К	нтд							■						
	вг								■					
	ки	▒	▒							■				

А/П	А/П	Ш						К				С		
		тд1	тд2	тд3	тд4	пц	сч	нтд	вг	ки	к	жп	с	ск
	к													
С	жп													
	с													
	ск													

М.А. Шолохов: из 15 пар собственных произведений однородными оказались 4 пары – (тд2, тд1), (тд3, тд2), (тд4, тд3) и (пц, тд4).

Д.Ф. Крюков: из 6 пар собственных произведений однородной оказалась 1 пара – (к, вг).

А.С. Серафимович: из 3 пар его произведений однородными оказались 2 пары – (ск, жп) и (с, жп).

Между М.А. Шолоховым и Д.Ф. Крюковым выявились 4 однородные пары – (ки, тд1), (ки, тд2), (к, тд3) и (к, тд4).

У А.С. Серафимовича обнаруживается связь с М.А. Шолоховым в виде появления 4-х однородных пар (с, тд2), (с, тд3), (жп, тд2) и (жп, тд3).

Пять однородных связей появились у А.С. Серафимовича с Д.Ф. Крюковым – (ск, вг), (жп, вг), (ск, к), (с, к) и (жп, к).

7. Заключение. Таблицы 3.7-3.15 по существу детально характеризуют ситуацию с однородными и неоднородными парами произведений, вычисленными посредством γ-классификатора. Они позволяют, в частности, сделать следующие выводы.

7.1. Произведения М. Шолохова, Д. Крюкова и А. Серафимовича не укладываются в рамки $\mathbb{H}$-гипотезы об однородности авторских произведений и неоднородности произведений различных авторов, что подтверждается показателями таблицы 3.16.

Таблица 3.16. – Число однородных пар авторских произведений при различных цифровых портретах

Цифровые портреты на основе	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	Число пар собственных текстов
Число однородных пар **М.А. Шолохова**	4	5	2	2	3	4	4	2	4	15
Число однородных пар **Д.Ф. Крюкова**	1	1	3	3	4	1	1	3	1	6
Число однородных пар **А.С. Серафимовича**	1	1	1	1	1	1	0	0	2	3

В этой таблице цифры 1-й строки обозначают (в соответствии с таблицей 3.5) названия элементов текста, на основе которых сформированы цифровые портреты произведений. Во 2-й, 3-й и 4-й строках указываются числа однородных авторских произведений при различных цифровых портретах. Из таблицы следует, что ***не только М.А. Шолохову, но также и двум другим писателям, свойственна преимущественно неоднородность собственных произведений***.

7.2. Число однородных пар произведений зависит от ЦП, применяемых для

количественного описания текстов. Таблица 3.17 на данных М. Шолохова и Д. Крюкова подтверждает это.

Таблица 3.17. – Число однородных пар произведений Шолохова – Крюкова

Цифровые портреты на основе	1	2	3	4	5	6	7	8	9
Число однородных пар	1	3	3	2	3	4	4	3	4

В этой таблице цифры 1-й строки имеют тот же смысл, что и в предыдущей таблице. Во 2-й строке указывается число однородных пар при описании текстов теми или иными ЦП.

7.3. В таблице 3.18 представлена более детальная информация об однородности произведений двух донских писателей. В ней применяются сокращенные обозначения тестов, предложенные в вводной части настоящей работы, в частности, (тд3, нтд), (тд3, вг) и (тд4, вг) – сокращенные названия, соответственно, однородных пар Шолохова – Крюкова «Тихий Дон, том 3 – На тихом Дону», «Тихий Дон, том 3 – В глубине» и «Тихий Дон, том 4 – В глубине».

Таблица 3.18. – Встречаемость однородных пар произведений Шолохова – Крюкова при различных цифровых портретах текстов

№	ЦП текстов на основе	**(тд3, нтд)**	**(тд3, вг)**	**(тд4, вг)**
1	униграмм	–	–	V
2	биграмм	–	V	V
3	униграмм с пробелом	V	V	V
4	биграмм с пробелом	V	V	–
5	триграмм с пробелом	V	V	–
6	словоформ	V	–	–
7	знаков пунктуаций	V	–	–
8	униграмм и зн. пункт.	V	–	V
9	длин предл. (в словах)	–	–	–

В этой таблице символами «V» и «–» показываются, в каких цифровых портретах встречаются и не встречаются те или иные указанные однородные произведения. Другие пары, такие как (тд1, вг), (тд1, к), (тд1, ки), (тд2, вг), (тд2, к), (тд2, ки), (тд3, к) и (тд4, к), также встречаются в разных ЦП, однако не более 2 раз.

7.4. Поскольку связь творчества М.А. Шолохова с А.С. Серафимовичем не столь выпукла как с его соплеменником Д.Ф. Крюковым, в данной работе она не подвергается подробному обсуждению.

Результаты §§ 3.1.1. – 3.1.3. опубликованы в [41-А, 44-А, 52-А].

§ 3.2. Исследование статистических закономерностей определения языка текстов

§ 3.2.1. Распознавание языка произведения с помощью γ-классификатора

В настоящем параграфе на примере модельной коллекции текстов устанавливается применимость γ-классификатора для автоматического

распознавания языка произведения на основе частотности алфавитных букв. С момента своего появления в 2017 году γ-классификатор получил широкое применение в решении самых разнообразных задач автоматической обработки текста, [286, 287].

В качестве экспериментального материала нами выбрана небольшая коллекция $\boldsymbol{C}$ из 10 произведений (текстов), среди которых

на английском языке:

W. Shakespeare «Romeo and Juliet» (en_1, 25832 *слова*);

M. Twain «A Connecticut Yankee in King Arthur's Court» (en_2, 117257 *слов*);

на казахском языке:

А. Құнанбаев «Абайдың қара сөздері (Суханҳои сиёҳи Абай)» (kz_1, 16751 *слово*),

А. Құнанбаев «Автобиография» (kz_2, 1709 *слов*);

на русском языке:

М.А. Горький «Дело Артамоновых» (ru_1, 71456 *слов*);

А.А. Фадеев «Разгром» (ru_2, 44409 *слов*);

на таджикском языке:

С. Айни «Аҳмади Девбанд» (tj_1, 7485 *слов*);

С. Турсун «Повести Камони Рустам» (tj_2, 4041 *слово*);

на узбекском языке:

С. Айни «Судхўрнинг ўлими (қисса)» (uz_1, 30543 *слова*);

Қ. Абдулла «Меҳробдан чаён (роман)» (uz_2, 59426 *слов*).

В представленном списке приведены имена авторов, названия их произведений и в конце, в скобках, выписывается обозначение произведения и его размер в количестве слов.

1. ЦПП. В качестве учётных элементов для описания произведений взяты

– 26 букв английского алфавита (a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z),

– 42 буквы казахского алфавита (а, ә, б, в, г, ғ, д, е, ё, ж, з, и, й, к, қ, л, м, н, ң, о, ө, п, р, с, т, у, ұ, ү, ф, х, h, ц, ч, ш, щ, ъ, ы, і, ь, э, ю, я),

– 33 буквы русского алфавита (а, б, в, г, д, е, ё, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ц, ч, ш, щ, ы, ь, ъ, э, ю, я),

– 35 букв таджикского алфавита (а, б, в, г, ғ, д, е, ё, ж, з, и, ӣ, й, к, қ, л, м, н, о, п, р, с, т, у, ӯ, ф, х, ҳ, ч, ҷ, ш, ъ, э, ю, я) и

– 35 букв узбекского алфавита (а, б, в, г, ғ, д, е, ё, ж, з, и, й, к, қ, л, м, н, о, п, р, с, т, у, ў, ф, х, ҳ, ц, ч, ш, ъ, ь, э, ю, я)

Для количественного описания всех произведений сформирован единый алфавит $\boldsymbol{A}$ из 70 букв (а, б, в, г, ғ, д, е, ё, ж, з, и, ӣ, й, к, қ, л, м, н, о, п, р, с, т, у, ӯ, ф, х, ҳ, ч, ҷ, ш, ъ, э, ю, я, ц, щ, ы, ь, ў, a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v,

w, x, y, z, ә, ң, ө, ұ), которым в таблице ASCII сопоставлены уникальные числовые коды.

Определение 1. ЦП *текста* *будем называть распределение частотности 70 букв алфавита* ***A*** *в пределах текста* (*произведения*).

ЦП текста T записывается в табличном виде:

$$\begin{array}{lll} N: & 1 \quad 2 \quad \ldots \quad m \\ P: & p_1 \quad p_2 \quad \ldots \quad p_m, \end{array} \tag{3.18}$$

в котором первая строка – порядковые номера букв алфавита $\boldsymbol{A}$ (m=70 – число элементов), а вторая − их относительные частоты встречаемости в тексте T, причём $\sum_{k=1}^{m} p_k = 1$.

ЦП представляется также в виде дискретной функции

$$F(s) = \sum_{k=1}^{s} p_k \qquad (s = 1, \cdots, m). \tag{3.19}$$

2. Расстояния между ЦП текстов. Пусть T_1, T_2 – произвольная пара текстов, характеризуемых на основе единого алфавита ***A,*** и

$$F^{(\alpha)}(s) = \sum_{k=1}^{s} p_k^{(\alpha)} - \tag{3.20}$$

соответствующие им ЦП, представленные дискретными функциями, $\alpha = 1, 2,$ и $s = 1, \cdots, m$ (=70).

Определение 2. *Расстоянием между текстами* T_1 *и* T_2 *называется положительное число* $\rho\,(T_1, T_2)$*, определяемое формулой*

$$\rho\,(T_1, T_2) = \sqrt{m/2}\ max_s \left| F^{(1)}(s) - F^{(2)}(s) \right|, \tag{3.21}$$

то есть расстояние между двумя текстами вычисляется как максимальное расстояние по оси ординат между их дискретными функциями $F^{(1)}(s)$ и $F^{(2)}(s)$, помноженное на весовой коэффициент $\sqrt{m/2}$. Отметим также, что равенство $\rho\,(T_1, T_2) = 0$ означает совпадение ЦП T_1 и T_2, но не самих текстов.

3. Гипотеза $\mathbb{H}$ «однородности» произведений. Она привлекается для того чтобы выделить характерную особенность текстов, предназначенную для построения математической модели распознавания языка произведений. Её мы формулируем в следующем виде.

ГИПОТЕЗА $\mathbb{H}$. *Произведения, написанные на одном языке, «однородны», а на разных языках – «неоднородны».*

Говоря об «однородности» произведений (текстов), мы имеем в виду их похожесть, одинаковость, сходность, однотипность, родственность и т.п.

4. Математическая модель $\mathbb{H}$-гипотезы. Пусть γ – некоторое положительное число.

Определение 3. *Тексты* T_1, T_2 *называются* ***γ- однородными*** **(***написанными на одном языке***)**, *если*

$$\rho\,(T_1\,,T_2) \le\ \gamma, \tag{3.22}$$

и ***γ- неоднородными*** **(***написанными на разных языках***)**, *если*

$$\rho\,(T_1\,,T_2) > \gamma. \tag{3.23}$$

Неравенства (3.22) и (3.23) являются математической интерпретацией (моделью) гипотезы $\mathbb{H}$.

Определение 4. *γ-классификатор – зависящий от одного вещественного параметра γ алгоритм принятия решения об отнесении пары текстов T_1 и T_2 к одному или двум разным языкам.*

Очевидно, что от значения γ зависит однородность или неоднородность любой пары текстов, следовательно, и степень выполнимости гипотезы. Принадлежность двух текстов одному языку в рамках математической модели означает справедливость неравенства (3.22), а двум разным языкам – справедливость неравенства (3.23). Гипотеза $\mathbb{H}$ может нарушаться для каких-то пар текстов одного и того же языка в случае, когда вместо неравенства (3.22) имеет место неравенство (3.23), а также в случае, когда какие-то два текста на разных языках удовлетворяют неравенству (3.22) вместо того, чтобы выполнялось неравенство (3.23).

Пусть $\tau = \tau(\gamma)$ – суммарное количество нарушений гипотезы $\mathbb{H}$ одновременно в двух случаях: невыполнения неравенства «однородности» в случае двух текстов, принадлежащих одному языку, и невыполнения неравенства «неоднородности» в случае двух текстов, принадлежащих разным языкам. Тогда для фиксированного γ *показатель выполнения гипотезы будет определяться величиной* π, задаваемой формулой

$$\pi = 1 - \tau(\gamma)/L\,, \tag{3.24}$$

где L – число взаимных расстояний между всеми парами текстов из коллекции $\boldsymbol{C}$ (в нашем случае $L = C_{10}^2 = 45$). Из этой формулы следует, что π может принимать значения из отрезка [0, 1], причём $\pi = 0$, если $\tau = L\,(= 45)$, и $\pi = 1$, если $\tau = 0$. В первом случае гипотезу $\mathbb{H}$ следует признать непригодной, а во втором – полностью согласованной с обучающей выборкой.

В связи с тем, что эффективность γ-классификатора зависит от значения параметра γ, представляет интерес найти такое его значение, при котором π принимает максимальное значение. *Именно в этом и заключается суть* ***настройки γ-классификатора*** *на данных обучающей выборки*. Если такая настройка будет приемлемой, то можно говорить о решении задачи обучения γ-

классификатора и его предрасположенности к распознаванию языков произведений самых разнообразных коллекций.

5. Итоговые результаты на примере модельной коллекции $\boldsymbol{C}$ получены далее путём последовательного выполнения следующих операций:

– вычисления ЦП (частотности букв алфавита $\boldsymbol{A}$) для всех 10 произведений модельной коллекции $\boldsymbol{C}$;

– вычисления по формулам (3.18), (3.19), (3.20) и (3.21) сорока пяти парных расстояний $\rho\,(T_1, T_2)$ между произведениями коллекции $\boldsymbol{C}$ (результаты расчетов приведены в следующей таблице):

Таблица 3.19. – Расстояния между текстами коллекции $\boldsymbol{C}$

Тексты		**En**		**Kz**		**Ru**		**Tj**		**Uz**	
		en_1	**en_2**	**kz_1**	**kz_2**	**ru_1**	**ru_2**	**tj_1**	**tj_2**	**uz_1**	**uz_2**
En	**en_1**										
	en_2	**0.1365**									
Kz	**kz_1**	5.5898	5.5898								
	kz_2	5.5966	5.5966	**0.1240**							
Ru	**ru_1**	5.9160	5.9160	0.5658	0.6141						
	ru_2	5.9147	5.9147	0.5638	0.6121	**0.0493**					
Tj	**tj_1**	5.9145	5.9145	0.7830	0.8328	0.8445	0.8838				
	tj_2	5.9152	5.9152	0.7800	0.8297	0.7783	0.8176	**0.1270**			
Uz	**uz_1**	5.9152	5.9152	0.6516	0.7011	0.9026	0.8616	0.5067	0.5617		
	uz_2	5.9159	5.9159	0.6554	0.7053	0.8188	0.7778	0.4229	0.4779	**0.1348**	

– вычисление с помощью алгоритма настройки γ-классификатора из § 1.4 оптимального интервала значений γ, для которого величина $\tau = \tau(\gamma)$ суммарного числа случаев нарушения гипотезы $\mathbb{H}$ достигает минимального значения и, следовательно, величина π (3.24) показателя выполнения гипотезы $\mathbb{H}$ принимает максимальное значение.

По данным таблицы 3.19 получены следующие результаты:

– совокупность всех пар расстояний размещается на отрезке [0.0493, 5.9160];

– оптимальный полуинтервал значений γ оказывается в пределах

$$\gamma^{\text{опт}} \in [0.1366;\ 0.4228);$$

в соответствии с определением 3 это значит, что если расстояние $\rho\,(T_1\,, T_2)$ между двумя текстами не превосходит значение $\gamma^{\text{опт}}$ из указанного полуинтервала, то пара текстов принадлежит одному и тому же языку; если же превосходит, то принадлежит разным языкам;

– отметим, что для всех (без исключения) произведений коллекции $\boldsymbol{C}$ полностью подтвердилась гипотеза $\mathbb{H}$ и её математическая интерпретация в виде определения 3, и потому получено

$$\tau = \tau_{min} = 0,$$

то есть ни одно из неравенств (3.22) и (3.23) не было нарушено;

– вследствие чего показатель эффективности предложенной в настоящей работе математической модели распознавания языка произведений оказался равным

$$\pi = \pi_{max} = 1.$$

6. Заключение. γ-классификатор с фиксированным значением $\gamma = \gamma^{\text{опт}}$ был протестирован на случайных выборках текстов и подтвердил 100%-ную способность к распознаванию языка.

Отметим также некоторые особенности, обнаруживаемые среди данных таблицы 3.19:

– самым близким оказалось расстояние между парой произведений на русском языке

$$\rho\ (\text{ru\_1}\ ,\ \text{ru\_2}) = 0.0493;$$

– расстояния между парами произведений для других языков оказались в пределах

от 0.1240 до 0.1365;

– расстояния от английских текстов до текстов на 4-х других языках оказались не менее 5.5000;

– расстояния между восемью произведениями (без учета текстов на английском языке) не превосходят 0.9026; вероятная причина в том, что алфавиты всех этих произведений на базе кириллицы.

Результаты данного параграфа опубликованы в [53-А].

§ 3.2.2. Об автоматическом идентификации языка текстов на основе кириллического алфавита

В данном параграфе на примере модельной коллекции из 10 текстов на пяти языках с использованием кириллической графики устанавливается применимость γ-классификатора для автоматического распознавания языка произведения на основе частотности общих 29 кириллических алфавитных букв.

В качестве экспериментального материала нами выбрана небольшая коллекция $\boldsymbol{C}$ из 10 произведений (текстов), среди которых

на казахском языке:

А. Қунанбаев «Абайдың қара сөздері (Чёрные речи Абая)» (kz1, 16751 *слово*),

А. Қунанбаев «Автобиография» (kz2, 1709 *слов*);

на киргизском языке:

Ч. Айтматов «Айтматовдун акыркы адашуусу жана айкөлдүгү (Последнее заблуждение и щедрость Айтматова)» (ky1, 1510 *слов*);

Ч. Айтматов «Өмүр баяны (Биография)» (ky2, 1012 *слова*);

на русском языке:

М.А. Горький «Дело Артамоновых» (ru1, 71456 *слов*);

А.А. Фадеев «Разгром» (ru2, 44409 *слов*);

на таджикском языке:

С. Айни «Аҳмади Девбанд» (Могучий Ахмад) (tj1, 7485 *слов*);

С. Турсун «Повести Камони Рустам» (Лук Рустама) (tj2, 4041 *слово*);

на узбекском языке:

С. Айни «Судхўрнинг ўлими (легенда)» (Смерть ростовщика) (uz1, 30543 *слова*);

Қ. Абдулла «Меҳробдан чаён (роман)» (Скорпион из алтаря) (uz2, 59426 *слов*).

В списке приведены имена авторов, названия произведений на родном языке и в переводе на русский, обозначения произведений и их размеры в количестве слов. Особенность коллекции в том, что в ней все тексты представлены в кириллической графике с использованием специфических символов ң, ө, ү – в киргизском, ә, ғ, қ, ң, ө, ұ, ү, h, i – в казахском, ғ, ӣ, қ, ӯ, ҷ – в таджикском и ғ, қ, ў, ҳ – в узбекском языках.

Из 33 букв кириллицы современного русского языка [272] общими для всех текстов являются 29, именно: а, б, в, г, д, е, ё, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ч, ш, ъ, э, ю, я.

Задача, которая представляет интерес, заключается в том, чтобы определить, насколько важным являются дополнения национальных алфавитов специфическими символами и возможно ли обойтись только лишь общими кириллическими буквами для автоматического распознавания языка, на котором написана та или иная печатная продукция.

1. ЦПП. В качестве учётных элементов для описания произведений взяты указанные для всех 5 языков 29 букв.

Определение 1. *ЦПТ будем называть распределение в нём частотности 29 букв.*

ЦП текста T записывается в табличном виде:

$$\begin{array}{lccccc} N: & 1 & 2 & \ldots & 29 \\ P: & p_1 & p_2 & \ldots & p_{29}, \end{array}$$

в котором первая строка – номера букв, расположенных в алфавитном порядке, а вторая – относительные частоты встречаемости букв в тексте T, причём $\sum_{k=1}^{29} p_k = 1$.

ЦП представляется также в виде дискретной функции

$$F(s) = \sum_{k=1}^{s} p_k \qquad (s = 1, \cdots, 29). \tag{3.25}$$

2. Расстояния между ЦПТ. Пусть T_1, T_2 – произвольная пара текстов,

характеризуемых на основе единого алфавита, и

$$F^{(\alpha)}(s) \;=\; \sum_{k=1}^{s} p_k^{(\alpha)} \;- \tag{3.26}$$

соответствующие им ЦП, представленные дискретными функциями, $\alpha = 1, 2,$ и $s = 1, \cdots, 29.$

Определение 2. *Расстоянием между текстами* T_1 *и* T_2 *называется положительное число* $\rho\,(T_1, T_2)$*, определяемое формулой*

$$\rho\,(T_1\,, T_2) = \sqrt{\frac{29}{2}}\; max_{\,s}\left|F^{(1)}(s) - \;F^{(2)}(s)\right|. \tag{3.27}$$

3. Гипотеза $\mathbb{H}$ «однородности» произведений. Она привлекается для того, чтобы выделить характерную особенность текстов, предназначенную для построения математической модели распознавания языка произведений. Её мы формулируем в следующем виде:

ГИПОТЕЗА $\mathbb{H}$. *Произведения, написанные на одном языке, «однородны», а на разных языках – «неоднородны».*

Говоря об «однородности» произведений (текстов), мы имеем в виду их похожесть, одинаковость, сходность, однотипность, родственность и т.п.

4. Математическая модель $\mathbb{H}$-гипотезы. Пусть γ – некоторое положительное число.

Определение 3. *Тексты* T_1, T_2 *называются* $\boldsymbol{\gamma}$*- однородными (написанными на одном языке), если*

$$\rho\,(T_1\,, T_2) \leq\; \gamma, \tag{3.28}$$

и $\boldsymbol{\gamma}$*- неоднородными (написанными на разных языках), если*

$$\rho\,(T_1\,, T_2) > \gamma. \tag{3.29}$$

Неравенства (3.28) и (3.29) являются математической интерпретацией (моделью) гипотезы $\mathbb{H}$.

Определение 4. *γ-классификатор – зависящий от одного вещественного параметра γ алгоритм принятия решения об отнесении пары текстов T_1 и T_2 к одному или двум разным языкам.*

Очевидно, что от значения γ зависит однородность или неоднородность любой пары текстов, следовательно, и степень выполнимости гипотезы. Принадлежность двух текстов одному языку в рамках математической модели означает справедливость неравенства (3.28), а двум разным языкам – справедливость неравенства (3.29). Гипотеза $\mathbb{H}$ может нарушаться для каких-то пар текстов одного и того же языка в случае, когда вместо неравенства (3.28) имеет место неравенство (3.29), а также в случае, когда какие-то два текста на разных языках удовлетворяют неравенству (3.28) вместо того, чтобы выполнялось неравенство (3.29).

Пусть $\tau = \tau(\gamma)$ – суммарное количество нарушений гипотезы $\mathbb{H}$ одновременно в двух случаях: невыполнения неравенства «однородности» в случае двух текстов, принадлежащих одному языку, и невыполнения неравенства «неоднородности» в случае двух текстов, принадлежащих разным языкам. Тогда для фиксированного γ показатель выполнения гипотезы будет определяться величиной π, задаваемой формулой

$$\pi = 1 - \tau(\gamma)/L\,,$$

где L – число взаимных расстояний между всеми парами текстов из коллекции $\boldsymbol{C}$ (в нашем случае $L = C_{10}^2 = 45$). Из этой формулы следует, что π может принимать значения из отрезка [0, 1], причём $\pi = 0$, если $\tau = L\,(= 45)$, и $\pi = 1$, если $\tau = 0$. В первом случае гипотезу $\mathbb{H}$ следует признать непригодной, а во втором – полностью согласованной с обучающей выборкой.

В связи с тем, что эффективность γ-классификатора зависит от значения параметра γ, представляет интерес найти такое его значение, при котором π принимает максимальное значение. Именно в этом и заключается суть настройки γ-классификатора на данных обучающей выборки. Если такая настройка будет приемлемой, то можно говорить о решении задачи обучения γ-классификатора и его предрасположенности к распознаванию языков произведений самых разнообразных коллекций.

5. Итоговые результаты на примере модельной коллекции $\boldsymbol{C}$ приведены далее путём последовательного выполнения следующих операций:

– вычисления ЦП (частотности букв алфавита $\boldsymbol{A}$) для всех 10 произведений модельной коллекции $\boldsymbol{C}$;

– вычисления по формулам (3.25), (3.26) и (3.27) сорока пяти парных расстояний $\rho\,(T_1, T_2)$ между произведениями коллекции $\boldsymbol{C}$ (результаты расчетов приведены в следующей таблице):

Таблица 3.20. – Расстояния между текстами коллекции $\boldsymbol{C}$

Автор (Произведения)		**Ky**		**Kz**		**Ru**		**Tj**		**Uz**	
		ky_1	**ky_2**	**kz_1**	**kz_2**	**ru_1**	**ru_2**	**tj_1**	**tj_2**	**uz_1**	**uz_2**
Ky	**ky_1**										
	ky_2	**0.0914**									
Kz	**kz_1**	0.3214	0.3416								
	kz_2	0.3530	0.3844	**0.1012**							
Ru	**ru_1**	0.4001	0.3307	0.5206	0.5499						
	ru_2	0.3719	0.3024	0.4924	0.5217	**0.0341**					
Tj	**tj_1**	0.3470	0.3974	0.2774	0.3067	0.5747	0.6009				
	tj_2	0.3063	0.3568	0.3152	0.3445	0.5341	0.5603	**0.0871**			
Uz	**uz_1**	0.3430	0.3787	0.2995	0.3422	0.5483	0.5200	0.3051	0.3429		
	uz_2	0.2680	0.3037	0.3743	0.4175	0.4829	0.4546	0.2841	0.2775	**0.0957**	

– вычисление с помощью алгоритма настройки γ-классификатора из § 1.4 оптимального интервала значений γ, для которого величина $\tau = \tau(\gamma)$ суммарного

числа случаев нарушения гипотезы $\mathbb{H}$ достигает минимального значения и, следовательно, величина π показателя выполнения гипотезы $\mathbb{H}$ принимает максимальное значение.

По данным таблицы 3.20 получены следующие результаты:

– совокупность всех пар расстояний размещается на отрезке [0.0871, 0.6009], при этом минимальное расстояние реализуется между произведениями С. Айни «Аҳмади Девбанд» и С. Турсуна «Повести Камони Рустам» на таджикском языке, а максимальное – между романами С. Айни «Аҳмади Девбанд» на таджикском языке и А.А. Фадеевым «Разгром» на русском языке;

– оптимальный полуинтервал значений γ оказывается в пределах

$$\gamma^{\text{опт}} \in [0.1013;\ 0.2679);$$

в соответствии с определением 3 это значит, что если расстояние $\rho\,(T_1\,, T_2)$ между двумя текстами не превосходит значение $\gamma^{\text{опт}}$ из указанного полуинтервала, то пара текстов принадлежит одному и тому же языку; если же превосходит, то принадлежат разным языкам;

– отметим, что для всех (без исключения) произведений коллекции $\boldsymbol{C}$ полностью подтвердилась гипотеза $\mathbb{H}$ и её математическая интерпретация в виде определения 3, и потому получено

$$\tau = \tau_{min} = 0,$$

то есть, ни одно из неравенств (3.28) и (3.29) не было нарушено;

– вследствие чего показатель эффективности предложенной в настоящей работе математической модели распознавания языка произведений оказался равным

$$\pi = \pi_{max} = 1.$$

6. Тестирование. Итак, настройка (обучение) γ-классификатора на данных модельной коллекции текстов $\boldsymbol{C}$ прошла успешно. Для тестирования классификатора выбраны следующие тексты:

на казахском языке:

Қ. Шакарем «Өмірбаяны (Автобиография)» (Text_Kz, 1359 *слов*);

на киргизском языке:

Ч. Айтматов «Айтматовдун жаңы чыгармасы табылды (Найдена новая работа Айтматова)» (Text_Ky, 205 *слов*);

на русском языке:

М.А. Шолохов «Судьба человека» (Text_Ru, 10891 *слово*);

на таджикском языке:

С. Улуғзода «Бежан ва Манижа» (Text_Tj, 19730 *слов*);

на узбекском языке:

Қ. Абдулла «Ўткан кунлар» (Text_Uz, 89319 *слов*).

После формирования ЦП произведений и вычисления расстояний по формуле (3.27) получена следующая таблица расстояний от каждого из текстов до всех 10 произведений исходной коллекции:

Таблица 3.21. – Расстояния между текстами коллекции и тестируемыми произведениями

Автор (Произведения)		**Text_Ky**	**Text_Kz**	**Text_Ru**	**Text_Tj**	**Text_Uz**
Ky	**ky_1**	**0.0983**	0.2777	0.4106	0.2745	0.2984
	ky_2	**0.1506**	0.3093	0.3411	0.3110	0.3319
Kz	**kz_1**	0.2991	**0.0904**	0.5311	0.3950	0.3773
	kz_2	0.2641	**0.1401**	0.5604	0.4243	0.4252
Ru	**ru_1**	0.3806	0.4386	**0.0341**	0.4883	0.5264
	ru_2	0.3524	0.4382	**0.0561**	0.5145	0.4982
Tj	**tj_1**	0.2663	0.2714	0.5994	**0.1235**	0.2948
	tj_2	0.2427	0.2390	0.5587	**0.1111**	0.3210
Uz	**uz_1**	0.2927	0.2670	0.5587	0.4227	**0.1022**
	uz_2	0.2177	0.3424	0.4933	0.3573	**0.0576**

В таблице 3.21 серым цветом выделены пары ячеек, которые соответствуют минимальным расстояниям от тестируемых объектов до элементов коллекции текстов. Полученные результаты показывают, что ближайшими соседями выбранной пятёрки произведений оказались как раз однородные с ними по языку пары текстов исходной коллекции.

Заключение. γ-классификатор с фиксированным значением $\gamma = \gamma^{\text{опт}}$ был протестирован на случайных выборках текстов и подтвердил 100%-ную способность к распознаванию языка.

Результаты данного параграфа опубликованы в [21-А].

§ 3.2.3. Об автоматическом определении языка текстов на основе латинского алфавита

В данном параграфе на примере модельной коллекции из 10 текстов на пяти языках с использованием латинской графики устанавливается применимость γ-классификатора для автоматического распознавания языка произведения на основе частотности общих 26 латинских алфавитных букв.

В наше время письменность на основе латинского алфавита получила широкое распространение среди романской, германской славянской, финно-угорской, тюркской, семитской и иранской групп языков, среди стран Индокитая, Зондского архипелага и Филиппин, Африки (южнее Сахары), Америки, Австралии и Океании, [273]. За исключением современного английского, для большинства других языков латинский алфавит из 26 букв (a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z) оказался недостаточным, в связи с чем для

отражения фонетических особенностей тех или иных языковых систем к базовой латинской графике были добавлены различные диакритические знаки, лигатуры и другие модификации букв. В настоящем параграфе изучается вопрос о том, возможно ли обойтись только лишь 26 латинскими буквами для автоматического распознавания языка, на котором написана та или иная печатная продукция.

В качестве экспериментального материала нами выбрана коллекция ***C*** из 10 произведений (текстов), среди которых

на английском языке: У. Шекспир «Romeo and Juliet» (en_1, 25832 *слова*), М. Твейн «A Connecticut Yankee in King Arthur's Court» (en_2, 21705 *слов*); *на венгерском языке:* J. Benzoni «Az átok» (vn_1, 12190 *слов*), J. Benzoni «A templomosok kincse» (vn_2, 20538 *слов*); *на латинском языке:* S. Boethius «De philosophiae consolatione» (lt_1, 24680 *слов*), IV. Carolus «Vita Caroli» (lt_2, 15144 *слова*); *на литовском языке:* A. Gutje «Mėlynas rūkas» (li_1, 20262 *слова*), A. Marinina «Triju ne desnis, часть 1» (li_2, 25486 *слов*); *на нидерландском языке:* P. Aspe «De kinderen van Chronos» (ni_1, 14082 *слова*), R. Jordan «Vuur uit de hemel» (ni_2, 19214 *слова*).

В списке приведены имена авторов, названия произведений на родном языке, обозначения произведений и их размеры в количестве слов. Особенность коллекции в том, что она охватывает всего лишь 5 европейских языков и все её тексты – на основе латинской графики с использованием дополнительных специфических символов: 9-х á, é, í, ó, ö, ő, ú, ü, ű – в венгерском (**vn**) и 6-и символов ą, č, į, š, ų, ž – в литовском (**li**) языках.

1. ЦП произведений. В качестве учётных элементов для описания произведений взяты указанные для всех 5 языков 26 латинских букв.

Определение 1. *Цифровым портретом текста будем называть распределение в нём частотности 26 букв.*

ЦП текста T записывается в табличном виде:

$$\begin{array}{lcccccc} N & : & 1 & 2 & . & . & . \quad 26 \\ P & : & p_1 & p_2 & . & . & . \quad p_{26}, \end{array} \tag{3.30}$$

в котором первая строка – номера букв, расположенных в алфавитном порядке, а вторая – относительные частотности букв в тексте T, причём $\sum_{k=1}^{26} p_k = 1$.

Одновременно с (3.30) ЦП представляется также в виде дискретной функции

$$F(s) = \sum_{k=1}^{s} p_k \qquad (s = 1, \cdots, 26). \tag{3.31}$$

2. Расстояния между ЦП текстов. Пусть T_1, T_2 – произвольная пара текстов, характеризуемых на основе единого алфавита, и

$$F^{(\alpha)}(s) = \sum_{k=1}^{s} p_k^{(\alpha)} - \tag{3.32}$$

соответствующие им ЦП, представленные дискретными функциями, $\alpha = 1, 2,$ и $s = 1, \cdots, 26.$

Определение 2. *Расстоянием между текстами* T_1 *и* T_2 *называется положительное число* $\rho\,(T_1, T_2)$*, определяемое формулой*

$$\rho\,(T_1\,, T_2) = \sqrt{\frac{26}{2}}\; max_{\,s}\left|F^{(1)}(s) - \; F^{(2)}(s)\right|. \tag{3.33}$$

3. Гипотеза $\mathbb{H}$ «однородности» произведений. Она привлекается для того, чтобы выделить характерную особенность текстов, предназначенную для построения математической модели распознавания языка произведений. Её мы формулируем в следующем виде.

ГИПОТЕЗА $\mathbb{H}$. *Произведения, написанные на одном языке, «однородны», а на разных языках – «неоднородны».*

Говоря об «однородности» произведений (текстов), мы имеем в виду их похожесть, одинаковость, сходность, однотипность, родственность и т.п.

4. Математическая модель $\mathbb{H}$-гипотезы. Пусть γ – некоторое положительное число.

Определение 3. *Тексты* $T_1,\ T_2$ *называются* γ*- однородными (написанными на одном языке), если*

$$\rho\,(T_1\,, T_2) \leq\ \gamma, \tag{3.34}$$

и γ*- неоднородными (написанными на разных языках), если*

$$\rho\,(T_1\,, T_2) > \gamma. \tag{3.35}$$

Неравенства (3.34) и (3.35) являются математической интерпретацией (моделью) гипотезы $\mathbb{H}$. Это значит, что в дальнейшем мы приступаем к распознаванию языков произведений с помощью математического аппарата, названного γ-классификатором, см. §§ 1.3-1.4.

Определение 4. γ*-классификатор – зависящий от одного вещественного параметра* γ *алгоритм принятия решения об отнесении пары текстов* T_1 *и* T_2 *к одному или двум разным языкам.*

Очевидно, что от значения γ зависит однородность или неоднородность любой пары текстов, следовательно, и степень выполнимости гипотезы. Принадлежность двух текстов одному языку в рамках математической модели означает справедливость неравенства (3.34), а двум разным языкам – справедливость неравенства (3.35). Гипотеза $\mathbb{H}$ может нарушаться для каких-то пар текстов одного и того же языка в случае, когда вместо неравенства (3.34) имеет место неравенство (3.35), а также в случае, когда какие-то два текста на разных языках удовлетворяют неравенству (3.34) вместо того, чтобы выполнялось неравенство (3.35).

Пусть $\tau = \tau(\gamma)$ – суммарное количество нарушений гипотезы $\mathbb{H}$ одновременно в двух случаях: невыполнения неравенства «однородности» в случае двух текстов, принадлежащих одному языку, и невыполнения неравенства

«неоднородности» в случае двух текстов, принадлежащих разным языкам. Тогда для фиксированного γ показатель выполнения гипотезы будет определяться величиной π, задаваемой формулой

$$\pi = 1 - \tau(\gamma)/L\,,$$

где L – число взаимных расстояний между всеми парами текстов из коллекции $\boldsymbol{C}$ (в нашем случае $L = C_{10}^2 = 45$). Из этой формулы следует, что π может принимать значения из отрезка [0, 1], причём $\pi = 0$, если $\tau = L\ (= 45)$, и $\pi = 1$, если $\tau = 0$. В первом случае гипотезу $\mathbb{H}$ следует признать непригодной, а во втором – полностью согласованной с обучающей выборкой.

В связи с тем, что эффективность γ-классификатора зависит от значения параметра γ, представляет интерес найти такое его значение, при котором π принимает максимальное значение. Именно в этом и заключается суть настройки γ-классификатора на данных обучающей выборки. Если такая настройка будет приемлемой, то можно говорить о решении задачи обучения γ-классификатора и его предрасположенности к распознаванию языков произведений самых разнообразных коллекций.

5. Итоговые результаты на примере коллекции $\boldsymbol{C}$ приведены далее. Им предшествовали следующие операции:

– предобработка экспериментальных данных путём удаления из всех произведений коллекции дополнительных сверх латинского алфавита букв;

– вычисления ЦП (3.30) (частотности 26 латинских букв) для всех 10 произведений модельной коллекции $\boldsymbol{C}$;

– вычисления по формулам (3.31), (3.32) и (3.33) сорока пяти парных расстояний $\rho\,(T_1, T_2)$ между произведениями коллекции $\boldsymbol{C}$ (результаты расчетов приведены в следующей таблице):

Таблица 3.22. – Расстояния между текстами коллекции $\boldsymbol{C}$

Тексты		**En**		**Vn**		**Lt**		**Li**		**Ni**	
		en_1	**en_2**	**vn_1**	**vn_2**	**lt_1**	**lt_2**	**li_1**	**li_2**	**ni_1**	**ni_2**
En	**en_1**										
	en_2	**0.0928**									
Vn	**vn_1**	0.1778	0.2700								
	vn_2	0.1969	0.2891	**0.0223**							
Lt	**lt_1**	0.2443	0.3219	0.3030	0.3178						
	lt_2	0.2133	0.2910	0.2380	0.2471	**0.0707**					
Li	**li_1**	0.3893	0.4669	0.2945	0.2799	0.2550	0.2389				
	li_2	0.4030	0.4806	0.3083	0.2936	0.2709	0.2428	**0.0445**			
Ni	**ni_1**	0.3348	0.2420	0.3547	0.3722	0.4152	0.3835	0.5464	0.5601		
	ni_2	0.3838	0.2910	0.3770	0.3945	0.4589	0.4113	0.5839	0.5976	**0.0491**	

– вычисление с помощью алгоритма настройки γ-классификатора оптимального интервала значений γ, для которого величина $\tau = \tau(\gamma)$ суммарного числа случаев нарушения гипотезы $\mathbb{H}$ достигает минимального значения и,

следовательно, величина π показателя выполнения гипотезы $\mathbb{H}$ принимает максимальное значение.

По данным таблицы 3.22 получены следующие результаты:

– совокупность всех пар расстояний размещается на отрезке [0.0223, 0.5976], при этом минимальное расстояние реализуется между двумя произведениями **vn_1** и **vn_2** на венгерском языке, а максимальное – между **li_2** на литовском и **ni_2** на нидерландском языках;

– оптимальный полуинтервал значений γ оказывается в пределах

$$\gamma^{\text{опт}} \in [0.0929;\ 0.1777);$$

в соответствии с определением 3 это значит, что если расстояние $\rho\,(T_1\,, T_2)$ между двумя текстами не превосходит значение $\gamma^{\text{опт}}$ из указанного полуинтервала, то пара текстов принадлежит одному и тому же языку (соответствующие расстояния в таблице помечены серым цветом); если же превосходит, то принадлежат разным языкам (соответствующие расстояния оставлены непомеченными);

– отметим, что для всех (без исключения) произведений коллекции $\boldsymbol{C}$ полностью подтвердилась гипотеза $\mathbb{H}$ и её математическая интерпретация в виде определения 3, и потому получено

$$\tau = \tau_{min} = 0,$$

то есть ни одно из неравенств (3.34) и (3.35) не было нарушено;

– вследствие чего показатель эффективности предложенной в настоящей работе математической модели распознавания языка произведений оказался равным

$$\pi = \pi_{max} = 1.$$

6. Тестирование. Итак, настройка (обучение) γ-классификатора на данных модельной коллекции текстов C прошла успешно. Для тестирования классификатора выбрано 5 текстов:

на английском языке **(En):** Дж. Лондон «The Call of the Wild» (Text_En, 31763 *слова*); *на венгерском языке* **(Vn):** A. Sztrugackij «A bíborszínű felhők bolygója» (Text_Vn, 25603 *слова*); *на латинском языке* **(Lt):** S. Lucilio «Ad Lucilium Epistulae Morales, часть 1» (Text_Lt, 11770 *слов*); *на литовском языке* **(Li):** A. Marinina «Triju ne desnis, часть 2» (Text_Li, 18507 *слов*); *и на нидерландском языке* **(Ni):** R. Jordan «Hart van de Winter» (Text_Ni, 18749 *слов*).

Для пяти произведений, предназначенных для тестирования, построены цифровые портреты (3.30) и затем по формулам (3.31), (3.32), (3.33) для каждого из них вычислены расстояния до 10 объектов коллекции $\boldsymbol{C}$. Соответствующие значения записаны в ячейках таблицы 3.23, расположенных на пересечениях строк и столбцов. Там же серым цветом выделены пары ячеек с минимальными

расстояниями между тестируемыми и содержащимися в коллекции произведениями.

Таблица 3.23. – Расстояния между текстами коллекции и тестируемыми произведениями

Тексты		Text_En	Text_Vn	Text_Lt	Text_Li	Text_Ni
En	**en_1**	**0.1758**	0.2324	0.2376	0.4042	0.3772
	en_2	**0.0969**	0.3246	0.3153	0.4818	0.2845
Vn	**vn_1**	0.3441	**0.0607**	0.2518	0.3095	0.3772
	vn_2	0.3632	**0.0508**	0.2665	0.2948	0.3947
Lt	**lt_1**	0.4188	0.3209	**0.0513**	0.2716	0.4729
	lt_2	0.3878	0.2502	**0.0507**	0.2435	0.4420
Li	**li_1**	0.5638	0.2339	0.2347	**0.0342**	0.6179
	li_2	0.5775	0.2476	0.2513	**0.0119**	0.6316
Ni	**ni_1**	0.1898	0.3881	0.3948	0.5613	**0.0715**
	ni_2	0.2389	0.4104	0.4330	0.5988	**0.0442**

Полученные результаты показывают, что ближайшими соседями выбранной пятёрки произведений оказались как раз однородные с ними по языку пары текстов исходной коллекции.

Заключение. Итак, **γ**-классификатор с фиксированным значением $\gamma = \gamma^{\text{опт}}$ на случайных выборках текстов с ЦП на основе частотности 26 базовых латинских букв подтвердил 100%-ную статистическую способность к распознаванию языков произведений.

Результаты данного параграфа опубликованы в [60-А].

§ 3.2.4. Тестирование γ-классификатора, настроенного на определение языков произведений на основе кириллического алфавита

В этом пункте на примере модельной коллекции текстов на белорусском, болгарском, русском, таджикском и рушанском языках, использующих кириллическую графику, устанавливается способность γ-классификатора на основе частотности общих 26 алфавитных кириллических букв автоматически распознавать язык произведения. С момента своего появления в 2017 году γ-классификатор из §§ 1.3 и 1.4 получил широкое применение в решении разнообразных задач автоматической обработки текста. Продолжаем тестирование количественных описаний текстов, начатое в работах [1-А-10-А], на предмет их пригодности для автоматического определения языка, на котором создано произведение.

В качестве экспериментального материала нами выбрана коллекция $\boldsymbol{C}$ из 10 произведений (текстов), среди которых

на белорусском языке:

Л. Станислав «Салярыс, часть 1» (be_1, 8497 *слов*);

Л. Станислав «Салярыс, часть 2» (be_2, 7041 *слово*);

на болгарском языке:

Н. Райнов «Неволя и богатство» (bo_1, 2565 *слов*);

Н. Райнов «Търговец и дяволи» (bo_2, 2567 *слов*);

на русском языке:

М.А. Шолохов «Судьба человека» (ru_1, 10891 *слово*);

Ф.А. Абрамов «Алька» (ru_2, 15668 *слов*);

на рушанском языке:

Л. Химатшоев «Даргилмодак» (rs_1, 7141 *слово*);

Б. Асалбеков «Дилдорум бинест» (rs_2, 6286 *слов*);

на таджикском языке:

С. Айни «Аҳмади Девбанд» (tj_1, 7485 *слов*);

С. Турсун «Повести Камони Рустам» (tj_2, 4041 *слово*).

В списке приведены имена авторов, названия произведений на родном языке, обозначения произведений и их размеры в количестве слов. Особенность коллекции в том, что в ней все тексты представлены в кириллической графике с использованием специфических символов ё, і, ў, ц, ы, ь – в белорусском, ё, ц, щ, ы, ь, ъ, э – в русском, ӑ, ӗ, ғ, ӷ, ҙ, ӣ, қ, ӧ, ҭ, ӯ, ү̊, ҳ̆, ц, ҷ, ҕ – в рушанском, ғ, ё, ӣ, қ, ӯ, ҳ, ҷ, ъ, э – в таджикском и ц, щ, ъ, ь – в болгарском языках.

Из 33 букв кириллицы современного русского языка [272] общими для всех текстов являются 26, именно: а, б, в, г, д, е, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ч, ш, ю, я.

Задача, которая представляет интерес, заключается в том, чтобы определить, насколько важными являются дополнения национальных алфавитов специфическими символами и возможно ли обойтись только лишь общими кириллическими буквами для автоматического распознавания языка, на котором написана та или иная печатная продукция.

1. ЦП произведений. В качестве учётных элементов для описания произведений взяты указанные для всех 5 языков 26 букв.

Определение 1. *ЦП текста будем называть распределение в нём частотности 26 букв.*

ЦП текста T представляется в табличном виде:

$$\begin{array}{lccccc} N: & 1 & 2 & . & . & . & 26 \\ P: & p_1 & p_2 & . & . & . & p_{26}, \end{array}$$

в котором первая строка – номера букв, расположенных в алфавитном порядке, а вторая – относительные частоты встречаемости букв в тексте T, причём $\sum_{k=1}^{26} p_k = 1$.

ЦП представляется также в виде дискретной функции

$$F(s) = \sum_{k=1}^{s} p_k \qquad (s = 1, \cdots, 26). \tag{3.36}$$

2. Расстояния между ЦП текстов. Пусть T_1, T_2 – произвольная пара

текстов, характеризуемых на основе единого алфавита, и

$$F^{(\alpha)}(s) \;=\; \sum_{k=1}^{s} p_k^{(\alpha)} \;- \tag{3.37}$$

соответствующие им ЦП, представленные дискретными функциями, $\alpha = 1, 2,$ и $s = 1, \cdots, 26.$

Определение 2. *Расстоянием между текстами* T_1 *и* T_2 *называется положительное число* $\rho\,(T_1, T_2)$*, определяемое по формуле*

$$\rho\,(T_1\,, T_2) = \sqrt{\frac{26}{2}}\; max_{\,s}\left|F^{(1)}(s) - \;F^{(2)}(s)\right|. \tag{3.38}$$

3. Гипотеза $\mathbb{H}$ **«однородности» произведений** привлекается для того, чтобы выделить характерную особенность текстов, предназначенную для построения математической модели распознавания языка произведений. Она формулируется в следующем виде.

ГИПОТЕЗА $\mathbb{H}$. *Произведения, написанные на одном языке, «однородны», а на разных языках – «неоднородны».*

Под «однородностью» произведений (текстов) мы понимаем их похожесть, одинаковость, сходность, однотипность, родственность и т.п.

4. Математическая модель $\mathbb{H}$**-гипотезы.** Пусть γ-некоторое положительное число.

Определение 3. *Тексты* $T_1,\ T_2$ *называются* γ*-однородными (написанными на одном языке), если*

$$\rho\,(T_1\,, T_2) \leq\; \gamma, \tag{3.39}$$

и γ*-неоднородными (написанными на разных языках), если*

$$\rho\,(T_1\,, T_2) > \gamma. \tag{3.40}$$

Неравенства (3.39) и (3.40) являются математической моделью гипотезы $\mathbb{H}$.

Определение 4. γ*-классификатор – зависящий от одного вещественного параметра* γ *алгоритм принятия решения об отнесении пары текстов* T_1 *и* T_2 *к одному или двум разным языкам.*

Очевидно, что от значения γ зависит однородность или неоднородность любой пары текстов, следовательно, и степень выполнимости гипотезы. Принадлежность двух текстов одному языку в рамках математической модели означает справедливость неравенства (3.39), а двум разным языкам – справедливость неравенства (3.40). Гипотеза $\mathbb{H}$ может нарушаться для каких-то пар текстов одного и того же языка в случае, когда вместо неравенства (3.39) имеет место неравенство (3.40), а также в случае, когда какие-то два текста на

разных языках удовлетворяют неравенству (3.39) вместо того, чтобы выполнялось неравенство (3.40).

Обозначим через $\tau = \tau(\gamma)$ – суммарное количество нарушений гипотезы $\mathbb{H}$ одновременно в двух случаях: невыполнения неравенства «однородности» в случае двух текстов, принадлежащих одному языку, и невыполнения неравенства «неоднородности» в случае двух текстов, принадлежащих разным языкам. Тогда для фиксированного γ показатель выполнения гипотезы будет определяться величиной π, задаваемой формулой

$$\pi = 1 - \tau(\gamma)/L\,,$$

где L – число взаимных расстояний между всеми парами текстов из коллекции $\boldsymbol{C}$ (в нашем случае $L = C_{10}^2 = 45$). Из этой формулы следует, что π может принимать значения из отрезка [0, 1], причём $\pi = 0$, если $\tau = L\,(= 45)$, и $\pi = 1$, если $\tau = 0$. В первом случае гипотезу $\mathbb{H}$ следует признать непригодной, а во втором – полностью согласованной с обучающей выборкой.

В связи с тем, что эффективность γ-классификатора зависит от значения параметра γ, представляет интерес найти такое его значение, при котором π принимает максимальное значение. Именно в этом и заключается суть настройки γ-классификатора на данных обучающей выборки. Если такая настройка будет приемлемой, то можно говорить о решении задачи обучения γ-классификатора и его предрасположенности к распознаванию языков произведений самых разнообразных коллекций.

5. Итоговые результаты на примере модельной коллекции $\boldsymbol{C}$ приведены далее путём последовательного выполнения следующих операций:

– вычисления ЦП (частотности букв алфавита $\boldsymbol{A}$) для всех 10 произведений модельной коллекции $\boldsymbol{C}$;

– вычисления по формулам (3.36), (3.37) и (3.38) сорока пяти парных расстояний $\rho\,(T_1, T_2)$ между произведениями коллекции $\boldsymbol{C}$ (результаты расчетов приведены в следующей таблице):

Таблица 3.24. – Расстояния между текстами коллекции $\boldsymbol{C}$

Тексты		Be		Bo		Ru		Rs		Tj	
		be_1	be_2	bo_1	bo_2	ru_1	ru_2	rs_1	rs_2	tj_1	tj_2
Be	be_1										
	be_2	**0.0511**									
Bo	bo_1	0.3071	0.2930								
	bo_2	0.2683	0.2970	**0.0992**							
Ru	ru_1	0.3841	0.4128	0.3148	0.2701						
	ru_2	0.2881	0.3167	0.2220	0.1773	**0.1232**					
Rs	rs_1	0.2655	0.2679	0.2143	0.2738	0.4092	0.3084				
	rs_2	0.2426	0.2450	0.2820	0.2604	0.3762	0.2801	**0.0747**			
Tj	tj_1	0.3420	0.2909	0.3679	0.4256	0.5771	0.4763	0.2080	0.2770		
	tj_2	0.2956	0.2445	0.3274	0.3850	0.5365	0.4358	0.1458	0.2147	**0.0807**	

– вычисление с помощью алгоритма настройки γ-классификатора [271] оптимального интервала значений γ, для которого величина $\tau = \tau(\gamma)$ суммарного числа случаев нарушения гипотезы $\mathbb{H}$ достигает минимального значения и, следовательно, величина π показателя выполнения гипотезы $\mathbb{H}$ принимает максимальное значение.

По данным таблицы 3.24 получены следующие результаты:

– совокупность всех пар расстояний размещается на отрезке [0.0511, 0.5771], при этом минимальное расстояние реализуется между произведениями на белорусском языке, а максимальное – между романами С. Айни «Аҳмади Девбанд» на таджикском языке и М.А. Шолохов «Судьба человека» на русском языке;

– оптимальный полуинтервал значений γ оказывается в пределах

$$\gamma^{\text{опт}} \in [0.1232;\ 0.1458);$$

в соответствии с определением 3 это значит, что если расстояние $\rho\,(T_1\,, T_2)$ между двумя текстами не превосходит значение $\gamma^{\text{опт}}$ из указанного полуинтервала, то пара текстов принадлежат одному и тому же языку; если же превосходит, то принадлежат разным языкам;

– отметим, что для всех (без исключения) произведений коллекции $\boldsymbol{C}$ полностью подтвердилась гипотеза $\mathbb{H}$ и её математическая интерпретация в виде определения 3, и потому получено

$$\tau = \tau_{min} = 0,$$

то есть, ни одно из неравенств (3.39) и (3.40) не было нарушено;

– вследствие чего показатель эффективности предложенной в настоящей работе математической модели распознавания языка произведений оказался равным

$$\pi = \pi_{max} = 1.$$

6. Тестирование. Итак, настройка (обучение) γ-классификатора на данных модельной коллекции текстов C прошла успешно. Для тестирования классификатора случайным образом выбраны следующие тексты:

на белорусском языке:

С. Давидович «Дзед-кіёк» (be_3, 1935 *слов*);

на болгарском языке:

Н. Райнов «Майчина грижа» (bo_3, 2644 *слова*);

на русском языке:

А.П. Гайдар «Голубая чашка» (ru_3, 6740 *слов*);

на рушанском языке:

А. Сайфиддин «Баҳор» (rs_3, 4683 *слова*);

на таджикском языке:

С. Улуғзода «Бежан ва Манижа» (tj_3, 19730 *слов*).

После формирования ЦП произведений и вычисления расстояний по формуле (3.38) получена следующая таблица расстояний от каждого из текстов до всех 10 произведений исходной коллекции:

Таблица 3.25. – Расстояния между текстами коллекции и тестируемыми произведениями

Тексты		**be_3**	**bo_3**	**ru_3**	**rs_3**	**tj_3**
Be	**be_1**	**0.1569**	0.3150	0.3155	0.2572	0.2640
	be_2	**0.1143**	0.2935	0.3467	0.2596	0.2260
Bo	**bo_1**	0.2654	**0.0646**	0.2696	0.2564	0.2854
	bo_2	0.3118	**0.1264**	0.2249	0.2758	0.3431
Ru	**ru_1**	0.4589	0.3554	**0.0724**	0.3916	0.4946
	ru_2	0.3581	0.2627	**0.0728**	0.2955	0.3938
Rs	**rs_1**	0.2353	0.2148	0.3452	**0.0664**	0.1519
	rs_2	0.2675	0.2557	0.3108	**0.0423**	0.2208
Tj	**tj_1**	0.2635	0.3171	0.5131	0.2415	**0.1145**
	tj_2	0.2012	0.2766	0.4726	0.1829	**0.1060**

В таблице 3.25 серым цветом выделены пары ячеек, которые соответствуют минимальным расстояниям от тестируемых объектов до элементов коллекции текстов. Полученные результаты показывают, что ближайшими соседями выбранной пятёрки произведений оказались как раз однородные с ними по языку пары текстов исходной коллекции.

Заключение. γ-классификатор с фиксированным значением $\gamma = \gamma^{\text{опт}}$ был протестирован на случайных выборках текстов и подтвердил 100%-ную способность к распознаванию языка.

Результаты данного параграфа опубликованы в [56-А].

§ 3.2.5. Тестирование γ-классификатора, настроенного на определение языков произведений на основе латинского алфавита

В данном параграфе на примере модельной коллекции из 10 текстов на пяти языках (английском, немецком, испанском, итальянском и французском) с использованием латинской графики устанавливается применимость γ-классификатора для автоматического распознавания языка произведения на основе частотности общих 26 латинских алфавитных букв. Математическая модель γ-классификатора представляется в виде триады. Её первым компонентом является ЦП текста – распределение в тексте частотности буквенных униграмм; вторым компонентом служит формула для вычисления расстояний между ЦПТ и третьим – алгоритм машинного обучения, реализующий гипотезу «однородности» произведений, написанных на одном языке, и «неоднородности» произведений, написанных на разных языках. Настройка алгоритма, использующего таблицу парных расстояний между всеми произведениями модельной коллекции, заключалась в определении оптимального значения вещественного параметра γ,

для которого минимизируется ошибка нарушения гипотезы «однородности». Для тестирования классификатора было выбрано дополнительно шесть случайных текстов, из которых пять на тех же языках, что и тексты модельной коллекции.

В наше время письменность на основе латинского алфавита получила широкое распространение среди романской, германской славянской, финно-угорской, тюркской, семитской и иранской групп языков, среди стран Индокитая, Зондского архипелага и Филиппин, Африки (южнее Сахары), Америки, Австралии и Океании, [273]. За исключением современного английского, для большинства других языков латинский алфавит из 26 букв (a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z) оказался недостаточным, в связи с чем для отражения фонетических особенностей тех или иных языковых систем к базовой латинской графике были добавлены различные диакритические знаки, лигатуры и другие модификации букв.

Задача, решением которой будем заниматься в данном параграфе, состоит о том, возможно ли обойтись только лишь 26 латинскими буквами для автоматического распознавания языка, на котором написана та или иная печатная продукция.

В качестве экспериментального материала, на котором развертывается наше исследование, выбрана небольшая коллекция ***C*** из 10 произведений (текстов), среди которых

на английском языке **(En):**

У. Шекспир «Romeo and Juliet» (Ромео и Джульетта, **en_1**, 25832 *слова*),

М. Твейн «A Connecticut Yankee in King Arthur's Court» (Янки из Коннектикута при дворе короля Артура, **en_2**, 117257 *слов*);

на немецком языке **(De):**

Г. Пиз «Schiff ohne Mannschaft» (Корабль без экипажа, **de_1**, 59695 *слов*),

Г. Диана «Das flammende Kreuz: Roman» (Пылающий Крест: Роман, **de_2**, 70104 *слова*);

на испанском языке **(Es):**

Д.Дж. Генрих «El ocaso de la magia» (Сумерки магии, **es_1**, 73300 *слов*),

В.Ф. Альберто «Oceano» (Океан, **es_2**, 103596 *слов*);

на итальянском языке **(It):**

Г. Эд «Elminster: la nascita di un mago» (Эльминстер: рождение волшебника, **it_1**, 127087 *слов*),

С. Роберт «Il paradosso del passato» (Парадокс прошлого, **it_2**, 69697 *слов*);

и на французском языке **(Fr):**

С. Жорж «Lavinia» (Лавиния, **fr_1**, 13151 *слово*),

Б. Мишель «Les Nymphéas noirs» (Черные водяные лилии, **fr_2**, 108137 *слов*).

Отметим, что сведения о текстах содержат имена авторов, названия их произведений в оригинале и в переводе на русский язык, а также сокращенные

обозначения произведений совместно с их размерами, определяемыми количеством слов. Особенность коллекции в том, что она охватывает всего лишь 5 европейских языков и все её тексты – на основе латинской графики с использованием дополнительных специфических символов: 4-х ä, ö, ß, ü – в немецком (**de**), одного символа ñ – в испанском (**es**), 10-и символов à, è, é, ì, í, î, ò, ó, ù, ú – в итальянском (**it**) и 14-и символов â, à, ç, é, ê, è, ë, î, ï, ô, û, ù, ü, ÿ – во французском (**fr**) языках.

Приступая к решению поставленной задачи, отметим, что в качестве исследовательского инструмента мы будем использовать *математическую триаду* в составе ЦП текстов, представляемых распределениями частотности 26 латинских букв, формулы для вычисления расстояний между текстам и алгоритма для выявления однородных текстов, см. §§ 1.3 и 1.4. Упомянутая триада с момента своего появления в 2017 году применялась, прежде всего, для распознавании авторства для различных вариантов ЦП текстов, [255-324]. В дополнение к сказанному уместно отметить, что в монографии [227] представлен обширный обзор работ по идентификации авторов текста на основе разнообразных ЦП текстов и применяемых методов классификации.

1. ЦПП. В качестве учётных элементов для описания произведений взяты указанные для всех 5 языков 26 латинских букв.

Определение 1. *ЦП текста будем называть распределение в нём частотности 26 букв.*

ЦП текста T записывается в табличном виде:

$$\begin{array}{ll} N: & 1 \quad 2 \quad \ldots \quad 26 \\ P: & p_1 \quad p_2 \quad \ldots \quad p_{26}, \end{array} \tag{3.41}$$

в котором первая строка – номера букв, расположенных в алфавитном порядке, а вторая – относительные частотности букв в тексте T, причём $\sum_{k=1}^{26} p_k = 1$.

Одновременно с (3.41) ЦП представляется также в виде дискретной функции

$$F(s) = \sum_{k=1}^{S} p_k \qquad (s = 1, \ldots, 26). \tag{3.42}$$

2. Расстояния между ЦП текстов. Пусть T_1, T_2 – произвольная пара текстов, характеризуемых на основе единого алфавита, и

$$F^{(\alpha)}(s) = \sum_{k=1}^{S} p_k{}^{(\alpha)} \quad - \tag{3.43}$$

соответствующие им ЦП, представленные дискретными функциями, $\alpha = 1, 2,$ и $(s = 1, \ldots, 26)$.

Определение 2. *Расстоянием между текстами* T_1 *и* T_2 *называется положительное число* $\rho(T_1, T_2)$*, определяемое формулой*

$$\rho(T_1, T_2) = \sqrt{\frac{26}{2}} \max_s \left| F^{(1)}(s) - F^{(2)}(s) \right|. \qquad (3.44)$$

3. Гипотеза Н **«однородности» произведений.** Она привлекается для того чтобы выделить характерную особенность текстов, предназначенную для построения математической модели распознавания языка произведений. Её мы формулируем в следующем виде.

ГИПОТЕЗА Н. *Произведения, написанные на одном языке, «однородны», а на разных языках – «неоднородны».*

Говоря об «однородности» произведений (текстов), мы имеем в виду их похожесть, одинаковость, сходность, однотипность, родственность и т.п.

4. Математическая модель Н**-гипотезы.** Пусть γ - некоторое положительное число.

Определение 3. *Тексты* T_1, T_2 *называются* γ*-однородными (написанными на одном языке), если*

$$\rho(T_1, T_2) \le \gamma, \qquad (3.45)$$

и γ *-неоднородными (написанными на разных языках), если*

$$\rho(T_1, T_2) > \gamma. \qquad (3.46)$$

Неравенства (3.45) и (3.46) являются математической интерпретацией (моделью) гипотезы Н. Это значит, что в дальнейшем мы приступаем к распознаванию языков произведений с помощью математического аппарата, названного γ-классификатором, см. § 1.4.

Определение 4. γ*-классификатор – зависящий от одного вещественного параметра* γ *алгоритм принятия решения об отнесении пары текстов* T_1 *и* T_2 *к одному или двум разным языкам.*

Очевидно, что от значения γ зависит однородность или неоднородность любой пары текстов, следовательно, и степень выполнимости гипотезы. Принадлежность двух текстов одному языку в рамках математической модели означает справедливость неравенства (3.45), а двум разным языкам – справедливость неравенства (3.46). Гипотеза Н может нарушаться для каких-то пар текстов одного и того же языка в случае, когда вместо неравенства (3.45) имеет место неравенство (3.46), а также в случае, когда какие-то два текста на разных языках удовлетворяют неравенству (3.45) вместо того, чтобы выполнялось неравенство (3.46).

Пусть $\tau = \tau(\gamma)$ – суммарное количество нарушений гипотезы Н одновременно в двух случаях: невыполнения неравенства «однородности» в случае двух текстов, принадлежащих одному языку, и невыполнения неравенства «неоднородности» в

случае двух текстов, принадлежащих разным языкам. Тогда для фиксированного γ показатель выполнения гипотезы будет определяться величиной π, задаваемой формулой

$$\pi = 1 - \tau(\gamma)/L,$$

где L – число взаимных расстояний между всеми парами текстов из коллекции ***C*** (в нашем случае $L = C_{10}^2 = 45$. Из этой формулы следует, что π может принимать значения из отрезка [0, 1], причём $\pi = 0$, если $\tau = L(= 45)$, и $\pi = 1$, если $\tau = 0$. В первом случае гипотезу H следует признать непригодной, а во втором – полностью согласованной с обучающей выборкой.

В связи с тем, что эффективность γ-классификатора зависит от значения параметра γ, представляет интерес найти такое его значение, при котором π принимает максимальное значение. Именно в этом и заключается суть настройки γ-классификатора на данных обучающей выборки. Если такая настройка будет приемлемой, то можно говорить о решении задачи обучения γ-классификатора и его предрасположенности к распознаванию языков произведений самых разнообразных коллекций.

5. Настройка классификатора на данных коллекции *C*. В процессе настройки выполняются следующие операции:

– предобработка экспериментальных данных путём удаления из всех произведений коллекции дополнительных сверх латинского алфавита букв;

– вычисления ЦП (3.41) (частотности 26 латинских букв) для всех 10 произведений модельной коллекции ***C***;

– вычисления по формулам (3.42), (3.43) и (3.44) сорока пяти парных расстояний $\rho(T_1, T_2)$ между произведениями коллекции ***C*** (результаты расчетов приведены в следующей таблице):

Таблица 3.26. – Расстояния между текстами коллекции ***C***

Тексты		**En**		**De**		**Es**		**It**		**Fr**	
		en_1	**en_2**	**de_1**	**de_2**	**es_1**	**es_2**	**it_1**	**it_2**	**fr_1**	**fr_2**
En	**en_1**										
	en_2	**0.0832**									
De	**de_1**	0.3949	0.3281								
	de_2	0.3817	0.3148	**0.0287**							
Es	**es_1**	0.3606	0.3030	0.2845	0.2963						
	es_2	0.3471	0.2895	0.3077	0.2945	**0.0450**					
It	**it_1**	0.2486	0.2302	0.2677	0.2579	0.1950	0.1814				
	it_2	0.2426	0.2243	0.2988	0.2928	0.2086	0.1951	**0.0378**			
Fr	**fr_1**	0.1354	0.1945	0.3982	0.3849	0.3205	0.2941	0.2628	0.2691		
	fr_2	0.1480	0.1833	0.4038	0.3920	0.3260	0.2995	0.2776	0.2773	**0.0299**	

– вычисление с помощью алгоритма настройки γ-классификатора

оптимального интервала значений γ, для которого величина $\tau = \tau(\gamma)$ суммарного числа случаев нарушения гипотезы H достигает минимального значения и, следовательно, величина π показателя выполнения гипотезы H принимает максимальное значение.

По данным таблицы 3.26 получены следующие результаты:

– совокупность всех пар расстояний размещается на отрезке [0.0287, 0.4038], при этом минимальное расстояние реализуется между двумя произведениями **de_1** и **de_2** на немецком языке, а максимальное – между **de_1** на немецком и **fr_2** на французском языках;

– оптимальный полуинтервал значений γ оказывается в пределах

$$\gamma^{onm} \in [0.0833;\ 0.1353); \qquad (3.47)$$

в соответствии с определением 3 это значит, что если расстояние $\rho(T_1, T_2)$ между двумя текстами не превосходит значение γ^{onm} из указанного полуинтервала, то пара текстов принадлежит одному и тому же языку (соответствующие расстояния в таблице помечены серым цветом); если же превосходит, то принадлежат разным языкам (соответствующие расстояния оставлены непомеченными);

– отметим, что для всех (без исключения) произведений коллекции ***C*** полностью подтвердилась гипотеза H и её математическая интерпретация в виде определения 3, и потому получено

$$\tau = \tau_{\min} = 0,$$

то есть ни одно из неравенств (3.45) и (3.46) не было нарушено;

– вследствие чего показатель эффективности предложенной в настоящей работе математической модели распознавания языка произведений оказался равным

$$\pi = \pi_{\max} = 1.$$

6. Тестирование. Итак, настройка (обучение) γ-классификатора на данных модельной коллекции текстов ***C*** прошла успешно. Для тестирования классификатора выбрано случайным образом 6 текстов:

на английском языке **(En):**

Дж. Лондон «The Call of the Wild» (Зов предков) (Text_En, 31763 *слова*);

на немецком языке **(De):**

М. Вилли «Die seltsamen Reisen des Marco Polo» (Странные путешествия Марко Поло) (Text_De, 126607 *слов*);

на испанском языке **(Es):**

Д. Арне «Misterioso» (Таинственный) (Text_Es, 106835 *слов*);

на итальянском языке **(It):**

Ш. Боб «Sfida al cielo» (Вызов небу) (Text_It, 101154 *слова*);

на французском языке **(Fr):**

К.С. Доминикович «Fantôme» (Призрак) (Text_Fr, 46089 *слов*);

и на румынском языке **(Ro):**

Т.Р. Руэл «Întoarcerea regelui» (Возвращение короля) (Text_Ro, 146266 *слов*).

Отметим, что сведения относительно выбранных произведений описаны по той же схеме, что и для элементов коллекции ***C***. В дополнение к предыдущему отметим, что в румынском языке (**ro**) латинский алфавит расширен на 5 символов, ă, â, î, ș. ț.

Для шести произведений, предназначенных для тестирования, построены цифровые портреты (3.41) и затем по формулам (3.42), (3.43), (3.44) для каждого из них вычислены расстояния до 10 объектов коллекции ***C***. Соответствующие значения записаны в ячейках таблицы 3.27, расположенных на пересечениях строк и столбцов. Там же серым цветом выделены пары ячеек с минимальными расстояниями между тестируемыми и содержащимися в коллекции произведениями.

Таблица 3.27. – Расстояния между текстами коллекции ***C*** и шестью случайно выбранными тестируемыми произведениями

Тексты		**Text_En**	**Text_De**	**Text_Es**	**Text_It**	**Text_Fr**	**Text_Ro**
En	**en_1**	**0.1592**	0.4069	0.3235	0.2378	0.1477	0.2084
	en_2	**0.0857**	0.3400	0.2659	0.2194	0.1905	0.1760
De	**de_1**	0.2599	**0.0305**	0.2659	0.2866	0.4235	0.2723
	de_2	0.2467	**0.0489**	0.2526	0.2734	0.4103	0.2663
Es	**es_1**	0.2674	0.3010	**0.0552**	0.1874	0.3250	**0.1707**
	es_2	0.2538	0.3197	**0.0430**	0.1738	0.2985	**0.1440**
It	**it_1**	0.1987	0.2882	0.1579	**0.0330**	0.3050	**0.1565**
	it_2	0.2365	0.3260	0.1715	**0.0281**	0.3047	**0.1563**
Fr	**fr_1**	0.2802	0.4101	0.2712	0.2501	**0.0460**	0.1933
	fr_2	0.2690	0.4158	0.2767	0.2604	**0.0448**	0.2033

Полученные результаты показывают, что ближайшими соседями [248, 249] первых пяти произведений оказались как раз однородные с ними по языку пары текстов исходной коллекции. Что касается текста на румынском языке (Text_Ro), то все его расстояния до десяти коллекционных текстов превзошли максимальное значение γ^{onm}, см. (3.47). Следовательно, как и ожидалось, для Text_Ro в коллекции не оказалось ни одного однородного объекта. Интересно, однако, отметить, что γ-классификатор указал в качестве её ближайших соседей два произведения es_1 и es_2 на испанском и два произведения it_1 и it_2 на итальянском языках.

Заключение. Итак, γ-классификатор с фиксированным значением $\gamma = \gamma^{onm}$ на случайных выборках текстов с ЦП на основе частотности 26 базовых латинских

букв подтвердил 100%-ную статистическую способность к распознаванию языков произведений.

Таким образом, математическая триада в составе ЦП текстов, представляемых распределениями частотности 26 латинских букв, формул (3.41) – (3.43) для вычисления расстояний между текстами и алгоритма для выявления однородных текстов, оказалась подходящей для эффективного решения поставленной задачи.

Автор выражает уверенность в том, что увеличение объема исходной коллекции текстов не станет препятствием для успешного применения γ-классификатора не только для распознавания языков, но также и для самых разнообразных однородностей текстовых документов.

Результаты данного параграфа опубликованы в [23-А, 47-А].

§ 3.2.6. К вопросу о метрической однородности текстов на славянских языках

В исследованиях Р. Грея и К. Аткинсона [4] посредством статистического анализа родственных слов, У. Чанга, Ч. Кэткарта, Д. Холла и А. Гарретта [5] с помощью статистического моделирования и А.С. Касьяна и А.В. Дыбо [6] на основе лексикостатистической классификации помимо обсуждения исторических вопросов представлены генеологические деревья, отражающие как родство, так и дивергенцию современных славянских языков. Таких деревьев достаточно много, они сходны в общих чертах и различны в небольших деталях, см. например, [6, 274]. Ареал прежде единого языка ныне разделился на три группы – восточную в составе белорусского, русского и украинского языков, западную – из чешского, словацкого, польского, кашубского и лужицких языков и южную, состоящую из болгарского, македонского, сербо-хорватского и словенского языков. В настоящем параграфе на примере случайно сформированной модельной коллекции из 26 текстов на 13 языках (по 2 произведения от каждого языка) устанавливается применимость γ-классификатора для автоматического распознавания принадлежности текстов той или иной группе славянских языков на основе частотности универсального для всех языков набора латинских символов. Математическая модель γ-классификатора представляется в виде триады, составленной из ЦП текста – распределения в тексте частотности латинских символьных унограмм; формулы для вычисления расстояний между ЦП текстами и алгоритма машинного обучения, реализующего гипотезу «однородности» произведений из одной группы языков и «неоднородности» произведений, принадлежащих разным группам языков. Настройка алгоритма, использующего таблицу парных расстояний между всеми произведениями модельной коллекции, осуществлялась путем подбора оптимального значения вещественного параметра γ, минимизирующего число ошибок нарушения

гипотезы «однородности». Обученный на текстах модельной коллекции γ-классификатор тестируется на предмет правильного отнесения случайного текста группе «однородных» с ним произведений. Для тестирования классификатора было выбрано 3 дополнительных случайных текста, по одному тексту для трёх разных групп славянских языков. Методом ближайшего (по расстоянию) соседа проверяется однородность с соответствующими парами одноязычных произведений, тем самым и однородность с соответствующей группой славянских языков.

Состояние работ по применению различных классификаторов, прежде всего методов нейронных сетей и машины опорных векторов, подробно описано в монографии [227]. На примере модельной случайно сформированной коллекции из 26 произведений на 13 славянских языках (по 2 произведения от каждого языка) решаются две задачи:

– путем подбора вещественного параметра γ настроить так называемый γ-классификатор, по возможности, для безошибочного распознавания принадлежности текстов соответствующей одной из трёх групп языков;

– для трёх дополнительных случайно выбранных произведений, принадлежащих различным группам, проверить правильность работы настроенного классификатора.

Прежде чем переходить к изучению задач, напомним основные понятия, связанные с компонентами триады.

I. Модельная коллекция текстов C, собранная случайным образом, представляет три группы славянских языков, причём от каждого языка по два произведения. В приводимом далее списке элементов коллекции C указываются имя автора, название его сочинения на родном языке и в скобках – используемый алфавит, аббревиатура сочинения и его размеры в количестве слов:

а) в восточнославянской группе

на белорусском языке:

Л. Станислав «Салярыс, часть 1» (кир., be_1, 8497 *слов*);

С. Давидович (Be): «Дзед-кіёк» (кир., be_2, 1935 *слов*);

на русском языке:

М.А. Шолохов (Ru): «Судьба человека» (кир., ru_1, 10891 *слово*);

Ф.А. Абрамов (Ru): «Алька» (кир., ru_2, 15668 *слов*);

на украинском языке:

В.Л. Кашин «Готується вбивство» (кир., uk1, 23771 *слово*);

М. Циба (Uk): «Акванавти, або Золота жила» (кир., uk_2, 20150 *слов*);

б) в западнославянской группе

на польском языке:

R.M. Wegner «Jeszcze może załopotać, часть 1» (лат., pl1, 10601 *слово*);

R.M. Wegner «Jeszcze może załopotać, часть 2» (лат., pl2, 9670 *слов*);

на чешском языке:

S. Lem «K Mrakům Magellanovým» (лат., cs1, 17552 *слова*);

B.S.R. Jordan «Bouře přichází» (лат., cs2, 17439 *слов*);

на словацком языке:

I.A. Jefremov «Na hranici Oekumeny» (лат., sv1, 13534 *слова*);

J. Jesenský «Demokrati» (лат., sv2, 17113 *слова*);

на кашубском языке:

D. Pioch «Biuletin Radzëznë Kaszëbsczégò Jãzëka» (лат., ks1, 12070 *слов*);

E. Breza «Prymas z Kaszub» (лат., ks2, 16871 *слово*);

в) в южнославянской группе:

на болгарском языке:

Н. Райнов «Неволя и богатство» (кир., bo1, 2565 *слов*);

Б. Джим (Во): «Фурията на принцепса, глава 1» (кир., bo_2, 2491 *слово*);

на боснийском языке:

И. Асимов «Немезис» (кир., bs1, 20035 *слов*);

Д. Вейнс «Mjесечев мољац» (кир., bs2, 10443 *слова*);

на сербском языке:

А. Кларк «Напеви далеке Земље» (кир., se1, 11129 *слов*);

Р.Л. Стивенсон «Црна стрела» (кир., se2, 15028 *слов*);

на словенском языке:

M. Hudnik «Kakor Kartagina» (лат., sl1, 14626 *слов*);

I. Koprivec «Josip Vidmar v oieh svojih sodobnikov» (лат., sl2, 16985 *слов*);

на македонском языке:

В. Тоциновски «Кочо Рацин – наша творечка и етичка мерка» (кир., mk1, 9047 *слов*);

Г. Прличев «Сердарот» (кир., mk2, 9478 *слов*);

на хорватском языке:

I.M. Andrić «Pročitani Pisci (Eseji i prikazi)» (лат., xr1, 26221 *слово*);

M. Lovrak «Vlak U Snijegu» (лат., xr2, 10522 *слова*).

II. ЦП произведений. В качестве элементов количественного образа произведений нами используются буквенные униграммы. Поскольку для славянских языков нет единого буквенного алфавита (в указанном списке 14 произведений на основе кириллического алфавита и 12 – на основе латинского), мы осуществляем предобработку алфавитов таким образом, чтобы выделить в них унифицированный набор символов. Среди 14 аналогов кириллических алфавитов общими оказались 26 букв: – «а, б, в, г, д, е, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ч, ш, ю, я»; между тем для 12 аналогов латинского алфавита – тоже 26 букв, но уже следующие «a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z». Из этих двух алфавитов был сформирован искусственный общий для всех текстов

алфавит из 22 символов «a, b, c, d, e, f, g, i, j, k, l, m, n, o, p, r, t, u, v, x, y, z», учитывающих сходных по написанию и по звучанию символы.

Теперь, когда хотя бы формально, все тексты описываются одним и тем же набором из 22 латинских символов, введем следующее

Определение 1. *Цифровым портретом какого-либо текста T на славянском языке будем называть распределение в нём частотности упомянутых 22 латинских символов.*

ЦП текста T записывается в табличном виде:

$$\begin{array}{lccccc} N: & 1 & 2 & . & . & . & 22 \\ P: & p_1 & p_2 & . & . & . & p_{22}, \end{array}$$

в котором первая строка – номера символов, расположенных в алфавитном порядке, а вторая – относительные частоты встречаемости символов в тексте T, причём $\sum_{k=1}^{22} p_k = 1$.

ЦП представляется также в виде дискретной функции

$$F(s) = \sum_{k=1}^{s} p_k \quad (s = 1, \cdots, 22). \tag{3.48}$$

III. Расстояния между ЦП текстов. Пусть T_1, T_2 – произвольная пара текстов из $\boldsymbol{C}$, характеризуемых на основе единого символьного алфавита, и

$$F^{(\alpha)}(s) = \sum_{k=1}^{s} p_k^{(\alpha)} - \tag{3.49}$$

соответствующие им ЦП, представленные дискретными функциями, $\alpha = 1, 2,$ и $s = 1, \cdots, 22.$

Определение 2. *Расстоянием между текстами T_1 и T_2 называется положительное число $\rho\,(T_1, T_2)$, определяемое формулой*

$$\rho\,(T_1, T_2) = \sqrt{\frac{22}{2}}\, max_s \left| F^{(1)}(s) - F^{(2)}(s) \right| \tag{3.50}$$

IV. Гипотеза $\mathbb{H}$ «однородности» произведений. Она привлекается для того чтобы выделить характерную особенность текстов, предназначенную для построения математической модели распознавания однородных групп произведений. Её мы формулируем в следующем виде.

ГИПОТЕЗА $\mathbb{H}$. *Любая пара произведений из одной и той же группы славянских языков «однородна», а из разных групп «неоднородна».*

Говоря об «однородности» произведений (текстов), мы имеем в виду их похожесть, одинаковость, сходность, однотипность, родственность и т.п.

V. Математическая модель $\mathbb{H}$-гипотезы. Пусть γ -некоторое

положительное число.

Определение 3. *Тексты* T_1, T_2 *называются* γ*-однородными (принадлежащими одной и той же группе славянских языков), если*

$$\rho\,(T_1\,, T_2) \leq\ \gamma, \tag{3.51}$$

и γ*-неоднородными (принадлежащими разным группам славянских языков), если*

$$\rho\,(T_1\,, T_2) > \gamma. \tag{3.52}$$

Неравенства (3.51) и (3.52) являются математической интерпретацией (моделью) гипотезы $\mathbb{H}$.

Определение 4. γ*-классификатор – это зависящий от одного вещественного параметра* γ *алгоритм принятия решения об отнесении пары текстов* T_1 *и* T_2 *к одной или двум разным группам славянским языкам.*

Очевидно, что от значения γ зависит однородность или неоднородность любой пары текстов, следовательно, и степень выполнимости гипотезы. Принадлежность двух текстов одной группе языков в рамках математической модели означает справедливость неравенства (3.51), а двум разным группам – справедливость неравенства (3.52). Гипотеза $\mathbb{H}$ может нарушаться для каких-то пар текстов одной и той же группе языков в случае, когда вместо неравенства (3.51) имеет место неравенство (3.52), а также в случае, когда какие-то два текста из разных групп удовлетворяют неравенству (3.51) вместо того, чтобы выполнялось неравенство (3.52).

Пусть $\tau = \tau(\gamma)$ – суммарное количество нарушений гипотезы $\mathbb{H}$ одновременно в двух случаях: невыполнения неравенства «однородности» в случае двух текстов, принадлежащих одной группе, и невыполнения неравенства «неоднородности» в случае двух текстов, принадлежащих разным группам. Тогда для фиксированного γ показатель выполнения гипотезы будем определять величиной π (3.53), задаваемой формулой

$$\pi = 1 - \tau(\gamma)/L\,, \tag{3.53}$$

где L – число взаимных расстояний между всеми парами текстов из коллекции $\boldsymbol{C}$ (в нашем случае $L = C_{26}^2 = 325$). Из этой формулы следует, что π может принимать значения из отрезка [0, 1], причём $\pi = 0$, если $\tau = L\,(= 325)$, и $\pi = 1$, если $\tau = 0$. В первом случае гипотезу $\mathbb{H}$ следует признать непригодной, а во втором – полностью согласованной с обучающей выборкой.

В связи с тем, что эффективность γ-классификатора зависит от значения параметра γ, представляет интерес найти такое его значение, при котором π принимает максимальное значение. *Именно в этом и заключается суть*

настройки γ-классификатора на данных обучающей выборки. Если такая настройка будет приемлемой, то можно говорить о решении задачи обучения γ-классификатора и его предрасположенности к распознаванию принадлежности пары произведений одной или же различным группам. Алгоритм настройки классификатора приведен в § 1.4.

VI. Предварительные результаты на примере модельной коллекции $\boldsymbol{C}$

приведены далее путём последовательного выполнения следующих операций:

– вычисления ЦП (частотности букв 22 общих латинских символов) для всех 26 произведений модельной коллекции $\boldsymbol{C}$;

– вычисления по формулам (3.48), (3.49) и (3.50) 325 парных расстояний $\rho\,(T_1, T_2)$ между произведениями коллекции $\boldsymbol{C}$ (результаты расчетов приведены в следующей таблице):

В соответствии с определением 3 это значит, что если расстояние $\rho(T_1\,, T_2)$ между двумя текстами не превосходит значение $\gamma^{\text{опт}} < 0.2160$, то пара текстов принадлежит одной и той же группе языков; если же $\rho(T_1\,, T_2)$ превосходит 0.2160, то принадлежит разным языкам.

Минимальное число нарушений оказалось равным $\tau = 45$. В таблице 3.28 ячейки нарушения гипотезы (3.51) «однородности» отмечены слабо серым цветом, а гипотезы (3.52) «неоднородности» серым цветом.

Теперь остается вычислить эффективность π классификатора по формуле (3.53):

$$\pi = 1 - \tau(\gamma^{\text{опт}})/L = 0.86$$

VII. Тестирование классификатора. После того как за счёт выбора оптимального значения γ произошла настройка классификатора и был отработан алгоритм, который в 86 случаях из 100 правильно соотносил элементы модельной коллекции к соответственной группе славянских языков, возникает естественный вопрос, а каковы будут результаты раскладки уже других славянских текстов, не входящих в коллекцию, по тем же самым трем языковым группам.

Для тестирования классификатора выбрано случайным образом 3 текста:

на украинском языке **(Uk)** – В.П. Бережной «Homo Novus» (кир., Text_Uk, 5768 *слов*);

на польском языке **(Pl)** – A. Szklarski «Tomek wśród łowców głów» (лат.,Text_Pl, 13635 *слов*);

на болгарском языке **(Bo)** – А. Каралийчев «Гулчечек» (кир., Text_Bo, 2436 *слов*).

Таблица 3.28. – Расстояния между текстами коллекции $\boldsymbol{C}$

Тексты		Восточнославянская подгруппа						Западнославянская подгруппа								Южнославянская подгруппа											
		be1	be2	ru1	ru2	uk1	uk2	pl1	pl2	cs1	cs2	sv1	sv2	ks1	ks2	bo1	bo2	bs1	bs2	se1	se2	sl1	sl2	mk1	mk2	xr1	xr2
Вост.	be1																										
	be2	0.13																									
	ru1	0.36	0.45																								
	ru2	0.27	0.35	0.09																							
	uk1	0.39	0.51	0.17	0.25																						
	uk2	0.36	0.47	0.13	0.21	0.04																					
Запад.	pl1	0.36	0.39	0.29	0.27	0.26	0.24																				
	pl2	0.33	0.36	0.28	0.26	0.28	0.25	0.03																			
	cs1	0.40	0.43	0.15	0.14	0.24	0.21	0.25	0.27																		
	cs2	0.34	0.37	0.13	0.12	0.27	0.24	0.21	0.23	0.06																	
	sv1	0.30	0.33	0.15	0.12	0.28	0.25	0.22	0.22	0.11	0.07																
	sv2	0.29	0.32	0.14	0.07	0.30	0.26	0.24	0.23	0.13	0.09	0.05															
	ks1	0.37	0.40	0.31	0.29	0.30	0.26	0.11	0.09	0.25	0.23	0.20	0.22														
	ks2	0.37	0.40	0.25	0.23	0.28	0.24	0.04	0.04	0.25	0.22	0.18	0.20	0.09													
Южн.	bo1	0.20	0.27	0.28	0.22	0.36	0.33	0.35	0.34	0.35	0.31	0.24	0.23	0.34	0.31												
	bo2	0.20	0.29	0.23	0.17	0.32	0.28	0.31	0.30	0.30	0.26	0.18	0.18	0.30	0.26	0.13											
	bs1	0.22	0.28	0.30	0.24	0.37	0.34	0.37	0.36	0.37	0.33	0.26	0.25	0.39	0.33	0.09	0.11										
	bs2	0.27	0.34	0.25	0.19	0.32	0.28	0.36	0.35	0.32	0.28	0.21	0.20	0.37	0.32	0.11	0.09	0.10									
	se1	0.23	0.30	0.27	0.21	0.35	0.31	0.36	0.35	0.34	0.30	0.22	0.22	0.38	0.32	0.09	0.09	0.05	0.08								
	se2	0.27	0.32	0.30	0.24	0.37	0.34	0.35	0.34	0.37	0.33	0.26	0.25	0.37	0.31	0.11	0.09	0.06	0.06	0.05							
	sl1	0.30	0.38	0.31	0.25	0.27	0.26	0.36	0.35	0.33	0.30	0.25	0.23	0.36	0.32	0.14	0.14	0.10	0.11	0.09	0.11						
	sl2	0.35	0.43	0.29	0.23	0.27	0.26	0.31	0.30	0.31	0.28	0.23	0.21	0.33	0.27	0.18	0.17	0.15	0.10	0.14	0.16	0.05					
	mk1	0.22	0.31	0.23	0.17	0.30	0.27	0.35	0.34	0.30	0.26	0.20	0.18	0.35	0.31	0.21	0.08	0.17	0.13	0.14	0.17	0.20	0.19				
	mk2	0.16	0.23	0.29	0.23	0.40	0.36	0.37	0.36	0.36	0.32	0.25	0.24	0.39	0.33	0.09	0.09	0.06	0.11	0.08	0.12	0.15	0.20	0.13			
	xr1	0.31	0.39	0.35	0.29	0.30	0.27	0.36	0.35	0.37	0.33	0.28	0.27	0.38	0.32	0.14	0.13	0.15	0.12	0.12	0.15	0.07	0.05	0.20	0.17		
	xr2	0.24	0.29	0.40	0.33	0.35	0.32	0.38	0.37	0.41	0.38	0.33	0.31	0.40	0.34	0.12	0.18	0.14	0.16	0.15	0.12	0.09	0.14	0.26	0.13	0.11	

– вычисление с помощью алгоритма настройки γ-классификатора из §§ 1.3-1.4 оптимального интервала значений γ, для которого величина $\tau = \tau(\gamma)$ суммарного числа случаев нарушения гипотезы $\mathbb{H}$ достигает минимального значения и, следовательно, величина π показателя выполнения гипотезы $\mathbb{H}$ принимает максимальное значение.

По данным таблицы 3.28 вычислен оптимальный полуинтервал значений γ

$$\gamma^{\text{опт}} \in [0.2142;\ 0.2160)$$

Для каждого произведения так же, как это было сделано для всех текстов модельной коллекции, построены ЦП на основе единого набора из 22 латинских символов. После чего по формуле (3.50) вычислены расстояния до всех 26 элементов модельной коллекции. Результаты показаны в таблице.

Таблица 3.29. – Расстояния между текстами коллекции ***C*** и тремя случайно выбранными произведениями

Тексты		**Text_Uk**	**Text_Pl**	**Text_Bo**
Восточная группа	**be1**	0.3421	0.3432	0.2031
	be2	0.4490	0.3742	0.2926
	ru1	0.1034	0.2699	0.2131
	ru2	0.1896	0.2517	0.1515
	uk1	0.0714	0.2398	0.3297
	uk2	**0.0511**	0.2190	0.2912
Западная группа	**pl1**	0.1612	0.1916	0.2013
	pl2	0.1791	0.2030	0.1836
	cs1	0.1856	0.2070	0.2844
	cs2	0.2125	0.1745	0.2445
	sv1	0.2238	0.2010	0.2090
	sv2	0.2391	0.2162	0.1619
	ks1	0.2347	0.1271	0.3233
	ks2	0.2158	**0.0862**	0.2946
Южная группа	**bo1**	0.3014	0.3389	0.1064
	bo2	0.2578	0.2901	**0.0510**
	bs1	0.3165	0.3521	0.1331
	bs2	0.2560	0.3386	0.1035
	se1	0.2918	0.3407	0.1115
	se2	0.3129	0.3303	0.1086
	sl1	0.2192	0.3418	0.1403
	sl2	0.2049	0.2971	0.1822
	mk1	0.2458	0.3232	0.1206
	mk2	0.3350	0.3500	0.0936
	xr1	0.2441	0.3419	0.1533
	xr2	0.2921	0.3661	0.2013

В ячейках таблицы на пересечении столбцов и строк приводятся значения расстояний между текстами. В первых трех столбцах ближайшими соседями текстов Text_Uk, Text_Pl и Text_Bo являются соответственно uk2, ks2 и bo2 на расстояниях соответственно 0.0511, 0.0862 и 0.0510 (в таблице отмечены серым цветом). Полученный результат показывает, что по методу ближайшего соседа три случайно выбранных произведения распределяются как раз по тем группам языков, которым они сами принадлежат.

Заключение. Итак, γ-классификатор с фиксированным значением $\gamma = \gamma^{\text{опт}}$ на случайных выборках текстов с ЦП на основе частотности 22 латинских символов подтвердил 86%-ную статистическую способность к распознаванию групп произведений на славянских языках. В свою очередь, метод ближайшего соседа показал возможность безошибочного распределения дополнительных

славянских произведений по восточной, западной и южной группам славянских языков.

Результаты данного параграфа опубликованы в [59-А, 61-А].

§ 3.3. Об однородности оригинала и его перевода

На примере модельной коллекции текстов на русском и таджикском языках и их переводов на таджикский и русский языки с помощью γ-классификатора и ЦП, характеризующих в текстах распределения частотности буквенных униграмм, исследуется статистическая «однородность» оригинальных и переводных произведений.

С точки зрения математического моделирования задачи проектирования автоматических систем распознавания плагиата, заимствования, авторства текстовых фрагментов, произведения и его перевода представляет собой грани единого целого, которые различаются, прежде всего, ЦП, формирующими количественный образ объектов исследований.

В данном параграфе таковым объектом является взаимосвязь двух типов творческой продукции – печатного произведения (оригинала) и его перевода (модели). Переводчик, моделирующий оригинальный текст, отражает в своей модели те или иные свойства оригинала и тем самым между двумя объектами устанавливает в некотором смысле однородность (нечто вроде единства, родства, подобия, сходства и т.п.).

Следующее утверждение представляется вполне естественным отражением взаимосвязи оригинала со своим переводом.

ГИПОТЕЗА $\mathbb{H}$. *Произведения и переводы произведений одного автора – «однородные», а разных авторов – «неоднородные».*

Для количественного описания «однородности» воспользуемся следующими понятиями, см. § 1.3.

Определение 3.3.1. ***Алфавит*** *– упорядоченное множество элементов текста.*

Примерами элементов текста могут служить буквы алфавита естественного языка, буквенные *N*-граммы и слоги, словоформы и словосочетания и многое другое. Для наших целей в качестве элементов алфавита будут использоваться упорядоченные по алфавиту буквенные униграммы.

Определение 3.3.2. ***Цифровым портретом текста*** *называется распределение частотности элементов алфавита.*

В нашем случае цифровым портретом произведений и их переводов является распределение частотностей буквенных униграмм. В качестве учётных элементов для описания произведений взяты:

– 33 буквы русского алфавита (а, б, в, г, д, е, ё, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ц, ч, ш, щ, ы, ь, ъ, э, ю, я) и

– 35 букв таджикского алфавита (а, б, в, г, ғ, д, е, ё, ж, з, и, ӣ, й, к, қ, л, м, н, о, п, р, с, т, у, ӯ, ф, х, ҳ, ч, ҷ, ш, ъ, э, ю, я).

Для количественного описания всех произведений сформирован единый алфавит из 39 букв (а, б, в, г, ғ, д, е, ё, ж, з, и, ӣ, й, к, қ, л, м, н, о, п, р, с, т, у, ӯ, ф, х, ҳ, ч, ҷ, ш, ъ, э, ю, я, ц, щ, ы, ь), которым в таблице ASCII сопоставлены уникальные числовые коды.

Говорят, что на множестве X элементов произвольной природы введено понятие *расстояния*, если каждой паре элементов x_1 и $x_2 \in X$ соотнесено вещественное число $\rho(x_1, x_2)$ такое, что

1. $\rho(x_1, x_2) \geq 0,\ \rho(x_1, x_2) = 0$ при $x_1 = x_2$ и если $x_1 = x_2$, то $\rho(x_1, x_2) = 0$,
2. $\rho(x_1, x_2) = \rho(x_2, x_1)$.

Пусть γ -некоторое положительное число.

Определение 3.3.3. *Пару текстов* T_1 *и* T_2 (будь то оригиналы или переводы) *назовём* ***γ- однородными****, если*

$$\rho\,(T_1\,, T_2) \leq \gamma, \tag{3.54}$$

и ***γ- неоднородными****, если*

$$\rho\,(T_1\,, T_2) > \gamma. \tag{3.55}$$

Неравенства (3.54), (3.55) по-существу являются математической моделью гипотезы $\mathbb{H}$[4]. Очевидно, что от значения γ зависит однородность или неоднородность любой пары текстов, а потому и степень выполнимости гипотезы.

Однородность всех текстов и переводов одного автора в рамках математической модели означает справедливость неравенства (3.54), а неоднородность любых двух текстов и переводов разных авторов – справедливость неравенства (3.55). Гипотеза $\mathbb{H}$ может нарушаться для каких-то пар текстов одного и того же автора в случае, когда вместо неравенства (3.54) имеет место неравенство (3.55), а также в случае, когда какие-то два произведения двух различных авторов удовлетворяют неравенство (3.54) вместо того, чтобы выполнялось неравенство (3.55).

Для конкретной коллекции текстов подсчёт суммарного количества $\tau = \tau(\gamma)$ нарушений гипотезы $\mathbb{H}$ позволяет для фиксированного γ оценить результативность гипотезы величиной π, вычисляемой по формуле

$$\pi = 1 - \tau(\gamma)/L\,, \tag{3.56}$$

[4] В исследованиях З.Д. Усманова она названа γ-классификатором.

где L – число взаимных расстояний между всеми парами текстов из коллекции. Из этой формулы следует, что π может принимать значения из отрезка [0, 1], причём $\pi = 0$, если $\tau = L$, и $\pi = 1$, если $\tau = 0$. В первом случае математическая модель оказывается непригодной, а во втором – полностью согласованной с гипотезой $\mathbb{H}$. В § 1.4 предложен алгоритм для нахождения оптимального γ, при котором π *достигает максимального значения.*

Модельная коллекция текстов, предназначенная для проверки плодотворности гипотезы $\mathbb{H}$ и её математической модели, была составлена из 6 художественных произведений:

на таджикском языке

А. Фирдоуси «Рустам ва Сӯҳроб» (АФ, Р&С_tj, 165 Кб),

А. Фирдоуси «Бежан бо Манижа» (АФ, Б&М_tj, 150 Кб),

М. Турсунзода «Ҳасани аробакаш» (МТ, ХА_tj, 93 Кб),

М. Турсунзода «Ҷони ширин» (МТ, ЧШ_tj, 21 Кб),

С. Айни «Дохунда» (СА, Д_tj, 752 Кб),

С. Айни «Марги Судхӯр» (СА, МС_tj, 524 Кб) и

на русском языке

А. Фирдоуси «Сказ о Сохрабе» (АФ, Р&С_ru, 189 Кб),

А. Фирдоуси «Сказ о Бижене и Мениже» (АФ, Б&М_ru, 169 Кб),

М. Турсунзода «Хасани аробекеш» (МТ, ХА_ru, 93 Кб),

М. Турсунзода «Дорогая моя» (МТ, ЧШ_ru, 22 Кб),

С. Айни «Дохунда» (СА, Д_ru, 619 Кб),

С. Айни «Смерть ростовщика» (СА, МС_ru, 465 Кб).

Для авторов и их произведений приняты обозначения, указываемые в скобках: первые две буквы – это инициалы авторов, вторые – шифры текстов, третьи – информация об объёмах произведений в килобайтах.

Обработка статистического материала включала в себя 4 этапа.

Этап 1. Создание компьютерной программы и вычисление с её помощью ЦП произведений, то есть распределений частотности буквенных униграмм по отдельности для всех упомянутых в предыдущем пункте текстов.

Этап 2. Создание компьютерной программы и вычисление с её помощью парных расстояний между ЦПП.

Этап 3. Настройка γ-классификатора. Существо настройки заключалось в том, чтобы определить такое значение вещественного параметра γ, при котором достигается максимальное значение критерия «γ-однородности» произведений. После чего устанавливается оценка π плодотворности гипотезы $\mathbb{H}$ и её математической модели или же распознавания степени однородности оригинальных и переводных произведений.

Этап 4. Обсуждение полученных результатов и заключение о приемлемости гипотезы $\mathbb{H}$ и γ-классификатора.

На этапе 1 для каждого произведения вычислялся ЦП

$$\begin{array}{lllll} \bar{N}: & 1 & 2 & \ldots & 39 \\ P: & p_1 & p_2 & \ldots & p_{39} \end{array}$$

в котором первая строка – номера букв, расположенных в алфавитном порядке, а вторая – относительные частоты встречаемости букв в тексте T, причём $\sum_{k=1}^{39} p_k = 1$.

ЦП представляется также в виде дискретной функции

$$F(s) = \sum_{k=1}^{s} p_k \qquad (s = 1, \cdots, 39).$$

Пусть T_1, T_2 – произвольная пара элементов из модельной коллекции и

$$F^{(\alpha)}(s) = \sum_{k=1}^{s} p_k^{(\alpha)} -$$

соответствующие им дискретные функции, $\alpha = 1, 2$ и $s = 1, \cdots, 39$.

На этапе 2 используется

Определение 3.3.4. *Расстояние* $\rho(T_1, T_2)$ *между текстами* T_1 *и* T_2 *– суть расстояние между их ЦП, определяемое по формуле,* см. § 1.3:

$$\rho(T_1, T_2) = \sqrt{39/2}\ max_s \left| F^{(1)}(s) - F^{(2)}(s) \right|.$$

По этой формуле подсчитаны 66 расстояний, значения которых показаны в таблице 3.30.

Таблица 3.30. – Расстояния между ЦП произведений

Автор (Произведения)		**АФ**				**МТ**				**СА**			
		Р&С (tj)	Б&М (tj)	Р&С (ru)	Б&М (ru)	ХА (tj)	ЧШ (tj)	ХА (ru)	ЧШ (ru)	Д (tj)	МС (tj)	Д (ru)	МС (ru)
АФ	Р&С (tj)												
	Б&М (tj)	**0.029**											
	Р&С (ru)	**0.533**	**0.528**										
	Б&М (ru)	**0.516**	**0.523**	**0.047**									
МТ	ХА (tj)	0.109	0.118	0.524	0.529								
	ЧШ (tj)	0.124	0.139	0.420	0.427	**0.115**							
	ХА (ru)	0.467	0.471	0.066	0.054	**0.478**	**0.376**						
	ЧШ (ru)	0.497	0.504	0.082	0.066	**0.511**	**0.409**	**0.043**					
СА	Д (tj)	0.160	0.152	0.573	0.556	0.065	0.160	0.507	0.536				
	МС (tj)	0.189	0.171	0.574	0.564	0.094	0.165	0.513	0.545	**0.036**			
	Д (ru)	0.478	0.473	0.080	0.058	0.473	0.371	0.050	0.049	**0.519**	**0.520**		
	МС (ru)	0.498	0.492	0.086	0.063	0.489	0.385	0.045	0.044	**0.538**	**0.539**	**0.039**	

На этапе 3 по алгоритму, см. § 1.4, создана программа, с помощью которой на основе данных таблицы 3.30 вычислен полуинтервал оптимальных значений параметра γ:

$$\gamma \in [0.043;\ 0.044), \qquad (3.57)$$

означающий, что ***для любого*** γ ***из этого полуинтервала*** суммарное число τ нарушений неравенств (3.54) и (3.55) становится минимальным, равным для модельной коллекции текстов числу 14 (ячейки нарушений упомянутых неравенств в таблице 3.30 отмечены серым цветом).

Далее по формуле (3.56), в которую нужно подставить $\tau = 14$ и $L = 66$, получаем, что качество классификатора для значений γ из полуинтервала (3.57) оценивается величиной $\pi = 0.79$.

Заключение. Таким образом, для рассматриваемой коллекции текстов математическая модель (γ-классификатор) подтверждает гипотезу $\mathbb{H}$ с точностью 79%. В этой связи интересно выяснить, какого рода ошибки не позволяют достигнуть максимального результата.

Обратимся, прежде всего, к неравенству (3.54). Оно означает однородность двух текстов, расстояние между которыми не больше γ. Согласно гипотезе этому условию должны удовлетворять расстояния между всеми произведениями и переводами одного автора. Авторов у нас – трое, у каждого – по 2 произведения и по 2 перевода его произведений. Следовательно, контролировать нужно 6 расстояний для каждого автора и 18 – для трёх авторов. Как показывает таблица 3.30, для С. Айни (СА) расстояния в 4-х ячейках, закрашенных серым цветом, не удовлетворяют неравенству (3.54); у А. Фирдоуси (АФ) и М. Турсунзода (МТ) таких ячеек по 5. Итого, из 18 расстояний 4 (22.2%) удовлетворяют и 14 (77.8%) не удовлетворяют неравенству (3.54). Это слишком скромный результат для того, чтобы считать неравенство (3.54) подходящей моделью для описания гипотеза $\mathbb{H}$.

Иная картина с выполнением неравенства (3.55). Согласно гипотезе этому условию должны удовлетворять расстояния между всеми произведениями и переводами разных авторов. Таких расстояний 48. Отметим, что для всех (без исключения) расстояний полностью подтвердилась гипотеза $\mathbb{H}$, таким образом, оно выполняется 100%.

И хотя итоговый результат выглядит удовлетворительным, в дальнейшем необходимо будет заняться уточнением гипотезы $\mathbb{H}$ и последующей модернизацией её математической модели.

Результаты данного параграфа опубликованы в [26-А].

§ 3.4. Определение шифр специальности с помощью символьных униграмм

В данном параграфе устанавливается применимость γ-классификатора для автоматического распознавания шифра специальности на основе распределения частотности униграмм. Были взяты научные труды, авторефераты разных ученых,

написанные на русском языке. Авторефераты были взяты в следующих научных областях: история, педагогика, политология, филология и экономика. Сконструированы ЦП и метрическое пространство научных произведений. В предположении уникальности шифр специальности устанавливаются пороговые значения метрики, на основе которых определяются классы «однородных» научных произведений. γ-классификатор дискретных случайных величин, подтвердивший высокую эффективность при идентификации авторства текстовых фрагментов в произведениях классической и современной поэзий, а также в современной прозе таджикского языка, см. главу 2 §§ 2.1-2.5, тестируется на предмет приспособляемости к распознаванию шифра специальности в научных трудах ученых. С момента появления в 2017 г. γ-классификатор [271] широко используется при решении различных задач автоматического распознавания текста, см., например, [255-324].

Для экспериментирования мы ограничились коллекцией из 10 авторефератов, принадлежащих 5 шифрам специальностей, по каждому шифру было взято по 2 автореферата:

шифр 07.00.02: (история):

1. Марков Ю.А. «Массовая бедность в западной Сибири в 1992-2000 гг.».

2. Кляченков Е.А. «Оппозиционная деятельность социалистов и анархистов на территории Орловской и Брянской губерний (октябрь 1917 г. – вторая половина 1920-х гг.)».

шифр 13.00.01: (педагогика):

1. Макарян А.А. «Педагогическое сопровождение развития толерантности в межличностном взаимодействии военнослужащих по призыву».

2. Шуткина Ж.А. «Организационно-педагогические условия формирования конкурентоспособности выпускников негосударственного ВУЗа».

шифр 23.00.01: (политология):

1. Бычков А.А. «Обоснование и кризис имперской идеи в XIV веке: Данте Алигьери, Уильям Оккам и Марсилий Падуанский».

2. Нежданов Д.В. «Метафора «политический рынок» как методологическая основа политических исследований».

шифр 10.01.01: (филология):

1. Розенсон Д.Э. «Творчество Исаака Бабеля в автобиографическом, мемуарном и иудейском контекстах».

2. Шкапа А.С. «Древнерусский памятник «Страсти Христовы»: литературная традиция и жанр».

шифр 08.00.01: (экономика):

1. Ермакова Е.М. «Особенности современного рынка труда в рыночной и переходной экономике».

2. Яськин А.В. «Институциональный фактор экономического выбора на современных рынках».

Совокупность 10 авторефератов будем называть ***А-коллекцией (модельной).***

В работе изучаются ***две задачи. Первая*** состоит в том, чтобы определить способность γ-классификатора (на базе распределения частотности буквенных униграмм) распознавать на обучающей выборке шифры специальности текстов в соответствии с указаниями учителя. ***Вторая задача*** заключается в том, чтобы оценить возможности уже настроенного γ-классификатора безошибочно идентифицировать шифры специальности для новых текстов.

Решение задачи 1. Нам понадобится ряд определений.

3.4.1. Цифровой портрет автореферата (ЦПА). В качестве учётных элементов мы выбрали частоты встречаемости в авторефератах буквенных униграмм.

Определение 3.4.1. *ЦПА назовем в нем частотное распределение 33 букв русского алфавита.*

ЦПА записывается в данном виде:

$$\begin{array}{lccccc} N: & 1 & 2 & . & . & . & 33 \\ P: & p_1 & p_2 & . & . & . & p_{33}, \end{array}$$

где первая строка – это количество букв, расположенных в алфавитном порядке, а вторая – относительная частота встречаемости букв, в автореферате T, причём $\sum_{k=1}^{33} p_k = 1$.

ЦПА также представляется как дискретная функция

$$F(s) = \sum_{k=1}^{s} p_k \qquad (s = 1, \cdots, 33). \tag{3.60}$$

3.4.2. Расстояния между ЦПА. Пусть – произвольная пара авторефератов, характеризуемых на основе алфавита, и

$$F^{(\alpha)}(s) = \sum_{k=1}^{s} p_k^{(\alpha)} - \tag{3.61}$$

соответствующий ЦПА, представленный дискретными функции, $\alpha = 1, 2,$ и $s = 1, \dots, 33.$

Определение 3.5.2. *Расстоянием между авторефератами T_1 и T_2 называется положительное число $\rho\,(T_1, T_2)$, определяемое по формуле*

$$\rho\,(T_1, T_2) = \sqrt{\frac{33}{2}}\, max_s \left| F^{(1)}(s) - F^{(2)}(s) \right|. \tag{3.62}$$

3.4.3. Гипотеза $\mathbb{H}$ «однородности» авторефератов привлекается с целью, чтобы выделения характерной особенности авторефератов, предназначенных для построения математической модели распознавания шифра специальности. Сформулируем это следующим образом.

ГИПОТЕЗА $\mathbb{H}$. *Авторефераты одного шифра специальности «однородные», а разных шифрах – «неоднородные».*

Говоря об «однородности» авторефератов, мы имеем их в виду подобие, одинаковость, сходство, единообразие, родство и т.п.

3.4.4. Математическая модель $\mathbb{H}$-гипотезы. Пусть γ – некоторое положительное число.

Определение 3.4.3. *Авторефераты T_1, T_2 называются γ-однородными (принадлежащими одному шифру), если*

$$\rho\,(T_1\,, T_2) \leq\ \gamma, \tag{3.63}$$

и γ- неоднородными (принадлежащими различным шифрам), если

$$\rho\,(T_1\,, T_2) > \gamma. \tag{3.64}$$

Неравенства (3.63) и (3.64) являются математической интерпретацией (моделью) гипотезы $\mathbb{H}$.

Определение 3.4.4. *γ-классификатор – в зависимости от одного вещественного параметра γ алгоритм принятия решения об присвоении пары авторефератов T_1 и T_2 одному или двум разным шифрам.*

Очевидно, неоднородность или однородность любой пары авторефератов зависит от значения γ и, следовательно, уровня выполнимости гипотезы. Принадлежность двух текстов к одному шифру в рамках математической модели означает справедливость неравенства (3.63), а два разных шифра – справедливость неравенства (3.64). Гипотеза $\mathbb{H}$ может быть нарушена для некоторых пар текстов одного и того же шифра в том случае, когда вместо неравенства (3.63) имеет место неравенство (3.64), а также в случае, когда в каких-то двух текстах на разных шифрах выполнялось неравенство (3.64).

Пусть $\tau = \tau(\gamma)$ – общее количество нарушений гипотезы $\mathbb{H}$ одновременно в двух случаях: невыполнения неравенства «однородность» в случае двух текстов, принадлежащих одну шифру, и невыполнения неравенства «неоднородность» в случае двух текстов, принадлежащих разным шифрам. Тогда при фиксированном γ показатель выполнения гипотезы будет определяться значением π, задаваемым формулой

$$\pi = 1 - \ \tau(\gamma)/L\,,$$

где L – число взаимных расстояний между всеми парами авторефератов из (в нашем случае $L = C_{10}^2 = 45$). Из этой формулы следует, что π может принимать значения из отрезка [0, 1], причём $\pi = 0$, если $\tau = L\ (= 45)$, и $\pi = 1$, если $\tau = 0$. В первом случае гипотезу $\mathbb{H}$ следует признать непригодной, а во втором – полностью соответствующей обучающей выборке.

Поскольку эффективность γ-классификатора зависит от параметра γ, представляет интерес найти такое значение, которое π принимает максимальным. В этом суть настройки γ-классификатора на данных исследуемой выборки.

3.4.5. Окончательные результаты для примера коллекции моделей *A* показаны ниже, предварительно выполнив следующие операции:

– вычисления ЦП (частота букв русского алфавиту) для 10 авторефератов коллекции моделей А;

– расчет по формулам (3.60), (3.61) и (3.62) 45 парных расстояний $\rho\ (T_1, T_2)$ между авторефератами коллекции ***A*** (результаты расчетов приведены в следующей таблице 3.31):

Таблица 3.31. – Расстояния между авторефератами коллекции ***A***

Шифры (Авторефераты)		**07.00.02**		**13.00.01**		**23.00.01**		**10.01.01**		**08.00.01**	
		1	2	1	2	1	2	1	2	1	2
07.00.02	1										
	2	**0.0891**									
13.00.01	1	0.0817	0.0646								
	2	0.1059	0.0792	**0.0615**							
23.00.01	1	0.1071	0.0821	0.0627	0.0998						
	2	0.0827	0.0443	0.0609	0.0644	**0.0393**					
10.01.01	1	0.1277	0.1737	0.1829	0.2336	0.1601	0.1693				
	2	0.1172	0.0757	0.1268	0.0925	0.0928	0.0741	**0.1749**			
08.00.01	1	0.1182	0.0961	0.1244	0.0901	0.1003	0.0716	0.2341	0.0592		
	2	0.1028	0.0715	0.0828	0.0771	0.0676	0.0471	0.2032	0.0737	**0.0591**	

– расчет с использованием алгоритма настройки γ -классификатора из § 1.4, задающего оптимальный интервал значений γ, при котором значение $\tau = \tau(\gamma)$ от общего числа случаев нарушения гипотезы $\mathbb{H}$ достигает минимального значения и, следовательно, значение π показателя выполнения гипотезы $\mathbb{H}$ принимает максимальное значение.

На основании данных таблицы 3.31 были получены следующие результаты:

– набор всех пар расстояний находится на отрезке [0.0393, 0.2341], при этом минимальное расстояние реализуется между шифрами 23.00.01 «Автореферат-1» и 23.00.01 «Автореферат-2», а максимальное – между шифрами 10.01.01 «Автореферат-1» и 08.00.01 «Автореферат-1»;

– половина оптимального интервала значений γ находится в пределах

$$\gamma^{\text{опт}} \in [0.0610;\ 0.0614);$$

Применять этот факт для выяснения метрической близости пары авторефератов T_1 $и$ T_2 необходимо следующим образом:

– если $\rho(T_1, T_2) < [0.0610;\ 0.0614)$, то T_1 $и$ T_2 однородный;

– если $\rho(T_1, T_2) > [0.0610;\ 0.0614)$, то T_1 $и$ T_2 неоднородный.

В табл. 3.31 закрашенные серым цветом ячейки (в данном случае их 6) показывают нарушение сформулированной гипотезы для соответствующих пар авторефератов, и потому получено

$$\tau = \tau_{min} = 6,$$

– в результате показатель эффективности предложенной в данной работе математической модели распознавания шифра авторефератов оказался равным

$$\pi = \pi_{max} = 0.87$$

Задача 2 «Тестирование». Итак, результаты предыдущего раздела показывают, что настройка (обучение) γ-классификатора в данной коллекции моделей *A* прошла успешно.

Для теста классификатора были выбраны следующие авторефераты (они все записаны под номером 3, чтобы показать, что они являются третьими авторефератами из соответствующих шифров специальности):

шифр 07.00.02:

3. Аракелян М.А. «Политическая полиция Российской империи в борьбе с революционным подпольем в 1881-1905 гг.»;

шифр 13.00.01:

3. Дуда И.В. «Формирование ценностных ориентаций больных сколиозом школьников в учебно-воспитательном процессе школы-интерната»;

шифр 23.00.01:

3. Андреев М.Г. «Роль средств массовой информации в формировании позитивного образа некоммерческих организаций в современной России»;

шифр 10.01.01:

3. Левина Е.Н. «Проблема биографизма в творчестве И.С. Тургенева 1840-1850-х годов»;

шифр 08.00.01:

3. Добролежа Е.В. «Управление ресурсным обеспечением экономики региона».

После формирования ЦП авторефератов, предназначенных для тестирования и расчета расстояний по формуле (3.62), была получена следующая таблица расстояний от каждого из протестированных авторефератов до всех 10 авторефератов исходной коллекции.

Таблица 3.32. – Расстояния между авторефератами коллекции и тестируемыми авторефератами

Шифры (Авторефераты)		**07.00.02**	**13.00.01**	**23.00.01**	**10.01.01**	**08.00.01**
		3	3	3	3	3
07.00.02	1	**0.0592**	0.1194	0.0472	0.1033	0.1071
	2	**0.0605**	0.0755	0.0795	0.1923	0.0851
13.00.01	1	0.0752	**0.1072**	0.0632	0.1513	0.0773
	2	0.0864	**0.0923**	0.0891	0.1532	0.0775
23.00.01	1	0.0937	0.0824	**0.0875**	0.1747	0.0713
	2	0.0681	0.0815	**0.0355**	0.1852	0.0599
10.01.01	1	0.1569	0.2116	0.1569	**0.0902**	0.2168
	2	0.0803	0.1264	0.0892	**0.1405**	0.0757
08.00.01	1	0.0931	0.1009	0.0896	0.1537	**0.0796**
	2	0.0778	0.0603	0.0908	0.2036	**0.0574**

В таблице 3.32 серым цветом показана пара ячеек, соответствующих минимальным расстояниям от тестируемых авторефератов до авторефератов коллекции $\boldsymbol{A}$.

Так для автореферата 07.00.02(3) ближайшим соседом оказался автореферат 07.00.02 (1);

для автореферата 13.00.01(3) ближайшим соседом оказался автореферат 08.00.01(2);

для автореферата 23.00.01(3) ближайшим соседом оказался автореферат 23.00.01(2);

для автореферата 10.01.01(3) ближайшим соседом оказался автореферат 10.01.01(1);

для автореферата 08.00.01(3) ближайшим соседом оказался автореферат 08.00.01(2).

По данным таблицы 3.32 метод ближайшего соседа безошибочно определяет шифры 4 тестируемых авторефератов из 5 и для 1 автореферата допускает ошибку.

Заключение. γ-классификатор с фиксированным значением $\gamma = \gamma^{\text{опт}}$ был протестирован на случайных выборках авторефератов и подтвердил 87%-ую способность к распознаванию шифра специальности авторефератов.

Результаты данного параграфа опубликованы в [32-А].

§ 3.5. К вопросу об автоматическом определении стилей и авторства произведений таджикско-персидской художественной литературы

На основе применения γ-классификатора к обработке 68 произведений 7 литературных школ устанавливаются оценки эффективности распознавания авторства и стилей в рамках таджикско-персидской литературы.

Объектом исследования настоящего параграфа является ряд шедевров персидской классической поэзии хорасанской, иракской и индийской литературных школ, дополненный произведениями школ классической прозы,

смешанного стиля, современной поэзии и современной прозы, которые также, как и предыдущие, следуют во времени друг за другом.

3.5.1. Состав коллекции текстов, обрабатываемый далее, приводится с указанием автора (в скобках – его аббревиатура и годы жизни) и названия произведений (в скобках – их сокращенные обозначения и размеры в количестве слов):

Хорасанская школа

– А. Рӯдакӣ (АР, 859-941): «Абёти пароканда» (АП, 2248 *слов*), «Қасоид» (Қ, 5054 *слова*);

– А. Дақиқӣ (АД, 930-977): «Осор» (О, 1309 *слов*);

– А. Фирдавсӣ (АФ, 932-1020): «Заҳҳок» (З, 5841 *слово*), «Достони Рустам ва Сӯҳроб» (Р&С, 16355 *слов*), «Достони Сиёвуш» (С, 30503 *слова*), «Достони Бежан бо Манижа» (Б&М, 14799 *слов*);

– А. Унсурӣ (АУ, 961-1039): «Қасидаҳо» (Қ, 3452 *слова*);

– Н. Хисрав (НХ, 1004-1088): «Маснавиҳо» (М, 1766 *слов*), «Шеърҳо» (Ш, 2190 *слов*);

– А. Тӯсӣ (АТ, 1010-1073): «Гаршосп» (Г, 1568 *слов*);

– С. Тирмизӣ (СБ, 1078-1147): «Осор» (О, 2183 *слова*).

Иракская школа

– У. Хайём (УХ, 1040-1123): «100-Рубоӣ» (100Р, 2541 *слово*), «301-Рубоӣ» (301Р, 5899 *слов*);

– С. Шерозӣ (СШ, 1184-1292): «Ғазалиёт қисми 1» (Ғ1, 16261 *слово*), «Ғазалиёт қисми 2» (Ғ2, 13001 *слово*);

– Ҷ. Румӣ (ҶР, 1207-1273): «Маснавии Маънавӣ Дафтари Аввал» (ММ1, 48713 *слова*), «Маснавии Маънавӣ Дафтари Дуввум» (ММ2, 41661 *слово*), «Маснавии Маънавӣ Дафтари Севум» (ММ3, 57787 *слов*), «Маснавии Маънавӣ Дафтари Чорум» (ММ4, 47285 *слов*);

– С. Совақӣ (СС, 1300-1376): «Ҷамшед ва Хуршед» (Ҷ&Х, 2133 *слова*);

– К. Хуҷандӣ (КХ, 1318-1401): «Ғазалиёт қисми 1» (Ғ1, 55179 *слов*), «Ғазалиёт қисми 2» (Ғ2, 51011 *слово*);

– Ҳ. Шерозӣ (ҲШ, 1325-1390): «Ғазалиёт қисми 1» (Ғ1, 33724 *слова*), «Ғазалиёт қисми 2» (Ғ2, 28923 *слова*);

– А. Ҷомӣ (АҶ, 1414-1492): «Лайлӣ ва Маҷнун» (Л&М, 33874 *слова*), «Юсуф ва Зулайхо» (Ю&З, 44483 *слова*).

Индийская школа

– С. Насафӣ (СН, 1634-1711): «Шаҳрошӯб Чарчинфурӯш» (ШЧ, 935 *слов*), «Баҳориёт Ҳайвонотнома Муш» (БҲМ, 2057 *слов*);

– Б. Зебуннисо (БЗ, 1639-1702): «Девони махфӣ қисми 1» (ДМ1, 21612 *слова*), «Девони махфӣ қисми 2» (ДМ2, 17016 *слов*);

– А. Бедил (АБ, 1644-1721): «Байтҳо» (Б, 1978 *слов*), «Девони Қитъаҳо» (Қ, 1305 *слов*);

– М. Бухоро̄ (МБ, 1720-1800): «Ғазалиёт» (Ғ, 984 *слова*).

Школа классической прозы

– А. Берун̄ (АБ, 973-1046): «Осор-ул-боқия 1» (ОБ1, 5758 *слов*), «Осор-ул-боқия 2» (ОБ2, 4749 *слов*);

– У. Кайковус (УК, 1001-1100): «Қобуснома Бобҳои 1-22» (Қ1-22, 23243 *слова*), «Қобуснома Бобҳои 23-44» (Қ23-44, 35244 *слова*);

– М. Ғазол̄ (МҒ, 1059-1111): «Насиҳат ул мулук» (НМ, 919 *слов*).

Школа смешанного стиля

– И. Балх̄ (ИБ, 1688-1749): «Ғазалиёт» (Ғ, 847 *слов*);

– А. Савдо (АС, 1823-1873): «Байтҳо» (Б, 2477 *слов*);

– А. Дониш (АД, 1826-1897): «Қасидаҳо 1-45» (Қ1-45, 2731 *слово*), «Қасидаҳо 46-93» (Қ46-93, 2323 *слова*);

– Т. Асирр̄ (ТА, 1864-1916): «Байтҳо» (Б, 442 *слова*);

– Н. Туғрал (НТ, 1865-1919): «Шеърҳо» (Ш, 604 *слова*);

– С. Айн̄ (СА, 1875-1954): «Байтҳо» (Б, 1247 *слов*).

Школа современной поэзии

– М. Лоҳурр̄ (МЛ, 1877-1938): «Паёми Форуқ» (ПФ, 1015 *слов*);

– М. Турсунзода (МТ, 1911-1977): «Ҳасани аробакаш» (ҲА, 8463 *слова*), «Садои Осиё» (СО, 879 *слов*);

– М. Миршакар (ММ, 1912-1993): «Достони Ишқи духтари кӯҳсор» (ИДК, 3720 *слов*), «Қишлоқи тиллоӣ» (ҚТ, 2227 *слов*);

– М. Қаноат (МҚ, 1932-2018): «Достони Оташ» (ДО, 3584 *слова*), «Ҳамосаи дод» (ҲД, 3875 *слов*);

– Л. Шералӣ (ЛШ, 1941-2000): «Катибаҳо» (К, 3290 *слов*), «Суханреза» (С, 3872 *слова*);

– А. Суруш (АС, р.1954): «Дафтари 1» (Д1, 7890 *слов*), «Дафтари 2» (Д2, 9322 *слова*);

– И. Фарзона (ИФ, р.1964): «101-Ғазал» (101Ғ, 9841 *слово*), «Мӯҳри гули мино» (МГМ, 41217 *слов*).

Школа современной прозы

– С. Айн̄ (СА, 1875-1954): «Аҳмади Девбанд» (АД, 7480 *слов*), «Одина» (О, 4000 *слов*), «Ёддоштҳо 1» (Ё1, 56312 *слова*);

– С. Улуғзода (СУ, 1911-1997): «Бежан ва Манижа» (Б&М, 19868 *слов*), «Рустам ва Сӯҳроб» (Р&С, 7328 *слов*);

– М. Шакурр̄ (МШ, 1925-2012): «Садри Бухоро» (СБ, 113592 *слова*), «Хуросон аст ин ҷо» (Х, 91202 *слова*);

– С. Турсун (СТ, р.1946): «Нисфирӯзӣ» (Н, 9936 *слов*), «Повести Камони Рустам» (ПКР, 4041 *слово*).

Таким образом, рассматриваемая нами коллекция текстов содержит 68 произведений 38 авторов.

3.5.2. Закономерности коллекционного материала, выявленные с помощью γ- классификатора, описанного в § 1.4. В применении к исследуемой коллекции определение значения γ производилось на основе ЦП всех 68 произведений, *характеризовавших частоты встречаемости буквенных триграмм с учетом пробелов*, и вполне приемлемой гипотезы о том, что *произведения одного автора "однородны", а разных авторов "неоднородны"*. В результате вычислений получено

$$\gamma \in [1.9056;\ 1.9264). \qquad (3.58)$$

Согласно §§ 1.3 и 1.4 это значит, что если расстояние $\rho\,(v_1\,, v_2)$ между текстами v_1 и v_2 меньше 1.9056, то v_1 и v_2 однородны, если же больше 1.9264, то v_1 и v_2 неоднородны и, наконец, при $\rho\,(v_1\,, v_2)$, принадлежащем указанному полуинтервалу, относительно v_1 и v_2 нельзя утверждать ни то, ни другое.

Применяя сказанное к анализу матрицы расстояний $\{\rho\,(v_1\,, v_2)\}$ между всеми 68 текстами (соответствующие данные не приводятся из-за большего размера матрицы), устанавливаем, что за исключением двух произведений С. Шерозӣ, которые объявляются неоднородными, все другие творения каждого из 37 авторов однородны между собой. С другой стороны, среди пар произведений 38 различных авторов лишь 31 пара оказалась однородной. Таким образом, в рамках γ-классификатора высказанная гипотеза об однородности текстов получила статистическое подтверждение в 99% случаях (общее число пар – 2278, нарушение обнаружено для 32 пар).

3.5.3. Закономерности стилей литературных школ выявлены также с помощью γ-классификатора. В данном пункте в отличие от предыдущего вычисление γ основывалось на иной гипотезе: *стили произведений авторов одной школы «однородны», а разных школ «неоднородны»*. Что касается ЦП произведений, то они оставлены без изменений.

В результате вычислений получено

$$\gamma \in [2.8324;\ \ 2.8338). \qquad (3.59)$$

Как и в предыдущем пункте это значит, что если расстояние $\rho\,(v_1\,, v_2)$ между текстами v_1 и v_2 меньше 2.8324, то v_1 и v_2 однородны по стилю, если же больше 2.8338, то v_1 и v_2 неоднородны. И, наконец, при $\rho\,(v_1\,, v_2)$, принадлежащем указанному полуинтервалу, относительно v_1 и v_2 нельзя утверждать ни то, ни другое.

И вновь, применяя сказанное к изучению матрицы расстояний $\{\rho\,(v_1\,,v_2)\}$ между 68 текстами (и здесь соответствующие данные не приводятся из-за большего размера матрицы), устанавливаем следующие результаты.

Внутри Хоросанской школы среди 66 различных пар произведений однородными по стилю оказалась 21 и неоднородными – 45 пар.

Внутри Иракской школы соответствующие данные таковы: различных пар произведений – 105, однородных по стилю – 35, неоднородных – 70 пар.

Внутри Индийской школы: различных пар произведений – 21, однородных по стилю – 6, неоднородных – 15 пар.

Внутри Школы классической прозы: различных пар произведений – 10, однородных по стилю – 4, неоднородных – 6 пар.

Внутри Школы смешанного стиля: различных пар произведений – 21, однородных по стилю – 4, неоднородных – 17 пар.

Внутри Школы современной поэзии: различных пар произведений – 78, однородных по стилю – 27, неоднородных – 51 пара.

Внутри Школы современной прозы: различных пар произведений – 36, однородных по стилю – 8, неоднородных – 28 пар.

Таким образом, внутри 7 школ из общего числа 337 пар произведений 232 пары оказались неоднородными по стилю, а 105 пар – однородными. Что касается пар произведений из разных школ, то среди 1941 пары подтвердили гипотезу 1719 (они оказались неоднородными) и 222 вошли в противоречие с ней (оказались однородными).

Итого, из 2278 различных пар, составленных из 68 произведений, выявлено 80 % (1824 = 1719 + 105), согласных с гипотезой и 454 пар, противоречащих ей.

ЗАМЕЧАНИЕ. Обратим внимание на различие значений γ в выражениях (3.58) и (3.59). Оно явилось следствием применения различных гипотез, использованных для вычисления γ: в первом случае – нацеленное на идентификацию авторов произведений, во втором – на отождествление авторских стилей.

3.5.4. Модельный вариант основоположников таджикско-персидских литературных школ. Изложенное в п. 3.5.3 на основе γ-классификатора исследование показало, что внутри каждой из научных школ неоднородных по стилю произведений оказалось больше, чем однородных. И хотя в целом полученные результаты выглядят вполне приемлемыми (80 % пар произведений подтвердили гипотезу), всё же предпочтительней была бы ситуация, в которой получилось наоборот: внутри школ число однородных по стилю пар было бы больше неоднородных. Тем не менее сам факт следует принять как должное и объяснение итоговой картины искать в решении экспертов увязывать стили произведений с периодами жизней их создателей.

В **заключение** приведем следующую таблицу, в которой показано всего лишь 4 стиля из 7.

Таблица 3.33. – Расстояния между ЦП произведений четырех стилей

Авторы	Классический хорасанский стиль						Классический иракский стиль				Классическая проза			Современная проза		
	АП	**Қ**	**З**	**Р&С**	**С**	**Б&М**	**100Р**	**301Р**	**Ғ1**	**Ғ2**	**Қ1-22**	**Қ23-44**	**НМ**	**АД**	**О**	**Ё1**
АП																
Қ	**1.79**															
З	**2.43**	**2.73**														
Р&С	**2.41**	**2.29**	**1.37**													
С	**2.61**	**2.79**	**1.91**	**1.58**												
Б&М	**2.72**	**2.58**	**1.39**	**0.77**	**1.77**											
100Р	3.11	2.84	5.44	4.24	5.51	4.71										
301Р	3.71	3.48	6.06	4.84	6.11	5.32	**1.11**									
Ғ1	3.80	3.58	6.06	4.82	6.10	5.30	**2.05**	**1.80**								
Ғ2	4.51	4.33	6.81	5.59	6.87	6.07	**2.37**	**2.00**	**0.99**							
Қ1-22	3.43	4.22	5.27	4.10	5.31	4.50	3.87	4.20	4.61	4.96						
Қ23-44	5.19	5.58	7.06	5.87	7.10	6.29	5.05	5.46	5.94	6.14	**1.93**					
НМ	5.96	6.15	7.08	6.55	7.49	6.67	5.40	6.00	6.42	6.10	**2.83**	**2.48**				
АД	7.18	6.91	7.32	7.38	8.04	6.83	6.67	6.84	7.18	7.73	5.01	4.43	6.17			
О	6.11	5.85	7.51	6.34	7.87	6.78	6.07	5.77	5.90	6.47	4.30	3.74	5.58	**1.57**		
Ё1	6.61	6.38	7.43	6.87	7.54	6.68	6.13	5.75	5.91	6.39	4.49	3.92	5.68	**1.78**	**1.65**	

Эта таблица извлечена из упомянутой ранее, размером 68 х 68, из которой удалены все произведения «небольших» объёмов, а также те, которые нарушали в особо «больших» масштабах гипотезу об однородности стилей. В таблице представлены 2 произведения А. Рудаки и 4 произведения А. Фирдоуси (Хорасанская школа), по 2 от У. Хайёма и Х. Шерози (Иракская школа), 2 от У. Кайковуса и 1 от М. Газоли (Школа классической прозы) и, наконец, 3 от С. Айни (Школа современной прозы). Для такого сочетания авторов со своими произведениями получено значение $\gamma \in$ [2.8324; 2.8385), незначительно отличающееся от (3.59), но для которого гипотеза об однородности стилей выполняется на все 100%. Именно по этой причине итоговая таблица интерпретируется в качестве модельного варианта основоположников таджикско-персидских литературных школ.

Результаты данного параграфа опубликованы в [16-А].

§ 3.6. Выводы по главе 3

Кроме распознавания авторства на основе γ-классификатора З.Д. Усманова и МБС, в этой главе также рассматривалось распознавание других признаков

однородности, таких как тематики текста, язык, группа языков, оригинал и его перевод, стиль произведений и шифры научных работ на основе различных ЦПТ. Обученный на модельной коллекции γ-классификатор показал высокую точность в распознавании однородности произведений для различных признаков. В свою очередь, методом ближайшего (по расстоянию) соседа все новые тексты подтвердили свою однородность с соответствующими парами произведений различных признаков. В следующей главе исследуется подобная задача не в модельной коллекции текстов, а в корпусах произведений художественной литературы, для которой удастся ли получить удовлетворительный результат решения рассматриваемой задачи.

ГЛАВА 4. ИССЛЕДОВАНИЕ СТАТИСТИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ РАСПОЗНАВАНИЯ ОДНОРОДНЫХ ТЕКСТОВ В КОРПУСАХ ХУДОЖЕСТВЕННЫХ ЛИТЕРАТУРНЫХ ПРОИЗВЕДЕНИЙ

В главе 2 на примере модельной коллекции текстов путём подбора оптимального значения γ удалось обучить γ-классификатор относительно успешного распознавания однородности произведений. По ходу выполнения работы обнаружилось непредвиденное: оптимальное γ – это вовсе не единственное число, а полуинтервал числовых значений. Предварительные, однако, бессистемные исследования показали, что увеличение количества произведений влияет на размеры упомянутого полуинтервала, причём в сторону уменьшения его длины и что, возможно, в пределе он стремится к единственному значению. Поэтому, чтобы выявить такое явление в настоящей главе исследуется распознавание однородности текстов не в модельной коллекции текстов, а в корпусах произведений художественной литературы, для которого удастся ли получить удовлетворительный результат решения рассматриваемой задачи. На примере корпуса рассмотрены случаи с 5, 10, 20 предполагаемыми авторами или языками, а также с 10, 20, 40 текстами, выявляются особенности применения γ-классификатора при распознавании автора или языка текста. Для тестирования классификатора дополнительно было выбрано несколько случайных текстов, которые составлены теми же авторами или языками, что и тексты коллекции. Методом ближайшего (по расстоянию) соседа тестируемые тексты проверяется на однородность с соответствующими парами произведений языков или авторов. Наша цель будет состоять не только в том, чтобы выявить различия в размерах и расположениях оптимальных полуинтервалов γ, но также и в определении числа нарушений гипотезы однородности, вычислении коэффициента эффективности распознавания авторов или языков по их произведениям в целом и, возможно, минимальным фрагментам. Приводятся результаты экспериментов по применению γ-классификатора на корпусе текстов художественной литературы.

§ 4.1. Исследование статистических закономерностей определения языка произведений на основе кириллического алфавита в корпусах текстов художественной литературы

В данном параграфе устанавливается применимость γ-классификатора для автоматического распознавания языка произведения на основе частотности общих 26 кириллических алфавитных букв на примере корпуса, состоящего из 70 текстов на 20 языках (по 8 произведений на 5 языках: белорусском, болгарском, русском, таджикском и украинском, и по 2 произведения на других 15 языках) с использованием кириллической графики. Математическая модель γ-

классификатора представляется в виде триады. Её первым компонентом является ЦПТ – распределение в тексте частотности буквенных униграмм; в качестве второго компонента служит формула для вычисления расстояний между ЦП текстов и третий компонент – алгоритм машинного обучения, реализующий гипотезу «однородности» произведений, написанных на одном языке, и «неоднородности» произведений, написанных на разных языках. Настройка алгоритма, использующего таблицу парных расстояний между всеми произведениями коллекции текстов, заключалась в определении оптимального значения вещественного параметра γ, для которого минимизируется ошибка нарушения гипотезы «однородности». На примере корпуса рассмотрены случаи с 5, 10, 20 предполагаемыми языками, а также с 10, 20, 40 текстами, выявляются особенности применения γ-классификатора при распознавании языка текста. Для тестирования классификатора дополнительно было выбрано три случайных текста, которые составлены теми же языками, что и тексты коллекции. Методом ближайшего (по расстоянию) соседа три случайных текста проверяются на однородность с соответствующими парами одноязычных произведений.

В наше время письменность на основе кириллического алфавита получила широкое распространение среди монгольской, славянской, тюркской, уральской и иранской групп языков таких стран, как Россия, Таджикистан, Молдавия, Азербайджан, Узбекистан, Туркменистан, Казахстан, Киргизия, Северная Македония, Приднестровская Молдавская Республика, Сербия, Украина, Черногория и Южная Осетия, [272]. За исключением современного русского, для большинства других языков кириллический алфавит из 33 букв (а, б, в, г, д, е, ё, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ц, ч, ш, щ, ы, ь, ъ, э, ю, я) оказался недостаточным для обозначения всех звуков этих языков, в связи с чем для отражения фонетических особенностей тех или иных языковых систем к базовой кириллической графике были добавлены различные диакритические знаки, лигатуры и другие модификации букв.

Задача, решением которой будем заниматься в этом параграфе, состоит в том, чтобы определить, можно ли обойтись только лишь 26 (а, б, в, г, д, е, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ч, ш, ю, я) кириллическими буквами для автоматического распознавания языка, на котором написана та или иная печатная продукция.

Приступая к решению поставленной задачи, отметим, что в качестве исследовательского инструмента мы будем использовать *математическую триаду* в составе ЦП текстов, представляемых распределениями частотности 26 кириллических букв, формулы для вычисления расстояний между текстами и алгоритма для выявления однородных текстов, см. §§ 1.3 и 1.4. Упомянутая триада с момента своего появления в 2017 году применялась, прежде всего, для

распознавания авторства для различных вариантов ЦП текстов [255-324]. В дополнении к сказанному уместно отметить, что в монографии [227] представлен обширный обзор работ по идентификации авторов текстов на основе разнообразных ЦП текстов и применяемых методов классификации. Отметим, что ранее аналогичный вопрос изучался именно для символьных (буквенных) униграмм в §§ 3.2.2 и 3.2.4. Существенным моментом в сравнении с нашим предыдущим исследованием в § 3.2 является изучение вопроса не в модельной коллекции текстов, а в корпусах произведений художественной литературы, для которого удается получить удовлетворительный результат решения рассматриваемой задачи.

Состояние работ по применению различных классификаторов, прежде всего методов нейронных сетей и машин опорных векторов, подробно описано в монографии [227]. В настоящем параграфе на примере корпуса, состоящего из 70 произведений на 20 разных языках, решаются две задачи:

– *настроить так называемый γ-классификатор, по возможности, для безошибочного распознавания принадлежности текстов соответствующих языков путем подбора вещественного параметра γ;*

– *проверить правильность работы настроенного классификатора для трёх дополнительных случайно выбранных произведений, принадлежащих различным языкам.*

Прежде чем переходить к изучению задач, напомним основные понятия, связанные с компонентами триады.

4.1.1. Корпус текстов $\boldsymbol{C}$ для исследований. В приводимом далее списке элементов коллекции $\boldsymbol{C}$ указываются имя автора, название его сочинения на родном языке и в скобках – аббревиатура сочинения и его размеры в количестве слов:

на башкирском языке (Ba): Ф.А. Сайфуллин «Юғалтыу һағыштары, часть 1» (ba_1, 11543 *слова*); Ф.А. Сайфуллин «Юғалтыу һағыштары, часть 2» (ba_2, 12112 *слова*);

на белорусском языке (Be): Л. Станислав «Салярыс, часть 1» (be_1, 8497 *слов*); Л. Станислав «Салярыс, часть 2» (be_2, 7041 *слово*); С. Давидович «Дзед-кіёк» (be_3, 1935 *слов*); У. Миллер «Гадзіна памяці» (be_4, 5051 *слово*); А. Моруа «Гатэль Танатос Палац» (be_5, 4035 *слов*); Г. Угаров «Вярнуць адкрыццё» (be_6, 3602 *слова*); Н. Ткачев «Жарынка» (be_7, 3331 *слово*); В. Мудров «Калодзеж» (be_8, 10104 *слова*);

на болгарском языке (Bo): Н. Райнов «Неволя и богатство» (bo_1, 2565 *слов*); Н. Райнов «Търговец и дяволи» (bo_2, 2567 *слов*); Н. Райнов «Майчина грижа» (bo_3, 2644 *слова*); Б. Джим «Фурията на принцепса, глава 1» (bo_4, 2491 *слово*); А. Каралийчев «Гулчечек» (bo_5, 2436 *слов*); А. Катцу «Глад, глава 1-3» (bo_6,

7890 *слов*); А. Катцу «Глад, глава 4-6» (bo_7, 9893 *слова*); Д.М. Юрий «И сам воинът е воин» (bo_8, 12151 *слово*);

на ваханском языке (Vs): «Афсонаҳо» (vs_1, 6701 *слово*); «Шеърҳо» (vs_2, 9809 *слов*);

на казахском языке (Kz): А. Қунанбаев «Абайдың қара сөздері» (kz_1, 16751 *слово*); А. Қунанбаев «Автобиаграфия» (kz_2, 1709 *слов*);

на языке коми (Ko): А. Вурдов «Лумпа туй (Рассказ)» (ko_1, 2108 *слов*); А. Вурдов «Эзысь сир войт» (ko_2, 1946 *слов*);

на кыргызском языке (Ky): Ч. Айтматов «Айтматовдун акыркы адашуусу жана айкөлдүгү (Последнее заблуждение и щедрость Айтматова)» (ky_1, 1510 *слов*); Ч. Айтматов «Өмүр баяны, биография» (ky_2, 1012 *слова*);

на молдавском языке (Ml): «Лимба матернэ — флоаре етернэ, часть 1» (ml_1, 5689 *слов*); «Лимба матернэ — флоаре етернэ, часть 2» (ml_2, 4557 *слов*);

на монгольском языке (Mn): «Надаар тоглосон хайр (Жүжгийн зохиол)» (mn_1, 2609 *слов*); «Театр» (mn_2, 2949 *слов*);

на русском языке (Ru): М.А. Шолохов «Судьба человека» (ru_1, 10891 *слово*); Ф.А. Абрамов «Алька» (ru_2, 15668 *слов*); А.П. Гайдар «Голубая чашка» (ru_3, 6740 *слов*); А.П. Гайдар «Судьба барабанщика» (ru_4, 7191 *слово*); М.А. Горький «Коновалов» (ru_5, 15620 *слов*); Ф.Д. Крюков «В родных местах» (ru_6, 9541 *слово*); Ф.Д. Крюков «Казачка» (ru_7, 12818 *слов*); И.С. Тургенев «Муму» (ru_8, 8425 *слов*);

на сербском языке (Se): А. Кларк «Напеви далеке Земље» (se_1, 11129 *слов*); Р.Л. Стивенсон «Црна стрела» (se_2, 15028 *слов*);

на таджикском языке (Tj): С. Айни «Аҳмади Девбанд» (tj_1, 7485 *слов*); С. Турсун «Повести Камони Рустам» (tj_2, 4041 *слово*); С. Улуғзода «Бежан ва Манижа» (tj_3, 19730 *слов*); А. Берунӣ «Осор-ул-боқия» (tj_4, 10509 *слов*); С. Турсун «Нисфирӯзӣ» (tj_5, 9958 *слов*); Ӯ. Кӯҳзод «Тахти равон ва тахтбардорон» (tj_6, 4774 *слова*); Ч. Айтматов «Чингиз ва Бибисоро» (tj_7, 8711 *слово*); Қ. Рустам «Азобҳои ҷаҳаннам» (tj_8, 9549 *слов*);

на татарском языке (Ta): Г. Тукай «Тугандаш шагыйрьнең элгәре турында» (ta_1, 2555 *слов*); «Кәҗә белән Сарык» (ta_2, 3612 *слова*);

на удмуртском языке (Ud): «Шоро-куспо югдур» (ud_1, 2361 *слово*); «Кионлэн пытьыез кузя» (ud_2, 2548 *слов*);

на узбекском языке (Uz): А. Ирисов «А. Сино. Ҳайй ибн Яқзон (фалсафий қисса)» (uz_1, 3049 *слов*); З.М. Бобур «Маҳрами асрор топмадим» (uz_2, 9493 *слова*);

на украинском языке (Uk): В.Л. Кашин «Готується вбивство» (uk_1, 23771 *слово*); В.Л. Кашин «День народження» (uk_2, 13636 *слов*); М. Циба «Акванавти, або Золота жила» (uk_3, 20150 *слов*); О. Бердник «Відслонити завісу часу!» (uk_4,

3598 *слов*); В.П. Бережной «Homo Novus» (uk_5, 5768 *слов*); В.П. Бережной «В космічній безвісті» (uk_6, 3921 *слово*); В.П. Бережной «Діти одного Сонця» (uk_7, 3317 *слов*); В.П. Бережной «Дарунки Шамбали» (uk_8, 4663 *слова*);

на чеченском языке (Ce): В. Хаджимурадов «Лакхарчу юьртара Іандаркъа» (ce_1, 4319 *слов*); В. Хаджимурадов «Зийнин «Майра к1ант» (Храбрый мальчик Зийны)» (ce_2, 5335 *слов*);

на чувашском языке (Cu): Л.Я. Агаков «Пăрахут анаталла каять» (cu_1, 2297 *слов*); «Аса илме те хăрушă» (cu_2, 2379 *слов*);

на шугнанском языке (Su): «Матнийен, часть 1» (su_1, 16906 *слов*); «Матнийен, часть 2» (su_2, 19354 *слова*);

на якутском языке (Ya): «Кыайыы хоһоонноро» (ya_1, 3823 *слова*); «Көмүс содула» (ya_2, 2441 *слово*).

4.1.2. ЦП произведений. В качестве учётных элементов для описания произведений взяты:

– башкирский язык: 42 буквы (а, б, в, г, ғ, д, ҙ, е, ё, ж, з, и, й, к, ҡ, л, м, н, ң, о, ө, п, р, с, ҫ, т, у, ү, ф, х, һ, ц, ч, ш, щ, ъ, ы, ь, э, ә, ю, я),

– белорусский язык: 32 буквы (а, б, в, г, д, е, ё, ж, з, і, й, к, л, м, н, о, п, р, с, т, у, ў, ф, х, ц, ч, ш, ы, ь, э, ю, я),

– болгарский язык: 30 букв (а, б, в, г, д, е, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ц, ч, ш, щ, ъ, ь, ю, я),

– ваханский язык: 46 букв (а, б, в, в̌, г, ғ̌, ғ, д, д̈, е, ё, ж, ж, з, з̌, и, й, к, қ, л, м, н, о, п, р, с, č, т, т̈, у, ф, х, х̌, ц, ц̌, ч, ӵ, ҷ, ҷ̈, щ, ш, ы, ә, э, ю, я),

– казахский язык: 42 буквы (а, ә, б, в, г, ғ, д, е, ё, ж, з, и, й, к, қ, л, м, н, ң, о, ө, п, р, с, т, у, ұ, ү, ф, х, һ, ц, ч, ш, щ, ъ, ы, і, ь, э, ю, я),

– язык коми: 35 букв (а, б, в, г, д, е, ё, ж, з, и, і, й, к, л, м, н, о, ӧ, п, р, с, т, у, ф, х, ц, ч, ш, щ, ъ, ы, ь, э, ю, я),

– кыргызский язык: 36 букв (а, б, в, г, д, е, ё, ж, з, и, й, к, л, м, н, ң, о, ө, п, р, с, т, у, ү, ф, х, ц, ч, ш, щ, ъ, ы, ь, э, ю, я),

– молдавский язык: 31 буква (а, б, в, г, д, е, ж, ӂ, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ц, ч, ш, ы, ь, э, ю, я),

– монгольский язык: 35 букв (а, б, в, г, д, е, ё, ж, з, и, й, к, л, м, н, о, ө, п, р, с, т, у, ү, ф, х, ц, ч, ш, щ, ъ, ы, ь, э, ю, я),

– русский язык: 33 буквы (а, б, в, г, д, е, ё, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ц, ч, ш, щ, ы, ь, ъ, э, ю, я),

– сербский язык: 30 букв (а, б, в, г, д, ђ, е, ж, з, и, ј, к, л, љ, м, н, њ, о, п, р, с, т, ћ, у, ф, х, ц, ч, џ, ш),

– таджикский язык: 35 букв (а, б, в, г, ғ, д, е, ё, ж, з, и, ӣ, й, к, қ, л, м, н, о, п, р, с, т, у, ӯ, ф, х, ҳ, ч, ҷ, ш, ъ, э, ю, я),

– татарский язык: 39 букв (а, ә, б, в, г, д, е, ё, ж, ж, з, и, й, к, л, м, н, ң, о, ө, п,

р, с, т, у, ү, ф, х, h, ц, ч, ш, щ, ъ, ы, ь, э, ю, я),

– удмуртский язык: 38 букв (а, б, в, г, д, е, ё, ж, ӝ, з, ӟ, и, й, й, к, л, м, н, о, ӧ, п, р, с, т, у, ф, х, ц, ч, ӵ, ш, щ, ъ, ы, ь, э, ю, я),

– узбекский язык: 35 букв (а, б, в, г, ғ, д, е, ё, ж, з, и, й, к, қ, л, м, н, о, п, р, с, т, у, ў, ф, х, ҳ, ц, ч, ш, ъ, ь, э, ю, я),

– украинский язык: 33 буквы (а, б, в, г, ґ, д, е, є, ж, з, и, і, ї, й, к, л, м, н, о, п, р, с, т, у, ф, х, ц, ч, ш, щ, ь, ю, я),

– чеченский язык: 34 буквы (а, б, в, г, д, е, ё, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ц, ч, ш, щ, ъ, ы, ь, э, ю, я, I),

– чувашский язык: 37 букв (а, ӑ, б, в, г, д, е, ё, ӗ, ж, з, и, й, к, л, м, н, о, п, р, с, ҫ, т, у, ӳ, ф, х, ц, ч, ш, щ, ъ, ы, ь, э, ю, я),

– шугнанский язык: 43 буквы (а, ā, б, в, в̌, г, ғ, г̌, ғ̌, д, д̌, е, ё, ж, з, з̌, и, ӣ, й, к, қ, л, м, н, о, п, р, с, т, т̌, у, ӯ, у̊, ф, х, ҳ, х̌, ц, ч, ҷ, ш, ъ, э),

– якутский язык: 38 букв (а, б, в, г, ҕ, д, е, ё, ж, з, и, й, к, л, м, н, ҥ, о, ө, п, р, с, h, т, у, ү, ф, х, ц, ч, ш, щ, ъ, ы, ь, э, ю, я).

Из 33 букв кириллицы современного русского языка общими для всех рассматриваемых текстов являются 26, именно: а, б, в, г, д, е, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ч, ш, ю, я.

Определение 4.1.1. *Цифровым портретом текста будем называть распределение в нём частотности 26 букв.*

ЦП текста T записывается в табличном виде:

$$\begin{aligned} N:&\quad 1 \quad 2 \quad \dots \quad 26 \\ P:&\quad p_1 \quad p_2 \quad \dots \quad p_{26}, \end{aligned} \tag{4.1}$$

в котором первая строка – номера букв, расположенных в алфавитном порядке, а вторая – относительные частотности букв в тексте T, причём $\sum_{k=1}^{26} p_k = 1$.

Одновременно с (4.1) ЦП представляется также в виде дискретной функции

$$F(s) = \sum_{k=1}^{S} p_k \qquad (s = 1, \dots, 26). \tag{4.2}$$

4.1.3. Расстояния между ЦПТ

Пусть T_1, T_2 – произвольная пара текстов, характеризуемых на основе единого алфавита, и

$$F^{(\alpha)}(s) = \sum_{k=1}^{S} p_k{}^{(\alpha)} \quad - \tag{4.3}$$

соответствующие им ЦП, представленные дискретными функциями, $\alpha = 1, 2,$ и $(s = 1, \dots, 26)$.

Определение 4.1.2. *Расстоянием между текстами* T_1 *и* T_2 *называется*

положительное число $\rho(T_1, T_2)$, *определяемое формулой*

$$\rho(T_1, T_2) = \sqrt{\frac{26}{2}} \max_S \left| F^{(1)}(s) - F^{(2)}(s) \right|. \qquad (4.4)$$

4.1.4. Гипотеза Н **«однородности» произведений**

Она привлекается для того, чтобы выделить характерную особенность текстов, предназначенную для построения математической модели распознавания языка произведений. Её мы формулируем в следующем виде.

ГИПОТЕЗА Н. *Произведения, написанные на одном языке, «однородны», а на разных языках – «неоднородны».*

Говоря об «однородности» произведений (текстов), мы имеем в виду их похожесть, одинаковость, сходность, однотипность, родственность и т.п.

4.1.5. Математическая модель Н **-гипотезы**

Пусть γ – некоторое положительное число.

Определение 4.1.3. *Тексты* T_1, T_2 *называются* γ*-однородными (написанными на одном языке), если*

$$\rho(T_1, T_2) \le \gamma, \qquad (4.5)$$

и γ*-неоднородными (написанными на разных языках), если*

$$\rho(T_1, T_2) > \gamma. \qquad (4.6)$$

Неравенства (4.5) и (4.6) являются математической интерпретацией (моделью) гипотезы Н. Это значит, что в дальнейшем мы приступаем к распознаванию языков произведений с помощью математического аппарата, названного γ-классификатором, описанной в §§ 1.3 и 1.4.

Определение 4.1.4. γ*-классификатор – зависящий от одного вещественного параметра* γ *алгоритм принятия решения об отнесении пары текстов* T_1 *и* T_2 *к одному или двум разным языкам.*

Очевидно, что от значения γ зависит однородность или неоднородность любой пары текстов, следовательно, и степень выполнимости гипотезы. Принадлежность двух текстов к одному языку в рамках математической модели означает справедливость неравенства (4.5), а к двум разным языкам – справедливость неравенства (4.6). Гипотеза Н может нарушаться для тех пар текстов одного и того же языка в случае, когда вместо неравенства (4.5) имеет место неравенство (4.6), а также в случае, когда два текста на разных языках удовлетворяют неравенству (4.5) вместо того, чтобы выполнялось неравенство (4.6).

Пусть $\tau = \tau(\gamma)$ – суммарное количество нарушений гипотезы H одновременно в двух случаях: невыполнения неравенства «однородности» в случае двух текстов, принадлежащих одному языку, и невыполнения неравенства «неоднородности» в случае двух текстов, принадлежащих разным языкам. Тогда для фиксированного γ показатель выполнения гипотезы будет определяться величиной π, задаваемой формулой

$$\pi = 1 - \tau(\gamma)/L,$$

где L – число взаимных расстояний между всеми парами текстов из подколлекции $\boldsymbol{C}$. Из этой формулы следует, что π может принимать значения из отрезка [0, 1], причём π=0, если τ=L, и π=1, если τ=0. В первом случае гипотезу H следует признать непригодной, а во втором – полностью согласованной с обучающей выборкой.

В связи с тем, что эффективность γ-классификатора зависит от значения параметра γ, представляет интерес найти такое его значение, при котором π принимает максимальное значение. Именно в этом и заключается суть настройки γ-классификатора на данных обучающей выборки. Если такая настройка будет приемлемой, то можно говорить о решении задачи обучения γ-классификатора и его предрасположенности к распознаванию языков произведений самых разнообразных коллекций.

4.1.6. Настройка классификатора на данных коллекции *C*. В процессе настройки выполняются следующие операции:

– предобработка экспериментальных данных путём удаления из всех произведений коллекции дополнительных сверх кириллического алфавита букв;

– вычисление ЦП (4.1) (частотности 26 кириллических букв) для всех 70 произведений коллекции ***C***;

– вычисление по формулам (4.2), (4.3) и (4.4) разных парных расстояний $\rho(T_1, T_2)$ между произведениями коллекции ***C*** (результаты проведенного эксперимента представлены в таблице 4.1);

Таблица 4.1. – Результаты экспериментов

Количество языков	**Количество текстов**	**Число взаимных расстояний – *L***	**τ-суммарное количество нарушений**	**Оптимальный γ-полуинтервал**	**π-эффективность распознавания языка**
5	10	45	0	[0.1455; 0.1638)	100
5	20	190	14	[0.1376; 0.1392)	93
5	40	780	63	[0.1375; 0.1377)	92
5	10	45	0	[0.1455; 0.1638)	100
10	20	190	3	[0.1455; 0.1508)	98
20	40	780	10	[0.1001; 0.1025)	99

– вычисление с помощью алгоритма настройки γ-классификатора оптимального интервала значений γ, для которого величина $\tau = \tau(\gamma)$ суммарного числа случаев нарушения гипотезы H достигает минимального значения и, следовательно, величина π показателя выполнения гипотезы H принимает максимальное значение.

На данных таблицы 4.1 получены следующие результаты:

– оптимальный полуинтервал значений γ оказывается в пределах

$$\gamma^{onm} \in [0.1001;\ 0.1638); \tag{4.7}$$

в соответствии с определением 4.1.3 это значит, что если расстояние $\rho(T_1, T_2)$ между двумя текстами не превосходит значение γ^{onm} из указанного полуинтервала, то пара текстов принадлежит к одному и тому же языку; если же превосходит, то принадлежат к разным языкам;

– наивысшее значение π=100% коэффициента эффективности распознавания языка текста реализуется на корпусах 5 языков с 10 текстами;

– коэффициент π эффективности распознавания языка произведений по объему выборок 5 языков с 20, 40 текстами определяется значениями от 92% до 93%, практически все нарушения имеются между текстами на русском и украинском языках. Это говорит о том, что эти языки очень близки;

– коэффициент π эффективности равен 98% и 99% при выборе корпуса текстов 10, 20 языков с 20, 40 текстами.

4.1.7. Тестирование. Итак, настройка (обучение) γ-классификатора на данных корпусах текстов ***C*** прошла успешно. Для тестирования классификатора выбрано случайным образом 3 текста:

на русском языке **(Ru):**

А.А. Фадеев «Разлив» (Text_Ru, 19075 *слов*);

на таджикском языке **(Tj):**

А. Фирдоуси «Рустам ва Сухроб» (Text_Tj, 16355 *слов*);

на узбекском языке **(Uz):**

К. Абдулла «Меҳробдан чаён» (Text_Uz, 59540 *слов*).

Отметим, что сведения относительно выбранных произведений описаны по той же схеме, что и для элементов коллекции ***C***.

Для трех произведений, предназначенных для тестирования, построены цифровые портреты (4.1) и затем по формулам (4.2), (4.3), (4.4) для каждого из них вычислены расстояния до 70 объектов коллекции ***C***. Соответствующие значения записаны в ячейках таблицы 4.2, расположенных на пересечениях строк и столбцов. Там же серым цветом выделены пары ячеек с минимальными расстояниями между тестируемыми и содержащимися в коллекции произведениями.

Таблица 4.2. – Расстояния между текстами коллекции ***C*** и 3 случайно выбранными тестируемыми произведениями

Тексты		**Text_Ru**	**Text_Tj**	**Text_Uz**
Ba	**ba_1**	0.2686	0.3834	0.4051
	ba_2	0.2609	0.4016	0.4100
Be	**be_1**	0.3734	0.2583	0.3851
	be_2	0.4020	0.2072	0.3663
	be_3	0.4348	0.1586	0.3321
	be_4	0.3855	0.1944	0.3329
	be_5	0.3997	0.1685	0.3394
	be_6	0.4377	0.1491	0.3095
	be_7	0.3959	0.1996	0.3102
	be_8	0.4065	0.2358	0.3769
Bo	**bo_1**	0.2715	0.2734	0.2006
	bo_2	0.2268	0.3311	0.2998
	bo_3	0.3122	0.2336	0.2553
	bo_4	0.2583	0.2794	0.2682
	bo_5	0.2122	0.2939	0.2912
	bo_6	0.2378	0.3489	0.3335
	bo_7	0.2208	0.3465	0.3358
	bo_8	0.2245	0.3343	0.3344
Vs	**vs_1**	0.2582	0.3486	0.4297
	vs_2	0.2058	0.3014	0.3495
Kz	**kz_1**	0.4363	0.3598	0.3516
	kz_2	0.4669	0.3903	0.3955
Ko	**ko_1**	0.2806	0.4461	0.5391
	ko_2	0.3271	0.4494	0.5424
Ky	**ky_1**	0.3593	0.2828	0.2428
	ky_2	0.2761	0.2856	0.2872
Ml	**ml_1**	0.2253	0.6556	0.5207
	ml_2	0.2036	0.6151	0.4802
Mn	**mn_1**	0.6009	0.2065	0.2844
	mn_2	0.3631	0.2277	0.3618
Ru	**ru_1**	0.0741	0.4826	0.4864
	ru_2	0.0915	0.3819	0.3959
	ru_3	0.0630	0.4186	0.4474
	ru_4	0.0543	0.4649	0.4746
	ru_5	0.0619	0.4891	0.4636
	ru_6	**0.0300**	0.4753	0.4391
	ru_7	0.0424	0.4641	0.4467
	ru_8	0.0386	0.4213	0.4232
Se	**se_1**	0.2647	0.3331	0.2917
	se_2	0.2984	0.3470	0.2996
Tj	**tj_1**	0.5530	0.1612	0.2571
	tj_2	0.5124	0.1164	0.2871
	tj_3	0.4705	**0.0507**	0.3628
	tj_4	0.4265	0.1205	0.2564
	tj_5	0.5070	0.1136	0.2985
	tj_6	0.4660	0.1623	0.2798
	tj_7	0.3734	0.1258	0.3097
	tj_8	0.4853	0.1219	0.2555
Ta	**ta_7**	0.3342	0.2696	0.2667
	ta_8	0.3341	0.3032	0.2915

Тексты		Text_Ru	Text_Tj	Text_Uz
Ud	**ud_1**	0.2521	0.5681	0.5555
	ud_2	0.3331	0.6342	0.5836
Uz	**uz_1**	0.4073	0.3368	**0.0723**
	uz_2	0.3109	0.2754	0.1917
Uk	**uk_1**	0.1056	0.4032	0.4647
	uk_2	0.1223	0.3728	0.3806
	uk_3	0.1361	0.3715	0.4308
	uk_4	0.1775	0.4385	0.5149
	uk_5	0.1043	0.3616	0.4575
	uk_6	0.0725	0.4129	0.4628
	uk_7	0.0652	0.4156	0.4419
	uk_8	0.1017	0.3745	0.4404
Ce	**ce_1**	0.4896	0.2912	0.3255
	ce_2	0.5094	0.2405	0.2918
Cu	**cu_1**	0.4617	0.5784	0.6572
	cu_2	0.4530	0.5741	0.6485
Su	**su_1**	0.3064	0.2407	0.2353
	su_2	0.2860	0.2454	0.2648
Ya	**ya_1**	0.2972	0.5055	0.5985
	ya_2	0.4695	0.3549	0.4409

В первых трех столбцах ближайшими соседями [7-9, 248, 249] текстов Text_Ru, Text_Tj и Text_Uz являются соответственно ru_6, tj_3 и uz_1 на расстояниях соответственно 0.0300, 0.0507 и 0.0723 (в таблице отмечены серым цветом). Интересно то, что эти расстояния меньше γ^{onm}, см. (4.7). Полученный результат показывает, что по методу ближайшего соседа три случайно выбранных произведения оказались как раз однородные с ними по языку пары текстов исходной коллекции.

Заключение. В данном параграфе была исследована применимость γ-классификатора не для модельной коллекции, а для корпуса, состоящего из 70 текстов на 20 языках. Анализ результатов показал, что методика на основе γ-классификатора способна достигать точности 100%, что является наилучшим результатом на сегодняшний день.

Таким образом, математическая триада в составе ЦП текстов, представляемых распределениями частотности 26 кириллических букв, формул (4.1)-(4.3) для вычисления расстояний между текстами и алгоритмом для выявления однородных текстов оказалась подходящей для эффективного решения поставленной задачи. Эти исследования показывают, что можно создать единый алфавит для языков.

В свою очередь, метод ближайшего соседа показал возможность безошибочного распределения дополнительных трех произведений по языкам.

Автор выражает уверенность в том, что еще увеличение объема исходной коллекции текстов не станет препятствием для успешного применения γ-классификатора не только для распознавания языков, но также и для самых

разнообразных однородностей текстовых документов.

§ 4.2. Исследование статистических закономерностей идентификация языка произведений на основе латинского алфавита в корпусах текстов художественной литературы

В данном параграфе устанавливается применимость γ-классификатора для автоматического распознавания языка произведения на основе частотности общих 26 латинских алфавитных букв на примере корпуса, состоящего из 70 текстов на 20 языках (по 8 произведений на 5 языках: английском, венгерском, латинском, литовском и голландском, и по 2 произведения на других 15 языках) с использованием латинской графики. Математическая модель γ-классификатора представляется в виде триады. Её первым компонентом является ЦПТ – распределение в тексте частотности буквенных униграмм; в качестве второго компонента служит формула для вычисления расстояний между ЦПТ и третий компонент – алгоритм машинного обучения, реализующий гипотезу «однородности» произведений, написанных на одном языке, и «неоднородности» произведений, написанных на разных языках. Настройка алгоритма, использующего таблицу парных расстояний между всеми произведениями коллекции текстов, заключалась в определении оптимального значения вещественного параметра γ, для которого минимизируется ошибка нарушения гипотезы «однородности». На примере корпуса рассмотрены случаи с 5, 10, 20 предполагаемыми языками, а также с 10, 20, 40 текстами, выявляются особенности применения γ-классификатора при распознавании языка текста. Для тестирования классификатора дополнительно было выбрано четыре случайных текста, которые составлены теми же языками, что и тексты коллекции. Методом ближайшего (по расстоянию) соседа четыре случайных текста проверяется на однородность с соответствующими парами одноязычных произведений. Приводятся результаты экспериментов по применению γ-классификатора на корпусе текстов художественной литературы.

В наше время письменность на основе латинского алфавита получила широкое распространение среди романской, германской славянской, финно-угорской, тюркской, семитской и иранской групп языков, среди стран Индокитая, Зондского архипелага и Филиппин, Африки (южнее Сахары), Америки, Австралии и Океании, [273]. За исключением современного английского, для большинства других языков латинский алфавит из 26 букв (a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z) оказался недостаточным, в связи с чем для отражения фонетических особенностей тех или иных языковых систем к базовой латинской графике были добавлены различные диакритические знаки, лигатуры и другие модификации букв.

Задача, решением которой будем заниматься в этом параграфе, состоит в том, чтобы определить, можно ли обойтись только лишь 26 латинскими буквами для автоматического распознавания языка, на котором написана та или иная печатная продукция.

Приступая к решению поставленной задачи, отметим, что в качестве исследовательского инструмента мы будем использовать *математическую триаду* в составе ЦПТ, представляемых распределениями частотности 26 латинских букв, формулы для вычисления расстояний между текстами и алгоритма для выявления однородных текстов, см. §§ 1.3-1.4. Упомянутая триада с момента своего появления в 2017 году применялась, прежде всего, для распознавания авторства для различных вариантов ЦП текстов [255-324]. В дополнении к сказанному уместно отметить, что в монографии [227] представлен обширный обзор работ по идентификации авторов текстов на основе разнообразных ЦП текстов и применяемых методов классификации. Отметим, что ранее аналогичный вопрос изучался именно для символьных (буквенных) униграмм в §§ 3.2.3 и 3.2.5. Существенным моментом в сравнении с нашим предыдущим исследованием в § 3.2 является изучение вопроса не в модельной коллекции текстов, а в корпусах произведений художественной литературы, для которого удается получить удовлетворительный результат решения рассматриваемой задачи.

Состояние работ по применению различных классификаторов, прежде всего методов нейронных сетей и машин опорных векторов, подробно описано в монографии [227]. В этом параграфе на примере корпуса, состоящего из 70 произведений на 20 разных языках, решаются две задачи:

– *настроить так называемый γ-классификатор, по возможности, для безошибочного распознавания принадлежности текстов соответствующих языков путем подбора вещественного параметра γ;*

– *проверить правильность работы настроенного классификатора для четырёх дополнительных случайно выбранных произведений, принадлежащих различным языкам.*

Прежде чем переходить к изучению задач, напомним основные понятия, связанные с компонентами триады.

4.2.1. Корпус текстов $\boldsymbol{C}$ для исследований. В приводимом далее списке элементов коллекции $\boldsymbol{C}$ указываются имя автора, название его сочинения на родном языке и в скобках – аббревиатура сочинения и его размеры в количестве слов:

на азербайджанском языке (Az): А. Айлисли «Daş yuxular» (az_1, 17569 *слов*); N. Gyandzhevi «Leyli-va-Macnun» (az_2, 21027 *слов*);

на английском языке (En): У. Шекспир «Romeo and Juliet» (en_1, 25832

слова); M. Твейн «A Connecticut Yankee in King Arthur's Court» (en_2, 21705 *слов*); Дж. Лондон «The Call of the Wild» (en_3, 16348 *слов*); A. Poe «The Mystery of Marie Roget» (en_4, 19688 *слов*); A. Poe «The Unparalleled Adventures of One Hans Pfaal» (en_5, 18561 *слово*); C. Dickens «A Christmas Carol» (en_6, 28257 *слов*); C. Dickens «Some Short Christmas Stories» (en_7, 20956 *слов*); D. Simak «Installment Plan» (en_8, 18087 *слов*);

на венгерском языке (Vn): J. Benzoni «Az átok» (vn_1, 12190 *слов*); J. Benzoni «A templomosok kincse» (vn_2, 20538 *слов*); A. Sztrugackij «A bíborszínű felhők bolygója» (vn_3, 25603 *слова*); K. Bulicsov «Kettészakított élet» (vn_4, 16763 *слова*); I. Jefremov «Csillaghajók» (vn_5, 24468 *слов*); S. Lem «Kiberiáda» (vn_6, 17832 *слова*); S. Lem «Pirx pilóta kalandjai» (vn_7, 15295 *слов*); G. Martinov «220 nap az űrhajón» (vn_8, 19465 *слов*);

на исландском языке (Is): J.R.R. Tolkien «Hobbitinn, часть 1-2» (is_1, 15619 *слов*); J.R.R. Tolkien «Hobbitinn, часть 3-5» (is_2, 14895 *слов*);

на испанском языке (Es): Д.Дж. Генрих «El ocaso de la magia» (es_1, 14401 *слово*); В.Ф. Альберто «Oceano» (es_2, 24611 *слово*);

на итальянском языке (It): Г. Эд «Elminster: la nascita di un mago, Parte I» (it_1, 23604 *слова*); C. Роберт «Il paradosso del passato» (it_2, 21074 *слова*);

на латинском языке (Lt): S. Boethius «De philosophiae consolatione» (lt_1, 24680 *слов*); IV. Carolus «Vita Caroli» (lt_2, 15144 *слова*); S. Lucilio «Ad Lucilium Epistulae Morales, Parte I» (lt_3, 11770 *слов*); S. Lucilio «Ad Lucilium Epistulae Morales, Parte II» (lt_4, 17326 *слов*); N. Hussoviani «Carmen de Bisontis» (lt_5, 7152 *слова*); S. Lucilio «Ad Lucilium Epistulae Morales, Parte III» (lt_6, 24291 *слово*); S. Lucilio «Ad Lucilium Epistulae Morales, Parte IV» (lt_7, 25667 *слов*); S. Lucilio «Ad Lucilium Epistulae Morales, Parte V» (lt_8, 21150 *слов*);

на литовском языке (Li): A. Gutje «Mėlynas rūkas» (li_1, 20262 *слова*); A. Marinina «Triju ne desnis, Parte I» (li_2, 25486 *слов*); A. Marinina «Triju ne desnis, Parte II» (li_3, 18507 *слов*); A. Marinina «Triju ne desnis, Parte III» (li_4, 20701 *слово*); D.R.R. Tolkinas «Žiedo brolija» (li_5, 13714 *слова*); D.R.R. Tolkinas «Dvi Tvirtovės, Parte I» (li_6, 15001 *слово*); D.R.R. Tolkinas «Karaliaus sugryžimas» (li_7, 16713 *слова*); D.R.R. Tolkinas «Dvi Tvirtovės, Parte II» (li_8, 14400 *слов*);

на немецком языке (De): Г. Пиз «Schiff ohne Mannschaft» (de_1, 25407 *слов*); Г. Диана «Das flammende Kreuz: Roman» (de_2, 24057 *слов*);

на нидерландском (голландском) языке (Ni): P. Aspe «De kinderen van Chronos» (ni_1, 14082 *слова*); R. Jordan «Vuur uit de hemel» (ni_2, 19214 *слова*); R. Jordan «Hart van de Winter» (ni_3, 18749 *слов*); R. Jordan «Viersprong van de Schemer» (ni_4, 12931 *слово*); R. Jordan «De Torens van Middernacht» (ni_5, 29318 *слов*); R. Jordan «Het Licht van Weleer» (ni_6, 29620 *слов*); A. West «Aardmagiër» (ni_7, 19563 *слова*); A. West «De watermagiër» (ni_8, 25483 *слова*);

на норвежском языке (No): A. Holte «En god soldat, Parte I» (no_1, 21703 *слова*); A. Holte «En god soldat, Parte II» (no_2, 24809 *слов*);

на польском языке (Pl): R.M. Wegner «Jeszcze może załopotać, Parte I» (pl_1, 10601 *слово*); R.M. Wegner «Jeszcze może załopotać, Parte II» (pl_2, 9670 *слов*);

на португальском языке (Pr): J. Belfort «A caçada ao lobo de Wall Street» (pr_1, 22728 *слов*); P. Coelho «As Valkirias» (pr_2, 17089 *слов*);

на румынском языке (Ro): Д. Роберт «În căutarea Cornului» (ro_1, 23951 *слово*); П. Камил «Ultima noapte de dragoste, întâia noapte de război» (ro_2, 20515 *слов*);

на словацком языке (Sv): I.A. Jefremov «Na hranici Oekumeny» (sv_1, 13534 *слова*); J. Jesenský «Demokrati» (sv_2, 17113 *слова*);

на финском языке (Fi): A. Paasilinna «Hirtettyjen kettujen metsä, Parte I» (fi_1, 21781 *слово*); A. Paasilinna «Hirtettyjen kettujen metsä, Parte II» (fi_2, 20049 *слов*);

на французском языке (Fr): С. Жорж «Lavinia» (fr_1, 13151 *слово*); Б. Мишель «Les Nymphéas noirs» (fr_2, 27621 *слово*);

на чешском языке (Cs): S. Lem «K Mrakům Magellanovým» (cs_1, 17552 *слова*); B.S.R. Jordan «Bouře přichází» (cs_2, 17439 *слов*);

на шведском языке (Sd): J. Flanagan «Den nya lärlingen» (sd_1, 30411 *слово*); L. Kepler «Paganinikontraktet» (sd_2, 19847 *слов*);

на языке эсперанто (Ep): J. Valano «Ĉu vi kuiras ĉine?» (ep_1, 29298 *слов*); J. Valano «Tien» (ep_2, 22012 *слова*).

4.2.2. ЦП произведений. В качестве учётных элементов для описания произведений взяты:

– азербайджанский язык: 32 буквы (a, b, c, ç, d, e, ə, f, g, ğ, h, x, ı, i, j, k, q, l, m, n, o, ö, p, r, s, ş, t, u, ü, v, y, z),

– английский язык: 26 букв (a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z),

– венгерский язык: 35 букв (a, á, b, c, d, e, é, f, g, h, i, í, j, k, l, m, n, o, ó, ö, ő, p, r, s, t, u, ú, ü, ű, v, q, w, x, y, z),

– исландский язык: 32 буквы (a, á, þ, æ, b, d, ð, e, é, f, g, h, i, í, j, k, l, m, n, o, ó, ö, p, r, s, t, u, ú, v, x, y, ý),

– испанский язык: 27 букв (a, b, c, d, e, f, g, h, i, j, k, l, m, n, ñ, o, p, q, r, s, t, u, v, w, x, y, z),

– итальянский язык: 36 букв (a, à, b, c, d, e, è, é, f, g, h, i, ì, í, î, j, k, l, m, n, o, ò, ó, p, q, r, s, t, u, ù, ú, v, w, x, y, z),

– латинский язык: 26 букв (a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z),

– литовский язык: 32 буквы (a, ą, b, c, č, d, e, ę, ė, f, g, h, i, į, y, j, k, l, m, n, o, p, r, s, š, t, u, ų, ū, v, z, ž),

– немецкий язык: 30 букв (a, ä, b, c, d, e, f, g, h, i, j, k, l, m, n, o, ö, p, q, r, s, ß, t, u, ü, v, w, x, y, z),

– нидерландский (Голландский) язык: 26 букв (a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z),

– норвежский язык: 29 букв (a, å, æ, b, c, d, e, f, g, h, i, j, k, l, m, n, o, ø, p, q, r, s, t, u, v, w, x, y, z),

– польский язык: 35 букв (a, ą, b, c, ć, d, e, ę, f, g, h, i, j, k, l, ł, m, n, ń, o, ó, p, q, r, s, ś, t, u, v, w, x, y, z, ź, ż),

– португальский язык: 38 букв (a, á, â, ã, à, b, c, ç, d, e, é, ê, f, g, h, i, í, j, k, l, m, n, o, ó, ô, õ, p, q, r, s, t, u, ú, v, w, x, y, z),

– румынский язык: 31 буква (a, ă, â, b, c, d, e, f, g, h, i, î, j, k, l, m, n, o, p, q, r, s, ș, t, ț, u, v, w, x, y, z),

– словацкий язык: 43 буквы (a, á, ä, b, c, č, d, ď, e, é, f, g, h, i, í, j, k, l, Ĺ, ľ, m, n, ň, o, ó, ô, p, q, r, ŕ, s, š, t, ť, u, ú, v, w, x, y, ý, z, ž),

– финский язык: 31 буква (a, å, ä, b, c, d, e, f, g, h, i, j, k, l, m, n, o, ö, p, q, r, s, š, t, u, v, w, x, y, z, ž),

– французский язык: 40 букв (a, â, à, b, c, ç, d, e, é, ê, è, ë, f, g, h, i, î, ï, j, k, l, m, n, o, ô, p, q, r, s, t, u, û, ù, ü, v, w, x, y, ÿ, z),

– чешский язык: 41 буква (a, á, b, c, č, d, ď, e, é, ě, f, g, h, i, í, j, k, l, m, n, ň, o, ó, p, q, r, ř, s, š, t, ť, u, ú, ů, v, w, x, y, ý, z, ž),

– шведский язык: 29 букв (a, å, ä, b, c, d, e, f, g, h, i, j, k, l, m, n, o, ö, p, q, r, s, t, u, v, w, x, y, z),

– язык эсперанто: 28 букв (a, b, c, ĉ, d, e, f, g, ĝ, h, ĥ, i, j, ĵ, k, l, m, n, o, p, r, s, ŝ, t, u, ŭ, v, z).

Из 26 букв латиницы современного английского языка общими для всех рассматриваемых текстов являются все 26, а именно: a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z.

Определение 4.2.1. *Цифровым портретом текста будем называть распределение в нём частотности 26 букв.*

ЦП текста T записывается в табличном виде:

$$\begin{array}{ll} N: & 1 \quad 2 \quad \ldots \quad 26 \\ P: & p_1 \quad p_2 \quad \ldots \quad p_{26}, \end{array} \tag{4.8}$$

в котором первая строка – номера букв, расположенных в алфавитном порядке, а вторая – относительные частотности букв в тексте T, причём $\sum_{k=1}^{26} p_k = 1$.

Одновременно с (4.8) ЦП представляется также в виде дискретной функции

$$F(s) = \sum_{k=1}^{S} p_k \qquad (s = 1, \ldots, 26). \tag{4.9}$$

4.2.3. Расстояния между ЦПТ. Пусть T_1, T_2 – произвольная пара текстов, характеризуемых на основе единого алфавита, и

$$F^{(\alpha)}(s) = \sum_{k=1}^{S} p_k^{(\alpha)} \quad - \tag{4.10}$$

соответствующие им ЦП, представленные дискретными функциями, $\alpha = 1, 2,$ и $(s = 1, \ldots, 26).$

Определение 4.2.2. *Расстоянием между текстами* T_1 *и* T_2 *называется положительное число* $\rho(T_1, T_2)$*, определяемое формулой*

$$\rho(T_1, T_2) = \sqrt{\frac{26}{2}} \max_s \left| F^{(1)}(s) - F^{(2)}(s) \right|. \tag{4.11}$$

4.2.4. Гипотеза Н **«однородности» произведений.** Она привлекается для того, чтобы выделить характерную особенность текстов, предназначенную для построения математической модели распознавания языка произведений. Её мы формулируем в следующем виде.

ГИПОТЕЗА Н. *Произведения, написанные на одном языке, «однородны», а на разных языках – «неоднородны».*

Говоря об «однородности» произведений (текстов), мы имеем в виду их похожесть, одинаковость, сходность, однотипность, родственность и т.п.

4.2.5. Математическая модель Н**-гипотезы.** Пусть γ – некоторое положительное число.

Определение 4.2.3. *Тексты* T_1, T_2 *называются* γ*-однородными (написанными на одном языке), если*

$$\rho(T_1, T_2) \le \gamma, \tag{4.12}$$

и γ*-неоднородными (написанными на разных языках), если*

$$\rho(T_1, T_2) > \gamma. \tag{4.13}$$

Неравенства (4.12) и (4.13) являются математической интерпретацией (моделью) гипотезы Н. Это значит, что в дальнейшем мы приступаем к распознаванию языков произведений с помощью математического аппарата, названного γ-классификатором, см. § 1.4.

Определение 4.2.4. γ*-классификатор – зависящий от одного вещественного параметра* γ *алгоритм принятия решения об отнесении пары текстов* T_1 *и* T_2 *к одному или двум разным языкам.*

Очевидно, что от значения γ зависит однородность или неоднородность любой пары текстов, следовательно, и степень выполнимости гипотезы.

Принадлежность двух текстов к одному языку в рамках математической модели означает справедливость неравенства (4.12), а к двум разным языкам – справедливость неравенства (4.13). Гипотеза H может нарушаться для тех пар текстов одного и того же языка в случае, когда вместо неравенства (4.12) имеет место неравенство (4.13), а также в случае, когда два текста на разных языках удовлетворяют неравенству (4.12) вместо того, чтобы выполнялось неравенство (4.13).

Пусть $\tau = \tau(\gamma)$ – суммарное количество нарушений гипотезы H одновременно в двух случаях: невыполнения неравенства «однородности» в случае двух текстов, принадлежащих одному языку, и невыполнения неравенства «неоднородности» в случае двух текстов, принадлежащих разным языкам. Тогда для фиксированного γ показатель выполнения гипотезы будет определяться величиной π, задаваемой формулой

$$\pi = 1 - \tau(\gamma)/L,$$

где L – число взаимных расстояний между всеми парами текстов из подколлекции $\boldsymbol{C}$. Из этой формулы следует, что π может принимать значения из отрезка [0, 1], причём π=0, если τ=L, и π=1, если τ=0. В первом случае гипотезу H следует признать непригодной, а во втором – полностью согласованной с обучающей выборкой.

В связи с тем, что эффективность γ-классификатора зависит от значения параметра γ, представляет интерес найти такое его значение, при котором π принимает максимальное значение. Именно в этом и заключается суть настройки γ-классификатора на данных обучающей выборки. Если такая настройка будет приемлемой, то можно говорить о решении задачи обучения γ-классификатора и его предрасположенности к распознаванию языков произведений самых разнообразных коллекций.

4.2.6. Настройка классификатора на данных коллекции *C*. В процессе настройки выполняются следующие операции:

– предобработка экспериментальных данных путём удаления из всех произведений коллекции дополнительных сверх латинского алфавита букв;

– вычисление ЦП (4.8) (частотности 26 латинских букв) для всех 70 произведений коллекции $\boldsymbol{C}$;

– вычисление по формулам (4.9), (4.10) и (4.11) разных парных расстояний $\rho(T_1, T_2)$ между произведениями коллекции $\boldsymbol{C}$ (результаты проведенного эксперимента представлены в таблице 4.3);

Таблица 4.3. – Результаты экспериментов

Количество языков	**Количество текстов**	**Число взаимных расстояний – *L***	**τ-суммарное количество нарушений**	**Оптимальный γ-полуинтервал**	**π-эффективность распознавания языка**
5	10	45	0	[0.0929; 0.1777)	100
5	20	190	0	[0.1759; 0.1777)	100
5	40	780	0	[0.1759; 0.1777)	100
5	10	45	0	[0.0929; 0.1777)	100
10	20	190	0	[0.0929; 0.1096)	100
20	40	780	0	[0.0929; 0.0962)	100

– вычисление с помощью алгоритма настройки γ-классификатора оптимального интервала значений γ, для которого величина $\tau = \tau(\gamma)$ суммарного числа случаев нарушения гипотезы H достигает минимального значения и, следовательно, величина π показателя выполнения гипотезы H принимает максимальное значение.

На данных таблицы 4.3 получены следующие результаты:

– оптимальный полуинтервал значений γ оказывается в пределах

$$\gamma^{onm} \in [0.0929;\ 0.1777); \qquad (4.14)$$

в соответствии с определением 4.2.3 это значит, что если расстояние $\rho(T_1, T_2)$ между двумя текстами не превосходит значение γ^{onm} из указанного полуинтервала, то пара текстов принадлежит к одному и тому же языку; если же превосходит, то принадлежат к разным языкам;

– отметим, что для всех (без исключения) произведений коллекции ***C*** полностью подтвердилась гипотеза H и её математическая интерпретация в виде определения 4.2.3, и потому получено

$$\tau = \tau_{\min} = 0,$$

то есть, ни одно из неравенств (4.12) и (4.13) не было нарушено;

– вследствие чего показатель эффективности предложенной в настоящей работе математической модели распознавания языка произведений оказался равным

$$\pi = \pi_{\max} = 1.$$

4.2.7. Тестирование. Итак, настройка (обучение) γ-классификатора на данных корпусах текстов ***C*** прошла успешно. Для тестирования классификатора выбрано случайным образом 4 текста:

на немецком языке (De): М. Вилли «Die seltsamen Reisen des Marco Polo» (Text_De, 126607 *слов*);

на испанском языке (Es): Д. Арне «Misterioso» (Text_Es, 106835 *слов*);

на французском языке (Fr): К.С. Доминикович «Fantôme» (Text_Fr, 46089 *слов*);

на итальянском языке (It): Ш. Боб «Sfida al cielo» (Text_It, 101154 *слова*).

Отметим, что сведения относительно выбранных произведений описаны по той же схеме, что и для элементов коллекции ***C***.

Для четырёх произведений, предназначенных для тестирования, построены цифровые портреты (4.8) и затем по формулам (4.9), (4.10), (4.11) для каждого из них вычислены расстояния до 70 объектов коллекции ***C***. Соответствующие значения записаны в ячейках таблицы 4.4, расположенных на пересечениях строк и столбцов. Там же серым цветом выделены пары ячеек с минимальными расстояниями между тестируемыми и содержащимися в коллекции произведениями.

Таблица 4.4. – Расстояния между текстами коллекции ***C*** и 4 случайно выбранными тестируемыми произведениями

Тексты		**Text_De**	**Text_Es**	**Text_Fr**	**Text_It**
Az	**az_1**	0.4176	0.2668	0.3952	0.2335
	az_2	0.4393	0.2827	0.3899	0.2396
En	**en_1**	0.4069	0.3235	0.1477	0.2378
	en_2	0.3141	0.2513	0.2019	0.1995
	en_3	0.2619	0.2109	0.2916	0.2237
	en_4	0.3653	0.2628	0.2036	0.2063
	en_5	0.3221	0.2422	0.1995	0.1813
	en_6	0.3501	0.2760	0.2133	0.1940
	en_7	0.3179	0.2536	0.2097	0.1872
	en_8	0.3154	0.2354	0.2410	0.1997
Vn	**vn_1**	0.4618	0.3196	0.2313	0.2821
	vn_2	0.4809	0.3244	0.2460	0.3044
	vn_3	0.5164	0.3752	0.2491	0.3056
	vn_4	0.4654	0.3255	0.2384	0.2562
	vn_5	0.4953	0.3338	0.2447	0.2835
	vn_6	0.4935	0.3583	0.2386	0.2938
	vn_7	0.5062	0.3509	0.2355	0.3024
	vn_8	0.5017	0.3379	0.2431	0.2860
Is	**is_1**	0.4565	0.5422	0.3508	0.3919
	is_2	0.4653	0.5509	0.3595	0.4007
Es	**es_1**	0.3023	**0.0572**	0.3264	0.1980
	es_2	0.3076	**0.0506**	0.3138	0.1872
It	**it_1**	0.2812	0.1506	0.3119	**0.0385**
	it_2	0.3349	0.1803	0.2910	**0.0370**
Lt	**lt_1**	0.4929	0.3270	0.1272	0.3306
	lt_2	0.4620	0.2961	0.0963	0.2731
	lt_3	0.4863	0.3204	0.1206	0.3026
	lt_4	0.4901	0.3242	0.1244	0.3004
	lt_5	0.5081	0.3549	0.1062	0.3616
	lt_6	0.5032	0.3373	0.1375	0.3005
	lt_7	0.4954	0.3295	0.1297	0.3095

Тексты		Text_De	Text_Es	Text_Fr	Text_It
	lt_8	0.4920	0.3262	0.1263	0.3043
Li	**li_1**	0.6379	0.4720	0.2722	0.3401
	li_2	0.6517	0.4858	0.2864	0.3538
	li_3	0.6528	0.4870	0.2871	0.3550
	li_4	0.6544	0.4886	0.2891	0.3566
	li_5	0.5879	0.4221	0.2877	0.2900
	li_6	0.6209	0.4681	0.2924	0.3331
	li_7	0.6132	0.4499	0.2742	0.3153
	li_8	0.5682	0.4074	0.2832	0.2724
De	**de_1**	**0.0297**	0.2666	0.4242	0.2874
	de_2	**0.0536**	0.2563	0.4139	0.2770
Ni	**ni_1**	0.1322	0.2705	0.3839	0.2402
	ni_2	0.1131	0.2554	0.4262	0.2755
	ni_3	0.1170	0.2491	0.4131	0.2778
	ni_4	0.1109	0.2549	0.4242	0.2757
	ni_5	0.1329	0.2681	0.4343	0.2889
	ni_6	0.1356	0.2785	0.4429	0.2992
	ni_7	0.1088	0.2462	0.4188	0.2577
	ni_8	0.1096	0.2530	0.4136	0.2616
No	**no_1**	0.3556	0.3940	0.2271	0.3057
	no_2	0.3520	0.3920	0.2252	0.3046
Pl	**pl_1**	0.4860	0.5240	0.5469	0.5504
	pl_2	0.4700	0.5118	0.5348	0.5382
Pr	**pr_1**	0.3778	0.1159	0.2511	0.1391
	pr_2	0.4525	0.1757	0.2376	0.1868
Ro	**ro_1**	0.2585	0.1428	0.2579	0.1275
	ro_2	0.2564	0.1673	0.2743	0.1232
Sv	**sv_1**	0.5033	0.3821	0.2486	0.2877
	sv_2	0.5133	0.3569	0.2590	0.2977
Fi	**fi_1**	0.7533	0.6178	0.4315	0.4929
	fi_2	0.7412	0.6121	0.4258	0.4873
Fr	**fr_1**	0.4101	0.2712	**0.0460**	0.2501
	fr_2	0.4263	0.2784	**0.0347**	0.2686
Cs	**cs_1**	0.6617	0.4442	0.2830	0.4460
	cs_2	0.5852	0.3743	0.2623	0.3696
Sd	**sd_1**	0.2895	0.2542	0.2829	0.1461
	sd_2	0.4032	0.3363	0.2310	0.1875
Ep	**ep_1**	0.5929	0.4270	0.3330	0.2950
	ep_2	0.5852	0.4194	0.3517	0.2874

В первых четырёх столбцах ближайшими соседями [7-9, 248, 249] текстов Text_De, Text_Es, Text_Fr и Text_It являются соответственно de_1, es_2, fr_2 и it_2 на расстояниях соответственно 0.0297, 0.0506, 0.0347 и 0.0370 (в таблице отмечены серым цветом). Интересно в том, что эти расстояния меньше γ^{onm}, см. (4.14). Полученный результат показывает, что по методу ближайшего соседа четыре случайно выбранных произведения оказались как раз однородные с ними по языку пары текстов исходной коллекции.

Заключение. В этом параграфе была исследована применимость γ-классификатора не для модельной коллекции, а для корпуса, состоящего из 70

текстов на 20 языках. Анализ результатов показал, что методика на основе γ-классификатора способна достигать точности 100%, что является наилучшим результатом на сегодняшний день.

Таким образом, математическая триада в составе ЦП текстов, представляемых распределениями частотности 26 латинских букв, формул (4.8) – (4.10) для вычисления расстояний между текстами и алгоритмом для выявления однородных текстов оказалась подходящей для эффективного решения поставленной задачи. Эти исследования показывают, что можно создать единый алфавит для языков.

В свою очередь, метод ближайшего соседа показал возможность безошибочного распределения дополнительных четырёх произведений по языкам.

Автор выражает уверенность в том, что еще увеличение объема исходной коллекции текстов не станет препятствием для успешного применения γ-классификатора не только для распознавания языков, но также и для самых разнообразных однородностей текстовых документов.

§ 4.3. Исследование статистических закономерностей определения автора произведений в корпусах текстов художественной литературы

В данном параграфе устанавливается применимость γ-классификатора для автоматического распознавания автора произведения на основе частотности 26 кириллических алфавитных букв на примере корпуса, состоящего из 70 поэтических текстов 20 таджикско-персидских авторов (по 8 произведений 5 авторов: А. Суруш, А. Фирдоуси, К. Худжанди, Л. Шерали и Дж. Руми, и по 2 произведения от других 15 авторов) с использованием кириллической графики. На примере корпуса рассмотрены случаи с 5, 10, 20 предполагаемыми авторами, а также с 10, 20, 40 текстами, выявляются особенности применения γ-классификатора при распознавании автора текста. Для тестирования классификатора дополнительно было выбрано три случайных текста, которые составлены теми же авторами, что и тексты коллекции. Методом ближайшего (по расстоянию) соседа три случайных текста проверяются на однородность с соответствующими парами произведений авторов. Приводятся результаты экспериментов по применению γ-классификатора на корпусе текстов художественной литературы.

С момента появления в 2017 г. γ-классификатор из §§ 1.3-1.4 широко используется при решении различных задач автоматического распознавания текста, см. например, [255-324]. Задача, решением которой будем заниматься в этом параграфе, состоит в том, чтобы определить, можно ли обойтись только лишь 26 (а, б, в, г, д, е, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ч, ш, ю, я) кириллическими буквами для автоматического распознавания автора

произведения. Отметим, что ранее аналогичный вопрос изучался именно для символьных (буквенных) униграмм в главе 2 § 2.1.1. Существенным моментом в сравнении с нашим предыдущим исследованием из § 2.1.1 является изучение вопроса не в модельной коллекции текстов, а в корпусах произведений художественной литературы, для которого удается получить удовлетворительный результат решения рассматриваемой задачи.

Состояние работ по применению различных классификаторов, прежде всего методов нейронных сетей и машин опорных векторов, подробно описано в монографии [227]. В этом параграфе на примере корпуса, состоящего из 70 произведений 20 авторов, решаются две задачи:

– *настроить так называемый γ-классификатор, по возможности, для безошибочного распознавания принадлежности текстов соответствующих авторов путем подбора вещественного параметра γ;*

– *проверить правильность работы настроенного классификатора для трёх дополнительных случайно выбранных произведений, принадлежащих различным авторам.*

Прежде чем перейти к изучению задач, напомним основные понятия, связанные с компонентами триады.

4.3.1. Корпус текстов $\boldsymbol{C}$ для исследований.

В приводимом далее списке элементов коллекции $\boldsymbol{C}$ указываются имя автора, название его сочинения на родном языке и в скобках – аббревиатура сочинения и его размеры в количестве слов:

– *А. Бедил (АБ):* «Ғазалиёт» (Ғ, 4700 *слов*), «Рубоӣ» (Р, 3962 *слова*);

– *А. Деҳлавӣ (АД):* «Маҷмӯи ғазалҳо, қисми 1» (МҒ1, 10307 *слов*), «Маҷмӯи ғазалҳо, қисми 2» (МҒ2, 9203 *слова*);

– *А. Рӯдакӣ (АР):* «Адабиёти пароканда» (АП, 5154 *слова*), «Рубоиёт» (Р, 5511 *слово*);

– *А. Суруш (АС):* «Дафтари аввал» (Д1, 6211 *слово*), «Дафтари дуввум» (Д2, 6177 *слов*), «Дафтари сеюм» (Д3, 6030 *слов*), «Дафтари чорум» (Д4, 6073 *слова*), «Дафтари панҷум» (Д5, 6580 *слов*), «Дафтари шашум» (Д6, 6179 *слов*), «Дафтари ҳафтум» (Д7, 5869 *слов*), «Дафтари ҳаштум» (Д8, 5908 *слов*);

– *А. Фирдавсӣ (АФ):* «Достони Бежан бо Манижа» (Б&М, 15862 *слова*), «Достони Разми Исфандиёр бо Рустам» (И&Р, 18716 *слов*), «Кайхусрав» (КВ, 18782 *слова*), «Достони Комуси Кашонӣ» (КК, 17454 *слова*), «Манучеҳр» (М, 22169 *слов*), «Достони Рустам ва Сӯҳроб» (Р&С, 16388 *слов*), «Достони Рустам бо Хоқони Чин» (Р&Х, 16722 *слова*), «Фаридун» (Ф, 12347 *слов*);

– *А. Ҷомӣ (АҶ):* «Лайлӣ ва Маҷнун» (Л&М, 23723 *слова*), «Юсуф ва Зулайхо» (Ю&З, 20212 *слова*);

– *Б. Зебуннисо (БЗ):* «Девони махфӣ, қисми 1» (ДМ1, 19646 *слов*), «Девони махфӣ, қисми 2» (ДМ2, 19489 *слов*);

– *И. Фарзона (ИФ):* «101-Ғазалҳо» (Ғ, 10198 *слов*), «Мӯҳри гули мино» (МГ, 22296 *слов*);

– *К. Хуҷандӣ (КХ):* «Ғазалиёт, қисми 1» (Ғ1, 13305 *слов*), «Ғазалиёт, қисми 2» (Ғ2, 13259 *слов*), «Ғазалиёт, қисми 3» (Ғ3, 12074 *слова*), «Ғазалиёт, қисми 4» (Ғ4, 14319 *слов*), «Ғазалиёт, қисми 5» (Ғ5, 14241 *слово*), «Ғазалиёт, қисми 6» (Ғ6, 14656 *слов*), «Ғазалиёт, қисми 7» (Ғ7, 13876 *слов*), «Ғазалиёт, қисми 8» (Ғ8, 14026 *слов*);

– *Л. Шералӣ (ЛШ):* «Куллиёт, қисми 1» (К1, 25280 *слов*), «Куллиёт, қисми 2» (К2, 22938 *слов*), «Куллиёт, қисми 3» (К3, 25129 *слов*), «Куллиёт, қисми 4» (К4, 24623 *слова*), «Куллиёт, қисми 5» (К5, 24152 *слова*), «Куллиёт, қисми 6» (К6, 16511 *слово*), «Куллиёт, қисми 7» (К7, 24370 *слов*), «Куллиёт, қисми 8» (К8, 19003 *слова*);

– *М. Қаноат (МҚ):* «Маҷмӯи шеърҳо, қисми 1» (МШ1, 6947 *слов*), «Маҷмӯи шеърҳо, қисми 2» (МШ2, 7936 *слов*);

– *М. Миршакар (ММ):* «Қишлоқи тиллоӣ, қисми 1» (ҚТ1, 10091 *слово*), «Қишлоқи тиллоӣ, қисми 2» (ҚТ2, 8404 *слова*);

– *М. Турсунзода (МТ):* «Мунтахаби Осор, қисми 1» (МО1, 19915 *слов*), «Мунтахаби Осор, қисми 2» (МО2, 19692 *слова*);

– *Н. Ганҷавӣ (НГ):* «Лайлӣ ва Маҷнун» (Л&М, 15557 *слов*), «Хусрав ва Ширин» (Х&Ш, 10410 *слов*);

– *Н. Қосим (НҚ):* «Малика Тӯрондухт, қисми 1» (МТ1, 8287 *слов*), «Малика Тӯрондухт, қисми 2» (МТ2, 8526 *слов*);

– *Н. Хисрав (НХ):* «Мунозира бо Худо» (МХ, 4719 *слов*), «Саодатнома» (С, 4157 *слов*);

– *С. Шерозӣ (СШ):* «Ғазалиёт, қисми 1» (Ғ1, 16016 *слов*), «Ғазалиёт, қисми 2» (Ғ2, 13266 *слов*);

– *У. Хайём (УХ):* «Рубоиёт, қисми 1» (Р1, 7224 *слов*), «Рубоиёт, қисми 2» (Р2, 6576 *слов*);

– *Ҳ. Шерозӣ (ҲШ):* «Ғазалиёт, қисми 1» (Ғ1, 21046 *слов*), «Ғазалиёт, қисми 2» (Ғ2, 19427 *слов*);

– *Ҷ. Румӣ (ҶР):* «Маснавии Маънавӣ, Дафтари 1» (ММ1, 24567 *слов*), «Маснавии Маънавӣ, Дафтари 2» (ММ2, 22270 *слов*), «Маснавии Маънавӣ, Дафтари 3» (ММ3, 24360 *слов*), «Маснавии Маънавӣ, Дафтари 4» (ММ4, 21341 *слово*), «Маснавии Маънавӣ, Дафтари 5» (ММ5, 21155 *слов*), «Маснавии Маънавӣ, Дафтари 6, қисми 1» (ММ6, 17490 *слов*), «Маснавии Маънавӣ, Дафтари 6, қисми 2» (ММ7, 17678 *слов*), «Маснавии Маънавӣ, Дафтари 6, қисми 3» (ММ8, 23112 *слова*).

4.3.2. ЦП произведений.

Из 35 букв кириллицы современного таджикского языка общими для всех рассматриваемых текстов являются 26, именно: а, б, в, г, д, е, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ч, ш, ю, я.

Определение 4.3.1. *Цифровым портретом текста будем называть распределение в нём частотности 26 букв.*

ЦП текста T записывается в табличном виде:

$$\begin{array}{ll} N: & 1 \quad 2 \quad \ldots \quad 26 \\ P: & p_1 \quad p_2 \quad \ldots \quad p_{26}, \end{array} \tag{4.15}$$

в котором первая строка – номера букв, расположенных в алфавитном порядке, а вторая – относительные частотности букв в тексте T, причём $\sum_{k=1}^{26} p_k = 1$.

Одновременно с (4.15) ЦП представляется также в виде дискретной функции

$$F(s) = \sum_{k=1}^{S} p_k \qquad (s = 1, \ldots, 26). \tag{4.16}$$

4.3.3. Расстояния между ЦПТ.

Пусть T_1, T_2 – произвольная пара текстов, характеризуемых на основе единого алфавита, и

$$F^{(\alpha)}(s) = \sum_{k=1}^{S} p_k^{(\alpha)} \quad - \tag{4.17}$$

соответствующие им ЦП, представленные дискретными функциями, $\alpha = 1, 2,$ и $(s = 1, \ldots, 26).$

Определение 4.3.2. *Расстоянием между текстами* T_1 *и* T_2 *называется положительное число* $\rho(T_1, T_2)$*, определяемое формулой*

$$\rho(T_1, T_2) = \sqrt{\frac{26}{2}} \max_s \left| F^{(1)}(s) - F^{(2)}(s) \right|. \tag{4.18}$$

4.3.4. Гипотеза Н **«однородности» произведений.**

Она привлекается для того, чтобы выделить характерную особенность текстов, предназначенную для построения математической модели распознавания автора произведений. Её мы формулируем в следующем виде.

ГИПОТЕЗА Н. *Произведения одного автора – «однородные», а разных авторов – «неоднородные».*

Говоря об «однородности» произведений (текстов), мы имеем в виду их похожесть, одинаковость, сходность, однотипность, родственность и т.п.

4.3.5. Математическая модель Н **-гипотезы.**

Пусть γ – некоторое положительное число.

Определение 4.3.3. *Тексты* T_1, T_2 *называются* γ*-однородными, если*

$$\rho(T_1, T_2) \le \gamma, \tag{4.19}$$

и γ*-неоднородными, если*

$$\rho(T_1, T_2) > \gamma. \tag{4.20}$$

Неравенства (4.19) и (4.20) являются математической интерпретацией (моделью) гипотезы Н. Это значит, что в дальнейшем мы приступаем к распознаванию авторов произведений с помощью математического аппарата, названного γ-классификатором, описанным в §§ 1.3-1.4.

Определение 4.3.4. *γ-классификатор – зависящий от одного вещественного параметра γ алгоритм принятия решения об отнесении пары текстов T_1 и T_2 к одному или двум разным авторам.*

Очевидно, что от значения γ зависит однородность или неоднородность любой пары текстов, следовательно, и степень выполнимости гипотезы. Однородность всех текстов одного автора в рамках математической модели означает справедливость неравенства (4.19), а неоднородность любых двух текстов разных авторов – справедливость неравенства (4.20). Гипотеза Н может нарушаться для каких-то пар текстов одного и того же автора в случае, когда вместо неравенства (4.19) имеет место неравенство (4.20), а также в случае, когда какие-то два текста двух различных авторов удовлетворяют неравенство (4.19) вместо того, чтобы выполнялось неравенство (4.20).

Пусть $\tau = \tau(\gamma)$ – суммарное количество нарушений гипотезы $\mathbb{H}$ одновременно в двух случаях: невыполнение неравенства «однородности» в случае двух текстов, принадлежащих одному автору, и невыполнение неравенства «неоднородности» в случае двух текстов, принадлежащих разным авторам. Тогда для фиксированного γ показатель выполнения гипотезы будет определяться величиной π, задаваемой формулой

$$\pi = 1 - \tau(\gamma)/L,$$

где *L* – число взаимных расстояний между всеми парами текстов из подколлекции ***C***. Из этой формулы следует, что π может принимать значения из отрезка [0, 1], причём π=0, если τ=*L*, и π=1, если τ=0. В первом случае гипотезу Н следует признать непригодной, а во втором – полностью согласованной с обучающей выборкой.

В связи с тем, что эффективность γ-классификатора зависит от значения параметра γ, представляет интерес найти такое его значение, при котором π принимает максимальное значение. Именно в этом и заключается суть настройки γ-классификатора на данных обучающей выборки. Если такая настройка будет приемлемой, то можно говорить о решении задачи обучения γ-классификатора и

его предрасположенности к распознаванию авторов произведений самых разнообразных коллекций.

4.3.6. Настройка классификатора на данных коллекции *C*.

В процессе настройки выполняются следующие операции:

– предобработка экспериментальных данных путём удаления из всех произведений коллекции дополнительных сверх кириллического алфавита букв;

– вычисление ЦП (4.15) (частотности 26 кириллических букв) для всех 70 произведений коллекции ***C***;

– вычисление по формулам (4.16), (4.17) и (4.18) разных парных расстояний $\rho(T_1, T_2)$ между произведениями коллекции ***C*** (результаты проведенного эксперимента представлены в таблице 4.5);

Таблица 4.5. – Результаты экспериментов

Количество авторов	**Количество текстов**	**Число взаимных расстояний – *L***	**τ-суммарное количество нарушений**	**Оптимальный γ-полуинтервал**	**π-эффективность распознавания автора**
5	10	45	1	[0.0435; 0.0479)	98
5	20	190	13	[0.0448; 0.0452)	93
5	40	780	75	[0.0451; 0.0452)	90
5	10	45	3	[0.0435; 0.0436)	93
10	20	190	7	[0.0357; 0.0364)	96
20	40	780	20	[0.0323; 0.0325)	97

– вычисление с помощью алгоритма настройки γ-классификатора оптимального интервала значений γ, для которого величина $\tau = \tau(\gamma)$ суммарного числа случаев нарушения гипотезы H достигает минимального значения и, следовательно, величина π показателя выполнения гипотезы H принимает максимальное значение.

На данных таблицы 4.5 получены следующие результаты:

– оптимальный полуинтервал значений γ оказывается в пределах

$$\gamma^{onm} \in [0.0323;\ 0.0479); \qquad (4.21)$$

в соответствии с определением 4.3.3 это значит, что если расстояние $\rho(T_1, T_2)$ между двумя текстами не превосходит значение γ^{onm} из указанного полуинтервала, то пара текстов принадлежит к одному и тому же автору; если же превосходит, то принадлежит к разным авторам;

– наивысшее значение π=98% коэффициента эффективности распознавания автора текста реализуется на корпусах 5 авторов с 10 текстами;

– коэффициент π эффективности распознавания автора произведений по объему выборок 5 авторов с 20, 40 текстами определяется значениями от 90% до 93%;

– коэффициент π эффективности равен 96% и 97% при выборе корпуса

текстов 10, 20 авторов с 20, 40 текстами.

4.3.7. Тестирование.

Итак, настройка (обучение) γ-классификатора на данных корпусах текстов ***C*** прошла успешно. Для тестирования классификатора выбрано случайным образом 3 текста:

– *А. Рӯдакӣ (АР):* «Қасоид» (Қ, 5060 *слов*);

– *А. Фирдавси (АФ):* «Подшоҳии Лӯҳросп» (Л, 9938 *слов*);

– *М. Турсунзода (МТ):* «Ҳасани аробакаш» (ҲА, 8515 *слов*).

Отметим, что сведения относительно выбранных произведений описаны по той же схеме, что и для элементов коллекции ***C***.

Для трех произведений, предназначенных для тестирования, построены цифровые портреты (4.15) и затем по формулам (4.16), (4.17), (4.18) для каждого из них вычислены расстояния до 70 объектов коллекции ***C***. Соответствующие значения записаны в ячейках таблицы 4.6, расположенных на пересечениях строк и столбцов. Там же серым цветом выделены пары ячеек с минимальными расстояниями между тестируемыми и содержащимися в коллекции произведениями.

Таблица 4.6. – Расстояния между текстами коллекции ***C*** и 3 случайно выбранными тестируемыми произведениями

Тексты		**АР**	**АФ**	**МТ**
		Қ	Л	ҲА
АБ	Ғ	0.1307	0.2089	0.0753
	Р	0.0927	0.1709	0.0552
АД	МҒ1	0.0614	0.0661	0.0959
	МҒ2	0.0553	0.1334	0.0759
АР	АП	0.0449	0.0635	0.1020
	Р	**0.0053**	0.1048	0.0753
АС	Д1	0.1452	0.2233	0.1089
	Д2	0.1291	0.2073	0.0853
	Д3	0.1012	0.1760	0.0717
	Д4	0.1242	0.2023	0.0729
	Д5	0.0694	0.1476	0.0727
	Д6	0.1484	0.2265	0.0929
	Д7	0.1090	0.1872	0.0615
	Д8	0.1203	0.1985	0.0649
АФ	Б&М	0.0501	0.0582	0.0960
	И&Р	0.0628	0.0492	0.1068
	КВ	0.0818	0.0395	0.1071
	КК	0.0820	0.0306	0.1236
	М	0.0713	0.0440	0.1200
	Р&С	0.0628	0.0595	0.0862
	Р&Х	0.0587	0.0513	0.1142
	Ф	0.0876	**0.0263**	0.1431
АҶ	Л&М	0.0539	0.1042	0.0726
	Ю&З	0.0413	0.0988	0.0790
БЗ	ДМ1	0.0912	0.1694	0.0496

Тексты		АР	АФ	МТ
		Қ	Л	ҲА
	ДМ2	0.1110	0.1892	0.0639
ИФ	Ғ	0.0678	0.1460	0.0949
	МГ	0.0803	0.1584	0.0832
КХ	Ғ1	0.0802	0.0927	0.1240
	Ғ2	0.1306	0.1039	0.1728
	Ғ3	0.0638	0.1420	0.0712
	Ғ4	0.1003	0.1785	0.0618
	Ғ5	0.0653	0.1434	0.0779
	Ғ6	0.1122	0.1904	0.0706
	Ғ7	0.0369	0.1151	0.0979
	Ғ8	0.0562	0.1344	0.0983
ЛШ	К1	0.0518	0.1287	0.0688
	К2	0.0605	0.1202	0.0934
	К3	0.0546	0.1327	0.0876
	К4	0.0643	0.1517	0.0747
	К5	0.0723	0.1411	0.0710
	К6	0.1060	0.1842	0.0686
	К7	0.0636	0.1435	0.0751
	К8	0.0779	0.1561	0.0517
МҚ	МШ1	0.0699	0.1452	0.0771
	МШ2	0.0593	0.1173	0.0666
ММ	ҚТ1	0.0410	0.1045	0.0560
	ҚТ2	0.0394	0.1147	0.0655
МТ	МО1	0.0563	0.0983	0.0650
	МО2	0.0665	0.1273	**0.0414**
НГ	Л&М	0.0519	0.0995	0.0967
	Х&Ш	0.0626	0.0534	0.1017
НҚ	МТ1	0.0735	0.1773	0.0551
	МТ2	0.0838	0.1863	0.0641
НХ	МХ	0.0420	0.0888	0.0833
	С	0.0374	0.0945	0.0748
СШ	Ғ1	0.0906	0.1688	0.0756
	Ғ2	0.1047	0.1654	0.0594
УХ	Р1	0.0553	0.1116	0.0974
	Р2	0.1028	0.1810	0.0699
ҲШ	Ғ1	0.0654	0.1436	0.1054
	Ғ2	0.1160	0.1942	0.0623
ЧР	ММ1	0.0689	0.0609	0.0979
	ММ2	0.0631	0.0604	0.1098
	ММ3	0.0550	0.0837	0.0799
	ММ4	0.0593	0.0802	0.1044
	ММ5	0.0331	0.0918	0.0806
	ММ6	0.0508	0.0698	0.0924
	ММ7	0.0614	0.0642	0.1032
	ММ8	0.0443	0.0757	0.0897

В первых трех столбцах ближайшими соседями текстов **АР**_Қ, **АФ**_Л и **МТ**_ҲА являются соответственно **АР**_Р, **АФ**_Ф и **МТ**_МО2 на расстояниях соответственно 0.0053, 0.0263 и 0.0414 (в таблице отмечены серым цветом). Интересно то, что эти расстояния меньше $\gamma^{опт}$, см. (4.21). Полученный результат

показывает, что по методу ближайшего соседа [7-9, 248, 249] три случайно выбранных произведения оказались как раз однородные с ними по автору пары текстов исходной коллекции.

Заключение. В настоящем параграфе была исследована применимость γ-классификатора не для модельной коллекции, а для корпуса, состоящего из 70 текстов 20 авторов. Анализ результатов показал, что методика на основе γ-классификатора способна достигать точности 98%, что является наилучшим результатом на сегодняшний день.

Таким образом, математическая триада в составе ЦП текстов, представляемых распределениями частотности 26 кириллических букв, формул (4.15)-(4.17) для вычисления расстояний между текстами и алгоритмом для выявления однородных текстов оказалась подходящей для эффективного решения поставленной задачи.

В свою очередь, метод ближайшего соседа показал возможность безошибочного распределения дополнительных трех произведений по авторам.

Автор выражает уверенность в том, что еще увеличение объема исходной коллекции текстов не станет препятствием для успешного применения γ-классификатора не только для распознавания авторов, но также и для самых разнообразных однородностей текстовых документов.

Результаты §§ 4.1. – 4.3. опубликованы в [62-А, 65-А, 67-А].

§ 4.4. Структура однородностей поэм произведения А. Фирдоуси «Шахнаме»

Введению и 63 поэмам произведения А. Фирдоуси «Шахнаме» сопоставляются цифровые портреты на основе распределений в них частностей букв кириллического алфавита таджикского языка. Воспользуемся агломеративным иерархическим алгоритмом классификации. В качестве расстояния между объектами примем метод γ-классификатора дискретных случайных чисел. В этом параграфе используется только два компонента. Полученные данные помещаем в таблицу (матрицу расстояний). С помощью метода ближайшего соседа по матрице расстояний осуществляется иерархическая кластеризация составных частей произведения.

Наши первые исследования творчества великого поэта А. Фирдоуси, представленного в произведении «Шахнаме» на таджикско-персидском языке в кириллической графике [277], были предприняты в публикациях [278, 40-А]. В [283] на основе обобщения формулы «золотого сечения», предложенного в [278], изучался вопрос о положении точки кульминации в трёх поэмах – о Нузаре, Рустаме и Сухробе и Сиёвуше. В [284] на примере одиннадцати поэм, оцифрованных с помощью пяти натуральных единиц измерения текста,

установлена статистическая неразличимость оригинала и его перевода [279] на русский язык.

В данном параграфе мы вновь обращаемся к творчеству А. Фирдоуси, но в отличие от работ [40-А, 283], в которых изучались зависимости между количествами словоупотреблений и словоформ, рассматриваем не отдельные поэмы, а произведение «Шахнаме» в полном объёме, и на основе информации о распределении частотностей буквенных униграмм займемся установлением взаимосвязей между различными частями произведения.

4.4.1. Исходный материал, использованный нами для исследования, состоял из «Вступления» и 63 поэм А. Фирдоуси «Шахнаме». Список составных частей в порядке, в котором они встречаются в «Шахнаме», представляется своими названиями, сопровождаемыми (в скобках) их сокращениями и размерами в словах:

Оғози китоб (ОК, 2680 *слов*); *Оғози достон* (ОД, 945 *слов*); *Ҳушанг* (Ҳ, 512 *слова*); *Таҳмурас* (Т, 537 *слов*); *Ҷамшед* (Д, 2363 *слова*); *Заҳҳок* (З, 5876 *слов*); *Фаридун* (Ф, 12347 *слов*); *Манучеҳр* (М, 22169 *слов*); *Нӯзар* (Н, 6612 *слова*); *Зави Таҳмосп* (ЗТ, 523 *слова*); *Гаршосп* (Г, 3006 *слов*); *Қайқубод* (ҚД, 2667 *слов*); *Кайковус* (К, 10865 *слов*); *Кори Кайковус ба шаҳри Барбаристон ва дигар достонҳо* (БД, 8434 *слова*); *Достони Рустам ва Сӯҳроб* (Р&С, 16388 *слов*); *Достони Сиёвуш* (С, 30541 *слово*); *Шикояти Фирдавсӣ аз пирии худ* (ШФ, 15862 *слова*); *Кайхусрав* (КВ, 18782 *слова*); *Достони Комуси Кашонӣ* (КК, 17454 *слова*); *Достони Рустам бо Хоқони Чин* (Р&Х, 16722 *слова*); *Достони Ҷанги Рустам бо Аквондев* (Р&А, 2604 *слова*); *Достони Бежан бо Манижа* (Б&М, 14884 *слова*); *Достони Дувоздаҳ Рух* (ДР, 27871 *слово*); *Подшоҳии Кайхусрав* (ПКВ, 35991 *слово*); *Подшоҳии Лӯҳросп* (Л, 9952 *слова*); *Подшоҳии Гуштосп* (ПГ, 16090 *слов*); *Ҳафт хони Исфандиёр* (ПИ, 9483 *слова*); *Достони Разми Исфандиёр бо Рустам* (И&Р, 18716 *слов*); *Достони Рустам ва Шағод* (Р&Ш, 3753 *слова*); *Подшоҳии Баҳмани Исфандиёр* (БИ, 1697 *слов*); *Подшоҳии Ҳумой* (ҲЙ, 3516 *слов*); *Подшоҳии Дороб* (ПД, 1460 *слов*); *Подшоҳии Доро писари Дороб* (ДД, 4950 *слов*); *Подшоҳии Искандар* (И, 21469 *слов*); *Подшоҳии Ашкониён* (А, 7976 *слов*); *Подшоҳии Сосониён (ПС,* 7043 *слова)*; *Подшоҳии Шопури Ардашер* (ША, 944 *слова*); *Подшоҳии Урмузди Шопур* (УШ, 978 *слов*); *Подшоҳии Баҳроми Урмузд* (БУ, 430 *слов*); *Подшоҳии Баҳроми Баҳром* (Б, 317 *слов*); *Подшоҳии Баҳроми Баҳромиён* (ББ, 139 *слов*); *Подшоҳии Нарсии Баҳром* (НБ, 281 *слово*); *Подшоҳии Урмузди Нарсӣ* (УН, 263 *слова*); *Подшоҳии Шопури Зулактоф* (ШЗ, 7099 *слов*); *Подшоҳии Ардашери Некӯкор* (АН, 181 *слово*); *Подшоҳии Шопур ибни Шопур* (ШШ, 352 *слова*); *Подшоҳии Баҳром писари Шопур* (БШ, 340 *слов*); *Подшоҳии Яздгирди Базагар* (ЯБ, 7524 *слова*); *Подшоҳии Баҳроми Гӯр* (БГ, 28726 *слов*); *Подшоҳии Яздгирд писари Баҳроми Гӯр* (Я, 277 *слов*); *Подшоҳии Ҳурмуз писари Яздгирд* (ҲЯ,

208 *слов*); *Подшоҳии Пирӯз писари Яздгирд* (ПЯ, 1485 *слов*); *Подшоҳии Балош писари Пирӯз* (БП, 2062 *слова*); *Подшоҳии Қуболи Пирӯз* (К&П, 4474 *слова*); *Подшоҳии Кисрои Нӯшинравон* (КН, 49721 *слово*); *Подшоҳии Ҳурмузд* (ҲД, 21002 *слова*); *Подшоҳии Хусрави Парвиз* (ХП, 45443 *слова*); *Подшоҳии Қуболи Парвиз* (ҚП, 6610 *слов*); *Подшоҳии Ардашери Ширӯй* (АШ, 629 *слов*); *Подшоҳии Фароин Гуроз* (ФГ, 711 *слово*); *Подшоҳии Пӯрондухт* (П, 253 *слова*); *Подшоҳии Озармдухт* (О, 110 *слов*); *Подшоҳии Фаррухзод* (ПФ, 309 *слов*); *Подшоҳии Яздгирд* (ПЯД, 9474 *слова*).

4.4.2. Обработка данных происходила в 3 этапа.

Этап 1. Для «Вступления» и 63 поэм построены согласно в § 1.3 цифровые портреты, характеризующие распределение частотности буквенных униграмм каждой части произведения.

Цифровые портреты представлены в табличном виде:

$$\begin{array}{lccccc} \bar{N} & : & 1 & 2 & \ldots & 35 \\ P & : & p_1 & p_2 & \ldots & p_{35}, \end{array}$$

в котором первая строка – порядковые номера 35 алфавитных букв (униграмм) таджикского языка; вторая строка – относительные частоты p_i букв ($i = \overline{1,\ 35}$), причём

$$\sum_{i=1}^{35} p_i = 1.$$

Этап 2. Вычисления согласно в § 1.3 расстояний $\rho(v_1, v_2)$ между всеми ЦП 64 составных частей произведения «Шахнаме» по формуле

$$\rho(v_1, v_2) = \sqrt{35/2}\, max_s \left|\sum_{i=1}^{s}\left(p_i^{(1)} - p_i^{(2)}\right)\right|,$$

в которой $p_i^{(1)}$ и $\mathrm{p_i^{(2)}}$– частотности буквы i (i = 1, … , 35) в поэмах v_1 и v_2 и s = 1, … , 35.

Этап 3. Определение на основе матрицы парных расстояний $\left[\rho\left(v_i, v_j\right)\right],\ i,j = \overline{1,64},$ методом ближайшего соседа, см., например, [248], структуры взаимных расположений составных частей. Итоговый результат представлен на рис. 4.1 в виде дендограммы, то есть «*дерево*», «ствол», «ветви» и «листья» которой строятся на основе матрицы расстояний $\left[\rho\left(v_i, v_j\right)\right]$. Построение выполняется агломеративным способом от «листьев к стволу» путём последовательного объединения каждой составной части произведения А. Фирдоуси, прежде всего, с ближайшим по расстоянию «соседом» в единый кластер, а затем уже совместно в более крупные подмножества.

На рис. 4.1 по оси абсцисс в сокращенных обозначениях размещены названия поэм по принципу ближайших друг к другу соседей, по оси ординат представлена

шкала взаимных расстояний между поэмами.

Из 64 составных единиц произведения «Шахнаме» особо «однородными» выглядят поэты «Подшоҳии Яздгирд» (ПЯД) и «Подшоҳии Кайхусрав» (ПКВ), между которыми расстояние $\rho((\text{ПЯД}),(\text{ПКВ})) = 0.0128$ оказалось минимальным в сравнении со всеми другими. Вместе с тем, на самом большом удалении расположились «Подшоҳии Ардашери Некӯкор» (АН) и «Подшоҳии Шопур ибни Шопур» (ШШ), который $\rho((\text{АН}),(\text{ШШ})) = 0.4021$. Возможная причина столь большого расстояния между ними состоит в том, что размеры этих поэм довольно незначительные, 181 *слово* в (АН) и 352 *слова* в (ШШ). На этом фоне в ПЯД содержится 9474 *слова*, а в (ПКВ) – 35991 *слово*.

Для среднего расстояния имеем $\rho = 0.0851$.

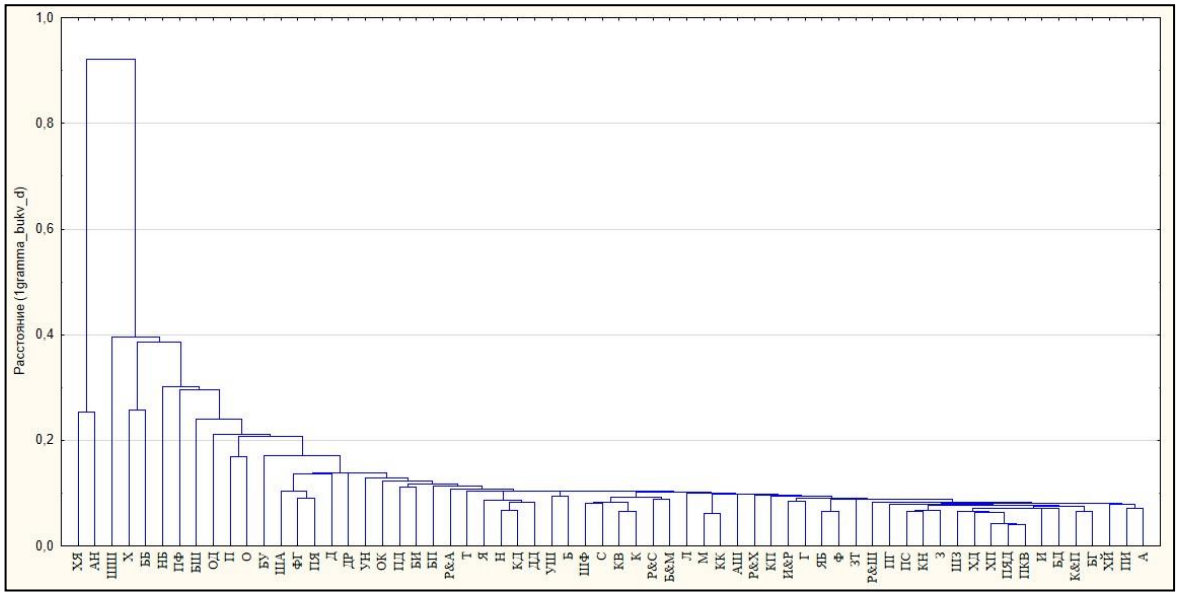


Рисунок 4.1. – Результаты иерархической классификации поэм в виде дендрограммы

Представляет интерес получить мнение квалифицированных литературоведов относительно иерархической классификации поэм произведения «Шахнаме», представленного на рис. 4.1.

Результаты данного параграфа опубликованы в [25-А].

§ 4.5. Оценка эффективности тестирование γ-классификатора для определения автора искусственно сгенерированных поэм «Шахнаме» А. Фирдоуси

Мы представляем исследование по обучению рекуррентных нейронных сетей поэмами «Шахнаме» А. Фирдоуси и генерацию новых поэм. Модель изучила долгосрочные зависимости и синтаксические характеристики корпуса. Эффективность классификации новых поэм в приложении TAJIK_TEXT_AUTHOR для определения автора текста устанавливается.

Как отмечается в §§ 2.1-2.5, при решении задач определения авторства текстов, использование $\gamma-$классификатора из §§ 1.3-1.4 подтвердило свою эффективность. До этого все исследования об эффективности применения и тестирования $\gamma-$ классификатора проводились с помощью коллекции текстов $\mathbb{T}_2$ (см. задача 2 [276]). $\mathbb{T}_2$ – это часть коллекции текста $\mathbb{T} = \{\mathrm{T_i}\}$, предназначенная для тестирования классификатора, для которого автор текста известен из списка $\mathbb{A} = \{\mathrm{A_i}\}$. Практическое применение и эффективность классификатора также можно оценить в условиях, при котором задан текст $\widetilde{\mathrm{T}}$, похожий на некоторый текст из коллекции $\mathbb{T} = \{\mathrm{T_j}\}$, где $\widetilde{\mathrm{T}} \notin \mathbb{T}$ и для которого априори автор $\widetilde{\mathrm{A}}$ неизвестен и $\widetilde{\mathrm{A}} \notin \mathbb{A}$.

Постановка задачи

1) Сгенерировать текст $\widetilde{\mathrm{T}}$ с помощью авто-регрессионной генерации последовательности символов (см. рис. 5 семплинг [280]), базирующийся на рекуррентных нейронных сетях, и обученные поэмами «Шахнаме» А. Фирдоуси [277].

2) Полученный текст $\widetilde{\mathrm{T}}$ использовать в качестве $\mathbb{T}_2$ для проверки эффективности $\gamma-$ классификатора.

1. Генерация текста $\widetilde{\boldsymbol{T}}$. Генерация текста является одним из приложений задачи языкового моделирования. Традиционные методы создания языковых моделей основаны посредством подсчета $N-gram$. Основной проблемой $N-gram$ моделей является *разреженность* (нехватка) данных [1], и это проблема была решена с помощью использования *рекуррентных нейронных сетей* (RNN), в частности рекуррентными нейронными сетями типа «*долгая краткосрочная память»* (*Long Short-Term Memory, LSTM*) [10], так как они хорошо подходят для моделирования последовательных данных [11]. Генерация текста состоит из 3 этапов:

1) Составление ЦП обучаемого корпуса.

2) Обучение рекуррентной нейронной сети в пакетном режиме.

3) Авто-регрессионная генерация последовательности символов.

1.1. Составление ЦП поэмы «Шахнаме» А. Фирдоуси. Пусть для некоторого естественного языка L с символами $S = \{s_i\} = A \cup P$, где $A-$ алфавит языка, а $P-$ знаки препинания и другие вспомогательные знаки, имеется корпус $C = \{c_n\}$, состоящий из n набора символов $c_k \in S,\ k = \overline{1,n}$.

В качестве корпуса C берем поэму «Шахнаме» А. Фирдоуси [277], длина которого равна $|C| = 3281624$, а количество различных символов в корпусе $|S| = 49$. Используя *унитарное кодирование* [281] каждый символ c_k закодируем 49-мерным вектором x_k, в котором прямой унитарный код $x_k[i] = 1$, где $i-$ позиция c_k в S.

Вектору $X = \{x_n\}$ сопоставим вектор $Y = \{y_n\}$ таким образом, что $y_j = x_{j+1}$

для $j = 1,2,3, \dots, n-1$ и $y_n = \vec{0}$. Следовательно, $\{X, Y\}$ является цифровым портретом поэмы «Шахнаме» А. Фирдоуси.

1.2. Обучение рекуррентной нейронной сети в пакетном режиме. Простая модель RNN, использованная в [280], хорошо подходит для коротких последовательных данных, но *не справляется с обучением хранения информации в течение длительных интервалов времени* [10, 2, 12]. Для экспериментального моделирования поэм произведения «Шахнаме» А. Фирдоуси будем использовать рекуррентную нейронную сеть типа LSTM [10].

LSTM является особым видом RNN и состоит из набора рекуррентных подключенных подсетей, известные как блоки памяти (гейт, вентиль). Эти блоки способны изучать, какие данные в последовательности важно сохранить или забыть. Таким образом, сеть сохраняет важную информацию по длинной цепочке для будущего прогнозирования.

В процессе обучения была использована сеть LSTM со следующими параметрами:

- *количество входных параметров:* 128
- *глубина слоя скрытых состояний:* 4
- *количество параметров в скрытом состоянии ячейки:* 1024
- *размер мини-партии* [13]*:* 128
- *скорость обучения:* 10^{-3}
- *количество итерации обучения модели:* 60.

1.3. Авто-регрессионная генерации последовательности символов. После окончания обучения модели LSTM с использованием метода семплирования [280] было сгенерировано 2 текста ($\tilde{T}_1, \tilde{T}_2$) длиной 1001 и 5002 слов. Полный текст версии с 5002 словами искусственно сгенерированной поэмы доступен в [282], и отрывок поэмы приведена ниже:

…

Ҳаме гуфт, к-«аз ранҷҳо беш ном,
Ба ман дил бипӯшид бояд ба дард.
Бубинам, ки бе дур карда сипоҳ,
Зи ганҷу сиёвахш бо бегазанд.
Набинад кас андар ҷаҳон низ ҷой,
Бимурд андар ин кор гуфтори хеш,
Ки дар бозгаштан бувад дастгоҳ,
Ба гетӣ магӯям, ки бо бежаност».

…

Для генерации нового текста была задана первоначальная последовательность символов, как «Подшоҳ» и «Ба». Из обученного текста и заданной начальной последовательности модель рассчитывает вероятность

появления следующего символа. Из искусственно сгенерированных символов (поэм) можно наблюдать, что модель научилась:

- генерировать правильные слова;
- генерировать новую строку и при этом длина строки поэмы в среднеарифметическом на 0.24 слов длинее, чем поэмы «Шахнаме»;
- открывать и закрывать кавычки;
- заканчивать строку с запятой в нечетных и с точкой в четных строках;
- ставить запятую перед словом «*ки*»;
- генерировать часто употребленные слова и последовательность символов как: «бад-ӯ», «бад-он», «к-аз», «к-эй», «к-ӯ», «гуфт:», «х(в)ар»;
- использование имен персонажей произведения как: «Афросиёб», «Хусрав», «Бежан», «Сиёвуш», «Равшан», «Рустам».

2. Распознавание автора текста в программном комплексе TAJIK_TEXT_AUTHOR (TTA) [283, 38-А]**.** Искусственно сгенерированные тексты $\tilde{T}_1$ и $\tilde{T}_2$, протестированные в программном комплексе **«TTA»**, привели к следующим результатам (таблица 4.7):

Таблица 4.7. – Эффективность искусственно сгенерированной поэмы

Искусственный текст	*N*-грамма	Эффек-тивность	А. Фирдоуси		Дж. Руми		…
			P&C	Б&M	MM1	MM2	
$\tilde{T}_1$ (1001 *слово*)	1gr. с пр.	93	0.0656	0.0596	0.1592	0.1524	
	2gr. с пр.	93	0.4204	0.3627	1.0171	0.9481	
	3gr. с пр.	96	2.5941	2.2471	6.1823	5.7627	
$\tilde{T}_2$ (5002 *слова*)	1gr. с пр.	100	0.0639	0.0499	0.1537	0.1472	
	2gr. с пр.	100	0.4048	0.3344	0.9528	0.9196	
	3gr. с пр.	100	2.4507	2.1481	6.0758	5.8107	

Продолжение Таблицы 4.7.

…	А. Суруш		С. Айни		С. Турсун		И. Фарзона	
	Д1	Д2	АД	О	Н	ПКР	101Г	МГМ
	0.1013	0.1072	0.1813	0.1509	0.1602	0.1668	0.1403	0.1287
	0.8961	0.9893	1.0852	1.0023	0.9609	1.0006	0.8956	0.8467
	5.3783	5.9337	6.5956	6.1436	5.7937	6.0593	5.5078	5.1866
	0.0961	0.1045	0.1475	0.1272	0.1514	0.1581	0.1351	0.1235
	0.8558	0.9489	0.8964	0.8686	0.9084	0.9481	0.8821	0.8195
	5.1459	5.6936	5.8311	5.5742	5.4814	5.7471	5.4464	5.0074

Программный комплекс использует $\gamma-$классификатор § 1.4 и оценивает эффективность использования униграмм, биграмм и триграмм при идентификации автора текста [6-А-8-А]. Объем искусственного текста удовлетворяет минимальному требованию, необходимому для распознавания его автора [10-А].

Результаты тестирования показали, что расстояния искусственно сгенерированных текстов $\tilde{T}_1$ и $\tilde{T}_2$ от произведений А. Фирдоуси (P&C, Б&M) при

использовании униграмм являются очень близкими (< 0.07). Также эффективность классификации автора текста составила 93-100%. Это свидетельствует о том, что качество искусственно сгенерированных текстов $\tilde{T}$ таково, что обученная рекуррентная нейронная сеть LSTM смогла научиться некоторым синтаксическим и стилистическим характеристикам поэмы «Шахнаме», и программный комплекс TTA с большой вероятностью смог предсказать А. Фирдоуси как автор этих текстов.

Результаты данного параграфа опубликованы в [22-А].

Заключение. Мы обучили рекуррентную нейронную сеть LSTM поэмами «Шахнаме» А. Фирдоуси и сгенерировали искусственные поэмы. Модель научилась характеристикам этого корпуса и при тестировании искусственных поэм программным комплексом TTA эффективность классификации автора текста составила 93-100%. При наличии эталона корпуса таджикского языка с большой вероятностью можно сказать, что рекуррентные нейронные сети LSTM позволяют создать языковую модель таджикского языка и разрешить некоторые задачи обработки естественного языка.

ГЛАВА 5. ИССЛЕДОВАНИЕ ВЛИЯНИЯ ПОРЯДКА ЦП ТЕКСТА НА РАСПОЗНАВАНИЕ ОДНОРОДНОСТИ ПРОИЗВЕДЕНИЯ

§ 5.1. О влиянии ЦП текста на определение автора произведения

На примере модельной коллекции, количественные описания произведений которой основываются на различных вариантах упорядочения буквенных N-грамм ($N = 1, 2, 3$) с пробелами, выявляются особенности применения γ-классификатора при распознавании автора текста.

Согласно Рудману [3] современный исследователь может использовать около тысячи разнообразных признаков текста и каждому сопоставлять свой определенный ЦП, формирующий количественный образ текста. В дальнейшем нас интересует специфические широко используемые в γ-классификаторах [259, 260] портреты на основе распределения частотностей элементов текста.

Поясним некоторые понятия, используемые в § 1.3.

Определение 5.1.1. ***Алфавит*** *– упорядоченное множество элементов текста.*

Примерами элементов текста могут служить буквы алфавита естественного языка, буквенные N-граммы и слоги, упорядоченные по алфавиту, длины слов и предложений, упорядоченные по возрастанию или убыванию длин, и т.д.

Определение 5.1.2. ***ЦП текста*** *назовём распределение частотности элементов алфавита.*

Примерами ЦП текста являются распределения частотностей символьных, буквенных и словоформных N-грамм, длин слов и предложений и т.д.

В настоящем параграфе на примерах модельных коллекций текстов устанавливаются особенности ЦП и γ-классификатора в зависимости от упорядочения алфавитных элементов.

5.1.1. Состав модельной коллекции текстов, заимствованной из § 2.1, представлен следующими произведениями

классической поэзии:

– А. Рӯдакӣ «Адабиёти пароканда» и «Қасоид»;

– А. Фирдавсӣ «Достони Рустам ва Сӯҳроб» и «Достони Бежан бо Манижа»;

– С. Шерозӣ «Ғазалиёт, қисми 1» и «Ғазалиёт, қисми 2»;

– Ҳ. Шерозӣ «Ғазалиёт, қисми 1» и «Ғазалиёт, қисми 2»;

– Ҷ. Румӣ «Маснавии Маънавӣ, Дафтари Аввал» и «Маснавии Маънавӣ, Дафтари Дуввум»;

современной поэзии:

– А. Суруш «Дафтари 1» и «Дафтари 2»;

– А. Шукӯҳӣ «Баргҳои тиллоӣ» и «Шохаи райҳон»;

– Г. Сафиева «Офтоб дар соя» и «Шӯъла дар санг»;

– И. Фарзона «101-Ғазал» и «Мӯҳри гули мино»;

– М. Турсунзода «Қиссаи Ҳиндустон» и «Ҳасани аробакаш» ;

современной прозы:

– А. Зоҳир «Бозгашт» и «Завол»;

– Г. Муҳаммадиева «Бӯи модар» и «Сафинаи муҳаббат»;

– М. Шакурӣ «Садри Бухоро» и «Хуросон аст ин ҷо»;

– С. Турсун «Нисфирӯзӣ» и «Повести Камони Рустам»;

– С. Айнӣ «Дохунда» и «Марги судхӯр».

Таким образом, модельная коллекция составлена из 3-х частей: классической и современной поэзий и современной прозы. В каждой части по 5 авторов, от каждого автора по 2 произведения.

5.1.2. Примеры текстовых элементов и их алфавитов. При изложении данного вопроса ограничимся рассмотрением простейших случаев, когда в качестве элементов текста выбираются N-граммы ($N = 1, 2, 3$) с пробелами.

Для униграмм ($N = 1$) естественных языков существующие алфавиты уже являются отсортированными в определенном порядке конечными множествами букв (также и с учётом пробела). Лексикографический порядок, аналогичный алфавитной сортировке, алфавитизирует также N-граммы ($N \geq 2$) и более сложные буквенно-символьные комбинации. Однако в дополнение к сказанному отметим, что такие комбинации, упорядоченные каким-либо другим способом, будут также называться алфавитными элементами текста. Как будет отмечено в п. 5.1.4, расстояния между ЦП текстов зависят от порядка элементов алфавита и потому не ясно, какому из допустимых алфавитов следует отдать предпочтение.

5.1.3. ЦПТ и расстояния между ними. После выбора фиксированного алфавита ЦП текста T удобно представлять в табличной форме:

$$\begin{array}{lcccccc} \overline{N} & : & 1 & 2 & . & . & . \quad m \\ P & : & p_1 & p_2 & . & . & . \quad p_m, \end{array} \tag{5.1}$$

в которой m – число элементов алфавита, строка $\overline{N}$ указывает номера упорядоченных элементов алфавита, а строка P – их относительные частоты встречаемости в T, причём

$$\sum_{k=1}^{m} p_k = 1.$$

ЦП можно задавать также дискретной функцией

$$F(s) = \sum_{k=1}^{s} p_k \qquad (s = 1, \cdots, m),$$

характеризующей распределение в тексте частот встречаемости элементов

алфавита.

Определение 5.1.3. *Расстоянием между двумя текстам называется расстояние между их ЦП, отнесенными к единому алфавиту.*

Пусть T_1, T_2 – произвольная пара текстов из коллекции $\mathbb{T}$ и

$$F^{(\alpha)}(s) = \sum_{k=1}^{s} p_k^{(\alpha)} - \quad (5.2)$$

соответствующие им дискретные функции, $\alpha = 1, 2$ и $s = 1, \cdots, m$.

Определение 5.1.4. *Расстоянием между текстами T_1 и T_2 называется положительное число $\rho(T_1, T_2)$, определяемое по формуле*

$$\rho(T_1, T_2) = \sqrt{m/2}\ max_s \left| F^{(1)}(s) - F^{(2)}(s) \right|, \quad (5.3)$$

то есть расстояние между двумя текстами вычисляется как максимальное расстояние по оси ординат между их дискретными функциями $F^{(1)}(s)$ и $F^{(2)}(s)$, помноженное на весовой коэффициент $\sqrt{m/2}$.

Замечание. Условие $\rho(T_1, T_2) = 0$ означает тождество ЦП текстов, то есть ЦПT_1 = ЦПT_2, но не $T_1 = T_2$, то есть идентичность текстов.

5.1.4. Обработка данных коллекционного материала, представленного в п. 5.1.1, состояла из 3 этапов.

Этап 1. Использование для всех произведений трёх частей коллекции трёх типов текстовых элементов:

- униграмм с учетом пробела (в таджикском языке 35 букв алфавита, потому общее число униграмм – 36):
- биграмм с учетом пробела (общее число таковых – $36^2 = 1296$):
- триграмм с учетом пробела (число таковых – $36^3 = 46656$).

Множества N-грамм ($N = 1, 2, 3$) в зависимости от упорядочения своих элементов рассматриваются в 4-х вариантах:

1) элементы располагаются в алфавитном порядке с пробелом в качестве последнего элемента алфавита (обозначается как ABC)[5];

2) элементы располагаются в порядке, обратном алфавитному с пробелом в качестве первого элемента алфавита (обозначается как CBA)[6];

3) элементы располагаются в порядке убывания их частотности в тексте (обозначается символом « ↘»);

4) элементы располагаются в порядке возрастания их частотности в тексте (обозначается символом « ↗ »).

Этап 2. Для каждого из 4-х вариантов упорядочения -грамм ($N = 1, 2, 3$) путём автоматической обработки формируются в табличном виде (5.1) цифровые

[5] Для биграмм и триграмм – с двумя и тремя пробелами в конце.
[6] Для биграмм и триграмм – с двумя и тремя пробелами в начале.

портреты всех произведений коллекции и затем по формулам (5.2) и (5.3) вычисляются расстояния между парами текстов на таджикском языке по отдельности из классической поэзии, современной поэзии и современной прозы. Из-за большого количества расстояний (таковых 135 = 3 х 45) мы не приводим итоговых результатов, однако обращаем внимание на тот факт, что расстояния, вычисляемые между любыми двумя текстами для различных вариантов расположения алфавитных элементов, оказываются в общем случае различными. В этом можно убедиться на простых примерах.

Этап 3. Настройка γ-классификатора – алгоритма, зависящего от одного вещественного параметра γ и устанавливающего в пределах модельной коллекции соответствие между текстами и их авторами. Существо настройки заключается в определении такого значения γ, при котором произведения одного автора «γ-однородны», а разных авторов – «γ-неоднородны». Однородность всех текстов одного автора в рамках математической модели означает справедливость неравенства

$$\rho\,(T_1\,, T_2) \leq\ \gamma, \tag{5.4}$$

а неоднородность любых двух текстов разных авторов – справедливость неравенства

$$\rho\,(T_1\,, T_2) > \gamma. \tag{5.5}$$

Ошибки в настройке γ-классификатора выявляется в случае, когда для каких-то пар текстов одного и того же автора вместо неравенства (5.4) имеет место неравенство (5.5), а также в случае, когда какие-то два произведения двух различных авторов удовлетворяют неравенство (5.4) вместо того, чтобы выполнялось неравенство (5.5).

Суммарное количество $\tau = \tau(\gamma)$ допущенных ошибок одновременно в двух случаях позволяет подсчитать величину π эффективности γ-классификатора при распознавании авторов текста по формуле

$$\pi = 1 -\ \tau(\gamma)/L\,, \tag{5.6}$$

где $L = 45$ – число взаимных расстояний между всеми парами произведений из классической и современной поэзий, а также из современной прозы. Детальное описание алгоритма для нахождения оптимального значения γ, при котором π (5.6) принимает максимальное значение, содержится в §§ 1.3-1.4.

Итоги применения трёх этапов автоматической обработки модельной коллекции текстов показаны в таблицах 5.1-5.3, соответственно для 3-х частей коллекции.

Таблица 5.1. – Значения π и γ для произведений классической поэзии

Элементы текста	Число элементов алфавита	Порядок элементов алфавита	π-эффективность распознавания автора	Оптимальный γ-полуинтервал
уни-граммы	36	А В С	0.98	[0.0354; 0.0447)
		С В А	0.98	[0.0354; 0.0447)
		по ↘	0.98	[0.0337; 0.0342)
		по ↗	0.98	[0.0337; 0.0342)
би-граммы	1296	А В С	0.98	[0.2987; 0.3551)
		С В А	0.98	[0.2987; 0.3551)
		по ↘	0.96	[0.2065; 0.2212)
		по ↗	0.96	[0.2065; 0.2212)
три-граммы	46656	А В С	1.00	[2.1630; 2.1648)
		С В А	1.00	[2.1630; 2.1648)
		по ↘	0.96	[1.2426; 1.4051)
		по ↗	0.96	[1.2426; 1.4051)

В этой таблице также, как и в двух последующих, в столбце 3 для описания порядка следования алфавитных элементов приняты обозначения, введенные в п. 5.1.4, этап 1.

Таблица 5.2. – Значения π и γ для произведений современной поэзии

Элементы текста	Число элементов алфавита	Порядок элементов алфавита	π-эффективность распознавания автора	Оптимальный γ-полуинтервал
уни-граммы	36	А В С	0.98	[0.0268; 0.0423)
		С В А	0.98	[0.0268; 0.0423)
		по ↘	0.98	[0.0384; 0.0415)
		по ↗	0.98	[0.0384; 0.0415)
би-граммы	1296	А В С	0.98	[0.2318; 0.2816)
		С В А	0.98	[0.2318; 0.2816)
		по ↘	0.98	[0.2484; 0.2745)
		по ↗	0.98	[0.2484; 0.2745)
три-граммы	46656	А В С	0.98	[1.3885; 1.7054)
		С В А	0.98	[1.3885; 1.7054)
		по ↘	0.98	[1.5556; 1.6453)
		по ↗	0.98	[1.5556; 1.6453)

Таблица 5.3. – Значения π и γ для произведений современной прозы

Элементы текста	Число элементов алфавита	Порядок элементов алфавита	π-эффективность распознавания автора	Оптимальный γ-полуинтервал
уни-граммы	36	А В С	0.96	[0.0285; 0.0336)
		С В А	0.96	[0.0285; 0.0336)
		по ↘	0.91	[0.0165; 0.0236)
		по ↗	0.91	[0.0165; 0.0236)

Элементы текста	Число элементов алфавита	Порядок элементов алфавита	π-эффективность распознавания автора	Оптимальный γ-полуинтервал
би-граммы	1296	A B C	0.93	[0.2216; 0.2272)
		C B A	0.93	[0.2216; 0.2272)
		по ↘	0.91	[0.2386; 0.2568)
		по ↗	0.91	[0.2386; 0.2568)
три-граммы	46656	A B C	0.96	[1.3379; 1.3412)
		C B A	0.96	[1.3379; 1.3412)
		по ↘	0.91	[0.7450; 1.3704)
		по ↗	0.91	[0.7450; 1.3704)

5.1.5. Заключение. Из представленных в 4-ой и 5-ой колонках результатов вычислений напрашиваются следующие выводы:

1) наивысшее значение $\pi = 1$ коэффициента эффективности распознавания автора текста реализуется для произведений классической поэзии на триграммах, упорядоченных как по ABC, так и по CBA;

2) значения коэффициентов π эффективности на основе порядков ABC и CBA расположения -грамм ($N = 1, 2, 3$) равны;

3) значения коэффициентов π эффективности на основе порядков расположения -грамм ($N = 1, 2, 3$) по убыванию (↘) или возрастанию (↗) также равны;

4) значение коэффициента π эффективности на основе порядка ABC и CBA расположения -грамм ($N = 1, 2, 3$) не ниже значения, основанного на порядке расположения N-грамм ($N = 1, 2, 3$) по убыванию (↘) или возрастанию (↗);

5) коэффициент π эффективности распознавания автора произведений современной поэзии как для любых N-грамм ($N = 1, 2, 3$), так и для всех вариантов их упорядочения определяется значением 0.98;

6) коэффициенты π для произведений современной прозы несколько ниже аналогичных значений для произведений классической и современной поэзии;

7) полуинтервалы оптимальных значений γ для двух противоположных порядков расположения N-грамм ($N = 1, 2, 3$) одинаковы.

Из огромного количества всевозможных вариантов упорядоченного расположения элементов текста было рассмотрено только четыре: два из них – связаны с алфавитным порядком, и два других – с учётом частотности элементов. Именно в этих двух случаях, прямого и обратного порядков упорядочения элементов, расстояния между любыми парами произведений оказывались равными, вследствие чего равными оказывались коэффициенты π эффективности γ-классификатора (см. п.п. 2 и 3 заключения), а также и полуинтервалы оптимальных значений γ (см. п. 7 заключения). В §§ 5.2 – 5.4 исследуются другие допустимые варианты.

Результаты данного параграфа опубликованы в [19-А].

§ 5.2. О влиянии порядка символьных униграмм на идентификации автора произведения

На примере модельной коллекции, количественные описания произведений которой основываются на различных вариантах упорядочения буквенных униграмм (с учётом и без учёта пробелов), выявляются особенности применения γ-классификатора при распознавании автора текста.

В данном параграфе на примерах модельных коллекций текстов устанавливаются особенности ЦП и γ-классификатора в зависимости от упорядочения алфавитных элементов. Отметим, что ранее аналогичный вопрос изучался именно для символьных (буквенных) униграмм, биграмм и триграмм с учетом пробела, см. § 5.1. В предыдущих исследованиях из огромного количества всевозможных вариантов упорядоченного расположения элементов текста было рассмотрено только четыре: два из них – связаны с алфавитным порядком, и два других – с учётом частотности элементов. Существенным моментом в сравнении с нашим предыдущим исследованием является изучение вопроса с учётом всех допустимых вариантов.

5.2.1. Примеры текстовых элементов и их алфавитов. При изложении данного вопроса ограничимся рассмотрением простейших случаев, когда в качестве элементов текста выбираются буквенные униграммы (с учётом и без учёта пробелов).

Для униграмм естественных языков существующие алфавиты уже являются отсортированными в определенном порядке конечными множествами букв (также и с учётом пробела). Лексикографический порядок, аналогичный алфавитной сортировке, алфавитизирует также N-граммы ($N \geq 2$) и более сложные буквенно-символьные комбинации. Однако в дополнение к сказанному отметим, что такие комбинации, упорядоченные каким-либо другим способом, будут также называться алфавитными элементами текста. Как будет отмечено в п. 5.2.3, расстояние между ЦПТ зависит от порядка элементов алфавита, и поэтому не ясно, какому из допустимых алфавитов следует отдать предпочтение. Поскольку таджикский алфавит состоит из 35 букв, то множество различных упорядочений элементов будет равно $35! \approx 1.03 * 10^{40}$, а для расширенного алфавита с учетом пробела – $36! \approx 3.72 * 10^{41}$. Общее количество упорядочений алфавитных элементов называется генеральной совокупностью. Количество упорядочений очень много и их рассмотрение достаточно трудоёмко, поэтому случайным образом выбирается 100 упорядочений для получения результатов, а 10 – для тестирования. Если выбранные 10 случаев упорядочений для тестирования совпали (эффективность и гамма) со 100 случаями упорядочения, то по выборке можно сделать выводы о

свойствах всей генеральной совокупности, то есть она должна быть представительной (репрезентативной).

5.2.2. ЦП текстов и расстояние между ними. После выбора фиксированного алфавита ЦП текста T удобно представлять в табличной форме:

$$\begin{array}{lll} \bar{N} & : & 1 \quad 2 \quad . \; . \; . \quad m \\ P & : & p_1 \quad p_2 \quad . \; . \; . \quad p_m, \end{array} \tag{5.7}$$

в которой m – число элементов алфавита, строка $\bar{N}$ указывает номера упорядоченных элементов алфавита, а строка P – их относительные частоты встречаемости в T, причём

$$\sum_{k=1}^{m} p_k = 1.$$

ЦП можно задавать также дискретной функцией

$$F(s) = \sum_{k=1}^{s} p_k \qquad (s = 1, \cdots, m),$$

характеризующей распределение в тексте частот встречаемости элементов алфавита.

Определение 5.2.1. *Расстоянием между двумя текстам называется расстояние между их ЦП, отнесенными к единому алфавиту.*

Пусть $T_1,\ T_2$ – произвольная пара текстов из коллекции $\mathbb{T}$ и

$$F^{(\alpha)}(s) \ = \ \sum_{k=1}^{s} p_k^{(\alpha)} \ - \tag{5.8}$$

соответствующие им дискретные функции, $\alpha = 1, 2$ и $s = 1, \cdots, m$.

Определение 5.2.2. *Расстоянием между текстами T_1 и T_2 называется положительное число $\rho\,(T_1\,, T_2)$, определяемое по формуле*

$$\rho\,(T_1\,, T_2) = \sqrt{m/2}\ max_{\,s}\left|F^{(1)}(s) - \ F^{(2)}(s)\right|, \tag{5.9}$$

то есть расстояние между двумя текстами вычисляется как максимальное расстояние по оси ординат между их дискретными функциями $F^{(1)}(s)$ и $F^{(2)}(s)$, помноженное на весовой коэффициент $\sqrt{m/2}$.

Замечание. Условие $\rho\,(T_1\,, T_2) = 0$ означает тождество ЦП текстов, то есть ЦПT_1 = ЦПT_2, но не $T_1 = \ T_2$, то есть идентичность текстов.

5.2.3. Обработка данных коллекционного материала, представленного в § 5.1 п. 5.1.1, состояла из 3 этапов.

Этап 1. Использование для всех произведений трёх частей коллекции двух типов текстовых элементов:

•униграмм без учёта пробелов (в таджикском языке 35 букв алфавита):

•униграмм с учетом пробела (число таковых 36).

Множества униграмм в зависимости от упорядочения своих элементов рассматриваются в 100 случайным образом выбранных вариантах.

Этап 2. Для каждого из 100 вариантов упорядочения униграмм путём автоматической обработки формируются в табличном виде (5.7) цифровые портреты всех произведений коллекции, и затем по формулам (5.8) и (5.9) вычисляются расстояния между парами текстов на таджикском языке по отдельности из классической поэзии, современной поэзии и современной прозы. Из-за большого количества расстояний (таковых 27000 = 2 х 3 х 100 х 45) мы не приводим итоговых результатов, однако обращаем внимание на тот факт, что расстояния, вычисляемые между любыми двумя текстами для различных вариантов расположения алфавитных элементов, оказываются в общем случае различными. В этом можно убедиться на простых примерах.

Этап 3. Настройка γ-классификатора – алгоритма, зависящего от одного вещественного параметра γ и устанавливающего в пределах модельной коллекции соответствие между текстами и их авторами. Сущность настройки заключается в определении такого значения γ, при котором произведения одного автора «γ-однородны», а разных авторов – «γ-неоднородны». Однородность всех текстов одного автора в рамках математической модели означает справедливость неравенства

$$\rho\,(T_1\,, T_2) \leq\ \gamma, \tag{5.10}$$

а неоднородность любых двух текстов разных авторов – справедливость неравенства

$$\rho\,(T_1\,, T_2) > \gamma. \tag{5.11}$$

Ошибки в настройке γ-классификатора выявляется в случае, когда для каких-то пар текстов одного и того же автора вместо неравенства (5.10) имеет место неравенство (5.11), а также в случае, когда какие-то два произведения двух различных авторов удовлетворяют неравенство (5.10) вместо того, чтобы выполнялось неравенство (5.11).

Суммарное количество $\tau = \tau(\gamma)$ допущенных ошибок одновременно в двух случаях позволяет подсчитать величину π эффективности γ-классификатора при распознавании авторов текста по формуле

$$\pi = 1 -\ \tau(\gamma)/L\,, \tag{5.12}$$

где $L = 45$ – число взаимных расстояний между всеми парами произведений из

классической и современной поэзий, а также из современной прозы. Детальное описание алгоритма для нахождения оптимального значения γ, при котором π (5.12) принимает максимальное значение, содержится в § 1.4.

Итоги применения трёх этапов автоматической обработки модельной коллекции текстов показаны в таблице 5.4, соответственно для 3-х частей коллекции.

Таблица 5.4. – Значения π и γ в зависимости от 100 случайно выбранных упорядочений алфавитных элементов для произведений трех коллекций

Элементы текста	Число элементов алфавита	π	**Классическая поэзия**		**Современная поэзия**		**Современная проза**	
			Частота π	γ	Частота π	γ	Частота π	γ
униграммы	35	0.87	0	[0.0136; 0.0703]	12	[0.0137; 0.0514]	1	[0.0079; 0.0412]
		0.89	0		11		3	
		0.91	12		36		39	
		0.93	43		31		31	
		0.96	29		9		24	
		0.98	13		1		2	
		1	3		0		0	
	36	0.87	0	[0.0097; 0.0590]	3	[0.0117; 0.0492]	0	[0.0075; 0.0346]
		0.89	0		3		2	
		0.91	14		29		33	
		0.93	38		29		41	
		0.96	33		28		17	
		0.98	12		6		6	
		1	3		2		1	

В этой таблице и таблице 5.5 в 1-м столбце показаны элементы текста, во 2-м столбце – число элементов алфавита, в 3-м столбце – эффективность, полученная во время выборки упорядочения алфавитных элементов. Затем следуют три блока (по два столбца в каждом), указывающие результаты для произведений из классической поэзии, современной поэзии и современной прозы. Первый и второй столбцы в блоках отмечают частоту встречаемости эффективности π в выборке и оптимальное значение γ. Сумма столбцов частоты встречаемости эффективности π в зависимости от выбора элемента текста равна 100, это количество выборки. Значение эффективности π для всех трех коллекций принимается в диапазоне от 87% до 100%, а γ оптимальный также достаточно близкий.

5.2.4. Тестирование классификатора.

После того, как за счёт выбора 100 случайным образом упорядоченных элементов алфавита определена эффективность π и оптимальное значение γ, возникает естественный вопрос, а каковы будут результаты уже других 10 выборов, случайным образом упорядочений алфавитных элементов, соответствует ли значение π и γ.

Для тестирования классификатора выбрано случайным образом 10

упорядочений алфавитных элементов. Каждое упорядочение алфавита также, как это было сделано для 100 выборов, применяется для трех коллекций. Результаты показаны в таблице.

Таблица 5.5. – Значения π и γ в зависимости от 10 случайно выбранных упорядочений алфавитных элементов для произведений трех коллекций

Элементы текста	Число элементов алфавита	π	**Классическая поэзия**		**Современная поэзия**		**Современная проза**	
			Частота π	γ	Частота π	γ	Частота π	γ
униграммы	35	0.87	0	[0.0255; 0.0570]	0	[0.0168; 0.0398]	0	[0.0099; 0.0298]
		0.89	0		2		0	
		0.91	0		2		3	
		0.93	5		3		3	
		0.96	5		3		4	
		0.98	0		0		0	
		1	0		0		0	
	36	0.87	0	[0.0247; 0.0515]	1	[0.0190; 0.0385]	0	[0.0136; 0.0359]
		0.89	0		0		0	
		0.91	0		3		3	
		0.93	5		2		3	
		0.96	3		2		2	
		0.98	2		2		2	
		1	0		0		0	

Полученный результат показывает, что совпали значения π и γ.

5.2.5. Заключение. Из представленных результатов вычислений получаем следующие выводы:

1. Символьные униграммы являются вполне приемлемыми количественными характеристиками для решения проблемы идентификации авторов текстов.

2. Учёт пробелов в униграммах повышает точность классификации.

3. γ-классификатор показал высокий уровень идентификации авторов от 87% до 100%.

4. По мере увеличения числа случайно выбранных упорядочений алфавита повышается эффективность идентификации.

Из огромного количества возможных вариантов упорядочения расположения элементов текста были рассмотрены только 110, из которых 100 – для получения результатов, 10 – для тестирования результатов. Другие допустимые варианты можно не рассматривать, потому что результаты 10 случайно выбранных упорядочений алфавита для тестирования совпали с результатами 100 упорядочений.

Таким образом, математическая триада в составе ЦП текстов, представляемых распределениями частотности униграмм, формул (5.7) – (5.9) для вычисления расстояний между текстами и алгоритма для выявления однородных текстов оказалась подходящей для эффективного решения поставленной задачи.

Автор выражает уверенность в том, что еще увеличение объема исходной коллекции текстов не станет препятствием для успешного применения γ-классификатора не только для распознавания авторов, но также и для самых разнообразных однородностей текстовых документов.

Результаты данного параграфа опубликованы в [30-А].

§ 5.3. О влиянии порядка символьных биграмм на определение автора произведения

На примере модельной коллекции, количественные описания произведений которой основываются на различных вариантах упорядочения буквенных биграмм (с учётом и без учёта пробелов), выявляются особенности применения γ-классификатора при распознавании автора текста.

В данном параграфе на примерах модельных коллекций текстов устанавливаются особенности ЦП и γ-классификатора в зависимости от упорядочения алфавитных элементов.

5.3.1. Примеры текстовых элементов и их алфавитов. При изложении данного вопроса ограничимся рассмотрением простейших случаев, когда в качестве элементов текста выбираются буквенные биграммы (с учётом и без учёта пробелов).

Для униграмм естественных языков существующие алфавиты уже являются отсортированными в определенном порядке конечными множествами букв (также и с учётом пробела). Лексикографический порядок, аналогичный алфавитной сортировке, алфавитизирует также N-граммы ($N \geq 2$) и более сложные буквенно-символьные комбинации. Однако в дополнение к сказанному отметим, что такие комбинации, упорядоченные каким-либо другим способом, будут также называться алфавитными элементами текста. Как будет отмечено в п. 5.3.3, расстояние между ЦП текстов зависит от порядка элементов алфавита, и поэтому не ясно, какому из допустимых алфавитов следует отдать предпочтение. Таджикский алфавит состоит из 35 букв, двухбуквенные комбинации которых определяют множество различных биграмм в количестве $1225=35^2$, то множество различных упорядочений элементов будет равно $1225!\approx7.93*10^{3252}$, а для расширенного алфавита с учетом пробела – $36^2!=1296!\approx1.11*10^{3473}$. Общее количество упорядочений алфавитных элементов называется генеральной совокупностью. Количество упорядочений очень много и их рассмотрение достаточно трудоёмко, поэтому случайным образом выбирается 100 упорядочений для получения результатов, а 10 – для тестирования. Если выбранные 10 случаев упорядочений для тестирования совпали (эффективность и гамма) со 100 случаями упорядочения, то по выборке можно сделать выводы о

свойствах всей генеральной совокупности, то есть она должна быть представительной (репрезентативной).

5.3.2. ЦП текстов и расстояние между ними. После выбора фиксированного алфавита ЦП текста T удобно представлять в табличной форме:

$$\begin{array}{lllllll} \bar{N} & : & 1 & 2 & . & . & . & m \\ P & : & p_1 & p_2 & . & . & . & p_m, \end{array} \qquad (5.13)$$

в которой m – число элементов алфавита, строка $\bar{N}$ указывает номера упорядоченных элементов алфавита, а строка P – их относительные частоты встречаемости в T, причём

$$\sum_{k=1}^{m} p_k = 1.$$

ЦП можно задавать также дискретной функцией

$$F(s) = \sum_{k=1}^{s} p_k \qquad (s = 1, \cdots, m),$$

характеризующей распределение в тексте частот встречаемости элементов алфавита.

Определение 5.3.1. *Расстоянием между двумя текстам называется расстояние между их ЦП, отнесенными к единому алфавиту.*

Пусть T_1, T_2 – произвольная пара текстов из коллекции $\mathbb{T}$ и

$$F^{(\alpha)}(s) = \sum_{k=1}^{s} p_k^{(\alpha)} - \qquad (5.14)$$

соответствующие им дискретные функции, $\alpha = 1, 2$ и $s = 1, \cdots, m$.

Определение 5.3.2. *Расстоянием между текстами* T_1 *и* T_2 *называется положительное число* $\rho\,(T_1, T_2)$*, определяемое по формуле*

$$\rho\,(T_1, T_2) = \sqrt{m/2}\ max_s \left| F^{(1)}(s) - F^{(2)}(s) \right|, \qquad (5.15)$$

то есть расстояние между двумя текстами вычисляется как максимальное расстояние по оси ординат между их дискретными функциями $F^{(1)}(s)$ и $F^{(2)}(s)$, помноженное на весовой коэффициент $\sqrt{m/2}$.

Замечание. Условие $\rho\,(T_1, T_2) = 0$ означает тождество ЦП текстов, то есть ЦПT_1 = ЦПT_2, но не $T_1 = T_2$, то есть идентичность текстов.

5.3.3. Обработка данных коллекционного материала, представленного в § 5.1 п. 5.1.1, состояла из 3 этапов.

Этап 1. Использование для всех произведений трёх частей коллекции двух типов текстовых элементов:

•биграмм без учёта пробелов (общее число таковых – 35^2 =1225):

•биграмм с учетом пробела (число таковых 36^2 =1296).

Множества биграмм в зависимости от упорядочения своих элементов рассматриваются в 100 случайным образом выбранных вариантах.

Этап 2. Для каждого из 100 вариантов упорядочения биграмм путём автоматической обработки формируются в табличном виде (5.13) цифровые портреты всех произведений коллекции, и затем по формулам (5.14) и (5.15) вычисляются расстояния между парами текстов на таджикском языке по отдельности из классической поэзии, современной поэзии и современной прозы. Из-за большого количества расстояний (таковых 27000 = 2 х 3 х 100 х 45) мы не приводим итоговых результатов, однако обращаем внимание на тот факт, что расстояния, вычисляемые между любыми двумя текстами для различных вариантов расположения алфавитных элементов, оказываются в общем случае различными. В этом можно убедиться на простых примерах.

Этап 3. Настройка γ-классификатора – алгоритма, зависящего от одного вещественного параметра γ и устанавливающего в пределах модельной коллекции соответствие между текстами и их авторами. Сущность настройки заключается в определении такого значения γ, при котором произведения одного автора «γ-однородны», а разных авторов – «γ-неоднородны». Однородность всех текстов одного автора в рамках математической модели означает справедливость неравенства

$$\rho\,(T_1\,,T_2) \leq\ \gamma, \tag{5.16}$$

а неоднородность любых двух текстов разных авторов – справедливость неравенства

$$\rho\,(T_1\,,T_2) > \gamma. \tag{5.17}$$

Ошибки в настройке γ-классификатора выявляются в случае, когда для каких-то пар текстов одного и того же автора вместо неравенства (5.16) имеет место неравенство (5.17), а также в случае, когда какие-то два произведения двух различных авторов удовлетворяют неравенство (5.16) вместо того, чтобы выполнялось неравенство (5.17).

Суммарное количество $\tau = \tau(\gamma)$ допущенных ошибок одновременно в двух случаях позволяет подсчитать величину π эффективности γ-классификатора при распознавании авторов текста по формуле

$$\pi = 1 -\ \tau(\gamma)/L\,, \tag{5.18}$$

где $L = 45$ – число взаимных расстояний между всеми парами произведений из

классической и современной поэзий, а также из современной прозы. Детальное описание алгоритма для нахождения оптимального значения γ, при котором π (5.18) принимает максимальное значение, содержится в § 1.4.

Итоги применения трёх этапов автоматической обработки модельной коллекции текстов показаны в таблице 5.6, соответственно для 3-х частей коллекции.

Таблица 5.6. – Значения π и γ в зависимости от 100 случайно выбранных упорядочений алфавитных элементов для произведений трех коллекций

Элементы текста	Число элементов алфавита	π	**Классическая поэзия**		**Современная поэзия**		**Современная проза**	
			Частота π	γ	Частота π	γ	Частота π	γ
биграммы	1225	0.87	1	[0.0709; 0.3389]	12	[0.0846; 0.2659]	0	[0.0523; 0.2498]
		0.89	0		15		1	
		0.91	37		32		62	
		0.93	38		26		28	
		0.96	20		10		8	
		0.98	3		3		1	
		1	1		2		0	
	1296	0.87	0	[0.0565; 0.4043]	4	[0.1204; 0.3201]	0	[0.0565; 0.3080]
		0.89	0		10		1	
		0.91	26		18		36	
		0.93	37		31		38	
		0.96	25		20		20	
		0.98	9		17		5	
		1	3		0		0	

В этой таблице и таблице 5.7 в 1-м столбце показаны элементы текста, во 2-м столбце – число элементов алфавита, в 3-м столбце – эффективность, полученная во время выборки упорядочения алфавитных элементов. Затем следуют три блока (по два столбца в каждом), указывающие результаты для произведений из классической поэзии, современной поэзии и современной прозы. Первый и второй столбцы в блоках отмечают частоту встречаемости эффективности π в выборке и оптимальное значение γ. Сумма столбцов частоты встречаемости эффективности π в зависимости от выбора элемента текста равна 100, это количество выборки. Значение эффективности π для всех трех коллекций принимается в диапазоне от 87% до 100%, а γ оптимальный также достаточно близкий.

5.3.4. Тестирование классификатора

После того, как за счёт выбора 100 случайным образом упорядоченных элементов алфавита определена эффективность π и оптимальное значение γ, возникает естественный вопрос, а каковы будут результаты уже других 10 выборов, случайным образом упорядочений алфавитных элементов, соответствует ли значение π и γ.

Для тестирования классификатора выбрано случайным образом 10

упорядочений алфавитных элементов. Каждое упорядочение алфавита также, как это было сделано для 100 выборов, применяется для трех коллекций. Результаты показаны в таблице.

Таблица 5.7. – Значения π и γ в зависимости от 10 случайно выбранных упорядочений алфавитных элементов для произведений трех коллекций

Элементы текста	Число элементов алфавита	π	**Классическая поэзия**		**Современная поэзия**		**Современная проза**	
			Частота π	γ	Частота π	γ	Частота π	γ
биграммы	1225	0.87	0	[0.0810; 0.3135]	1	[0.1279; 0.2154]	0	[0.0526; 0.1884]
		0.89	0		1		0	
		0.91	2		3		10	
		0.93	5		2		0	
		0.96	3		2		0	
		0.98	0		1		0	
		1	0		0		0	
	1296	0.87	0	[0.1212; 0.3323]	0	[0.1617; 0.3214]	0	[0.1025; 0.2436]
		0.89	0		1		0	
		0.91	2		2		3	
		0.93	2		2		2	
		0.96	3		4		4	
		0.98	3		1		1	
		1	0		0		0	

Полученный результат показывает, что совпали значения π и γ.

5.3.5. Заключение. Из представленных результатов вычислений получаем следующие выводы:

1. Символьные биграммы являются вполне приемлемыми количественными характеристиками для решения проблемы идентификации авторов текстов.

2. Учёт пробелов в биграммах повышает точность классификации.

3. γ-классификатор показал высокий уровень идентификации авторов от 87% до 100%.

4. По мере увеличения числа случайно выбранных упорядочений алфавита повышается эффективность идентификации.

Из огромного количества возможных вариантов упорядочения расположения элементов текста были рассмотрены только 110, из которых 100 – для получения результатов, 10 – для тестирования результатов. Другие допустимые варианты можно не рассматривать, потому что результаты 10 случайно выбранных упорядочений алфавита для тестирования совпали с результатами 100 упорядочений.

Таким образом, математическая триада в составе ЦП текстов, представляемых распределениями частотности биграмм, формул (5.13) – (5.15) для вычисления расстояний между текстами и алгоритмом для выявления однородных текстов оказалась подходящей для эффективного решения

поставленной задачи.

Автор выражает уверенность в том, что еще увеличение объема исходной коллекции текстов не станет препятствием для успешного применения γ-классификатора не только для распознавания авторов, но также и для самых разнообразных однородностей текстовых документов.

Результаты данного параграфа опубликованы в [66-А].

§ 5.4. О влиянии порядка символьных триграмм на идентификации автора произведения

На примере модельной коллекции, количественные описания произведений которой основываются на различных вариантах упорядочения буквенных триграмм (с учётом и без учёта пробелов), выявляются особенности применения γ-классификатора при распознавании автора текста.

В данном параграфе на примерах модельных коллекций текстов устанавливаются особенности ЦП и γ-классификатора в зависимости от упорядочения алфавитных элементов. Отметим, что ранее аналогичный вопрос изучался именно для символьных (буквенных) униграмм, биграмм и триграмм с учетом пробела, см. § 5.1. В предыдущих исследованиях из огромного количества всевозможных вариантов упорядоченного расположения элементов текста было рассмотрено только четыре: два из них – связаны с алфавитным порядком и два других – с учётом частотности элементов. Существенным моментом в сравнении с нашим предыдущим исследованием является изучение вопроса с учётом всех допустимых вариантов.

5.4.1. Примеры текстовых элементов и их алфавитов. При изложении данного вопроса ограничимся рассмотрением простейших случаев, когда в качестве элементов текста выбираются буквенные триграммы (с учётом и без учёта пробелов).

Как будет отмечено в п. 5.4.3, расстояние между ЦПТ зависит от порядка элементов алфавита, и поэтому не ясно, какому из допустимых алфавитов следует отдать предпочтение. Таджикский алфавит состоит из 35 букв, трехбуквенные комбинации которых определяют множество различных триграмм в количестве $42875=35^3$, то множество различных упорядочений элементов будет равно 42875!, а для расширенного алфавита с учетом пробела – 36^3!=46656!. Общее количество упорядочений алфавитных элементов называется генеральной совокупностью. Количество упорядочений очень много и их рассмотрение достаточно трудоёмко, поэтому случайным образом выбирается 100 упорядочений для получения результатов, а 10 – для тестирования. Если выбранные 10 случаев упорядочений для тестирования совпали (эффективность и гамма) со 100 случаями упорядочения, то по выборке можно сделать выводы о свойствах всей

генеральной совокупности, то есть она должна быть представительной (репрезентативной).

5.4.2. ЦПТ и расстояние между ними. После выбора фиксированного алфавита ЦП текста T удобно представлять в табличной форме:

$$\begin{array}{lllllll} \bar{N} & : & 1 & 2 & . & . & . \quad m \\ P & : & p_1 & p_2 & . & . & . \quad p_m, \end{array} \tag{5.19}$$

в которой m – число элементов алфавита, строка $\bar{N}$ указывает номера упорядоченных элементов алфавита, а строка P – их относительные частоты встречаемости в T, причём

$$\sum_{k=1}^{m} p_k = 1.$$

ЦП можно задавать также дискретной функцией

$$F(s) = \sum_{k=1}^{s} p_k \qquad (s = 1, \cdots, m),$$

характеризующей распределение в тексте частот встречаемости элементов алфавита.

Определение 5.4.1. *Расстоянием между двумя текстами называется расстояние между их ЦП, отнесенными к единому алфавиту.*

Пусть T_1, T_2 – произвольная пара текстов из коллекции $\mathbb{T}$ и

$$F^{(\alpha)}(s) = \sum_{k=1}^{s} p_k^{(\alpha)} - \tag{5.20}$$

соответствующие им дискретные функции, $\alpha = 1, 2$ и $s = 1, \cdots, m$.

Определение 5.4.2. *Расстоянием между текстами* T_1 *и* T_2 *называется положительное число* $\rho\,(T_1, T_2)$*, определяемое по формуле*

$$\rho\,(T_1, T_2) = \sqrt{m/2}\ max_s \left| F^{(1)}(s) - F^{(2)}(s) \right|, \tag{5.21}$$

то есть расстояние между двумя текстами вычисляется как максимальное расстояние по оси ординат между их дискретными функциями $F^{(1)}(s)$ и $F^{(2)}(s)$, помноженное на весовой коэффициент $\sqrt{m/2}$.

Замечание. Условие $\rho\,(T_1, T_2) = 0$ означает тождество ЦП текстов, то есть ЦПT_1 = ЦПT_2, но не $T_1 = T_2$, то есть идентичность текстов.

5.4.3. Обработка данных коллекционного материала, представленного в § 5.1 п. 5.1.1, состояла из 3 этапов.

Этап 1. Использование для всех произведений трёх частей коллекции двух типов текстовых элементов:

- триграмм без учёта пробелов (общее число таковых – 35^3 =42875):
- триграмм с учетом пробела (число таковых 36^3 =46656).

Множества триграмм в зависимости от упорядочения своих элементов рассматриваются в 100 случайным образом выбранных вариантах.

Этап 2. Для каждого из 100 вариантов упорядочения триграмм путём автоматической обработки формируются в табличном виде (5.19) цифровые портреты всех произведений коллекции, и затем по формулам (5.20) и (5.21) вычисляются расстояния между парами текстов на таджикском языке по отдельности из классической поэзии, современной поэзии и современной прозы. Из-за большого количества расстояний (таковых 27000 = 2 x 3 x 100 x 45) мы не приводим итоговых результатов, однако обращаем внимание на тот факт, что расстояния, вычисляемые между любыми двумя текстами для различных вариантов расположения алфавитных элементов, оказываются в общем случае различными. В этом можно убедиться на простых примерах.

Этап 3. Настройка γ-классификатора – алгоритма, зависящего от одного вещественного параметра γ и устанавливающего в пределах модельной коллекции соответствие между текстами и их авторами. Сущность настройки заключается в определении такого значения γ, при котором произведения одного автора «γ-однородны», а разных авторов – «γ-неоднородны». Однородность всех текстов одного автора в рамках математической модели означает справедливость неравенства

$$\rho\,(T_1\,,T_2) \leq\ \gamma, \tag{5.22}$$

а неоднородность любых двух текстов разных авторов – справедливость неравенства

$$\rho\,(T_1\,,T_2) > \gamma. \tag{5.23}$$

Ошибки в настройке γ-классификатора выявляются в случае, когда для каких-то пар текстов одного и того же автора вместо неравенства (5.22) имеет место неравенство (5.23), а также в случае, когда какие-то два произведения двух различных авторов удовлетворяют неравенство (5.22) вместо того, чтобы выполнялось неравенство (5.23).

Суммарное количество $\tau = \tau(\gamma)$ допущенных ошибок одновременно в двух случаях позволяет подсчитать величину π эффективности γ-классификатора при распознавании авторов текста по формуле

$$\pi = 1 -\ \tau(\gamma)/L\,, \tag{5.24}$$

где $L = 45$ – число взаимных расстояний между всеми парами произведений из классической и современной поэзий, а также из современной прозы. Детальное описание алгоритма для нахождения оптимального значения γ, при котором π (5.24) принимает максимальное значение, содержится в § 1.4.

Итоги применения трёх этапов автоматической обработки модельной коллекции текстов показаны в таблице 5.8, соответственно для 3-х частей коллекции.

Таблица 5.8. – Значения π и γ в зависимости от 100 случайно выбранных упорядочений алфавитных элементов для произведений трех коллекций

Элементы текста	Число элементов алфавита	π	**Классическая поэзия**		**Современная поэзия**		**Современная проза**	
			Частота π	γ	Частота π	γ	Частота π	γ
триграммы	42875	0.87	4	[0.4177; 1.6329]	12	[0.4791; 1.3703]	1	[0.3078; 1.2319]
		0.89	4		13		1	
		0.91	50		36		54	
		0.93	23		26		38	
		0.96	15		8		5	
		0.98	4		4		1	
		1	0		1		0	
	46656	0.87	0	[0.4205; 1.7642]	6	[0.5853; 1.7554]	0	[0.3019; 1.3990]
		0.89	2		9		3	
		0.91	34		36		47	
		0.93	42		30		43	
		0.96	15		16		6	
		0.98	7		3		1	
		1	0		0		0	

В этой таблице и таблице 5.9 в 1-м столбце показаны элементы текста, во 2-м столбце – число элементов алфавита, в 3-м столбце – эффективность, полученная во время выборки упорядочения алфавитных элементов. Затем следуют три блока (по два столбца в каждом), указывающие результаты для произведений из классической поэзии, современной поэзии и современной прозы. Первый и второй столбцы в блоках отмечают частоту встречаемости эффективности π в выборке и оптимальное значение γ. Сумма столбцов частоты встречаемости эффективности π в зависимости от выбора элемента текста равна 100, это количество выборки. Значение эффективности π для всех трех коллекций принимается в диапазоне от 87% до 100%, а γ оптимальный также достаточно близкий.

5.4.4. Тестирование классификатора

После того, как за счёт выбора 100 случайным образом упорядоченных элементов алфавита определена эффективность π и оптимальное значение γ, возникает естественный вопрос, а каковы будут результаты уже других 10 выборов, случайным образом упорядочений алфавитных элементов, соответствует ли значение π и γ.

Для тестирования классификатора выбрано случайным образом 10 упорядочений алфавитных элементов. Каждое упорядочение алфавита также, как это было сделано для 100 выборов, применяется для трех коллекций. Результаты показаны в таблице.

Таблица 5.9. – Значения π и γ в зависимости от 10 случайно выбранных упорядочений алфавитных элементов для произведений трех коллекций

Элементы текста	Число элементов алфавита	π	**Классическая поэзия**		**Современная поэзия**		**Современная проза**	
			Частота π	γ	Частота π	γ	Частота π	γ
триграммы	42875	0.87	0	[0.4694; 1.2826]	2	[0.6822; 1.1091]	0	[0.3627; 1.0337]
		0.89	1		2		0	
		0.91	3		3		6	
		0.93	4		2		2	
		0.96	2		1		2	
		0.98	0		0		0	
		1	0		0		0	
	46656	0.87	0	[0.5785; 1.5512]	1	[0.6711; 1.2874]	0	[0.4469; 1.1940]
		0.89	0		0		3	
		0.91	2		4		4	
		0.93	6		4		1	
		0.96	2		1		2	
		0.98	0		0		0	
		1	0		0		0	

Полученный результат показывает, что совпали значение π и γ.

5.4.5. Заключение. Из представленных результатов вычислений получаем следующие выводы:

1. Символьные триграммы являются вполне приемлемыми количественными характеристиками для решения проблемы идентификации авторов текстов.

2. Учёт пробелов в триграммах повышает точность классификации.

3. γ-классификатор показал высокий уровень идентификации авторов от 87% до 100%.

4. По мере увеличения числа случайно выбранных упорядочений алфавита повышается эффективность идентификации.

Из огромного количества возможных вариантов упорядочения расположения элементов текста были рассмотрены только 110, из которых 100 – для получения результатов, 10 – для тестирования результатов. Другие допустимые варианты можно не рассматривать, потому что результаты 10 случайно выбранных упорядочений алфавита для тестирования совпали с результатами 100 упорядочений.

Таким образом, математическая триада в составе ЦП текстов, представляемых распределениями частотности триграмм, формул (5.19) – (5.21) для вычисления расстояний между текстами и алгоритмом для выявления однородных текстов оказалась подходящей для эффективного решения поставленной задачи.

Автор выражает уверенность в том, что еще увеличение объема исходной коллекции текстов не станет препятствием для успешного применения γ-классификатора не только для распознавания авторов, но также и для самых разнообразных однородностей текстовых документов.

Результаты данного параграфа опубликованы в [31-А].

ГЛАВА 6. ПРОГРАММНЫЙ ПРОДУКТ «THR»

Программный комплекс **«THR»** (Text Homogeneity Recognition) предназначен для распознавания однородности текста. Программа **«THR»** – это автоматический определитель однородности текста, который поможет определить на каком языке написан, автора, тематики, УДК, шифра специальности, перевода и его оригинала, пола автора, поэзии или прозы, жанра и группы языков текста или даже фрагмента текста. Программа поддерживает более 100 языков. Открываемые файлы с текстами должны быть в кодировке Уникод или ANSI.

§ 6.1. Блок-схема программного система «THR»

Основные процедуры, входящие в состав программного комплекса, показаны на рисунке 6.1.

Программный комплекс «THR» начинает свою работу с ввода текста, обозначаемого буквой T, см. блок 1. В качестве T могут выступать тексты любого размера, в частности произведение в полном объеме, его фрагменты, а также короткие тексты.

В блоке 2 выполняются процедуры подготовки текста к последующим обработкам: преобразование к единому регистру, удаление символов, отличных от пробела и букв таджикского алфавита, также различных видов правок.

В блоке 3 определяется размер $\ell(T)$ текста T количеством слов:

– если $\ell(T) < 20$ слов, то программный комплекс направляет сообщение в блок 14 и прекращает работу;

– иначе – переход в блок 4.

В блоке 4 пользователю предлагается выбрать процентное сравнение или по формуле расстояние.

В блоке 5 пользователю предлагается выбрать признаки однородности.

В блоке 6 пользователю предлагается выбрать по своему желанию тот или иной элементы текста для распознавания T-текста.

В блоке 7 пользователь отмечает, какой из двух альтернатив – полным элементом текста или же только его высокочастотной частью – он предпочитает воспользоваться для распознавания текста T.

В блоке 8 производится вычисление распределения частотностей элементов T-текста и в блоке 9 пользователю вновь предоставляется выбор альтернативы либо сравнивать T со всеми произведениями базы данных, либо ограничиться некоторыми из них.

Отмеченные произведения вместе со значением γ, определяемым длиной $\ell(T)$, извлекаются из базы данных 11 в блок 10.

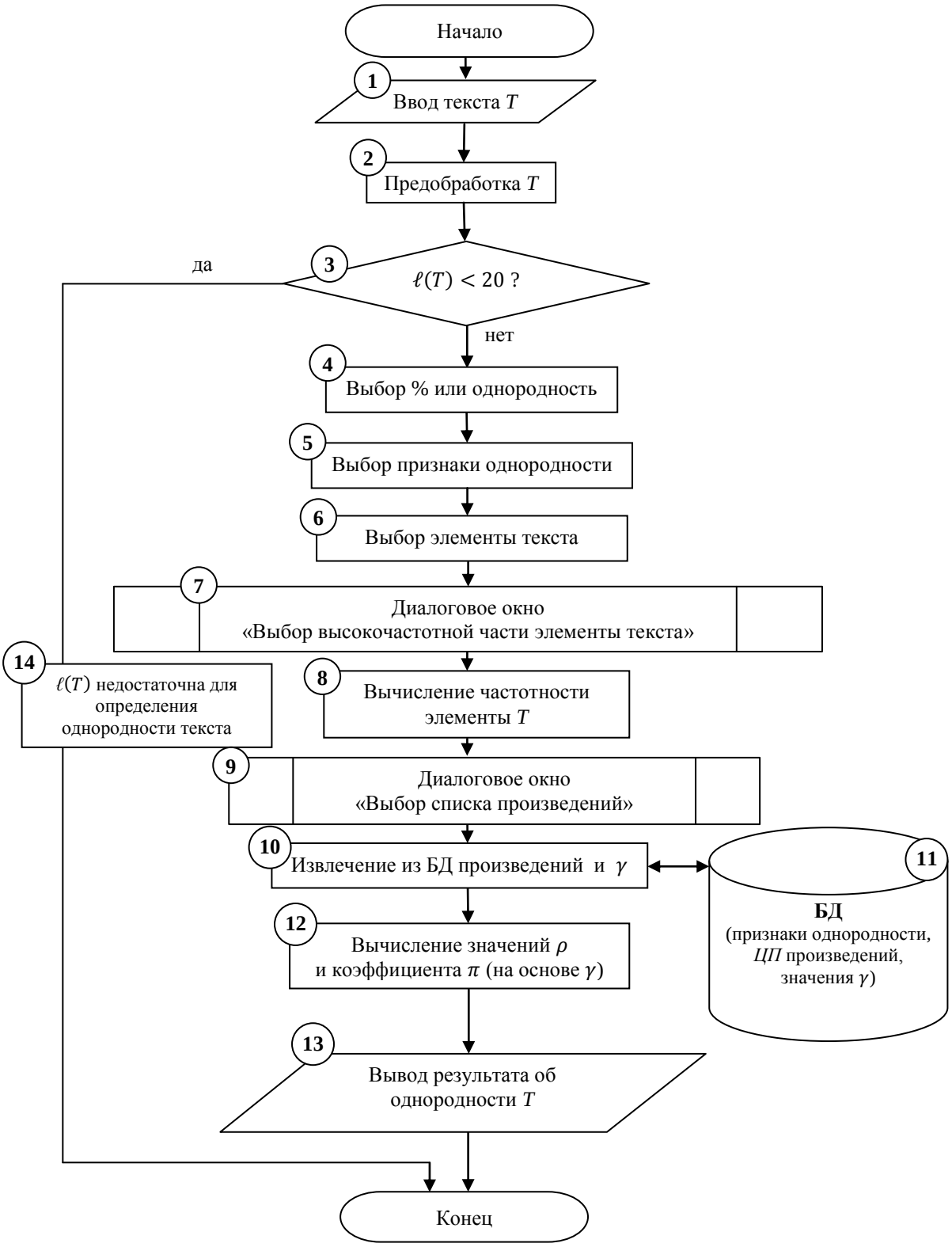


Рисунок 6.1. – Блок-схема программного комплекса для идентификации однородности текста

В блоке 12 производятся вычисления расстояний ρ между T-текстом и всеми извлеченными произведениями. Затем по значениям ρ и γ вычисляется показатель эффективности π. По полученным данным в блоке 13 выдаются итоговые результаты.

§ 6.2. БД для хранения произведений и их характеристик

Программный комплекс «THR» предназначен для определения однородности текстов из различных сфер человеческой деятельности, в этом и последующих параграфах его описание производится для случая, когда его база данных содержит художественные и научные произведения.

Во время экспериментов по определению однородности обработке подвергаются большие массивы текстовой информации, что может потребовать существенных временных затрат. Хранение текста в виде реляционных таблиц требует значительного количества времени лишь на этапе загрузки информации в базу данных. Вместе с тем это предоставляет ряд преимуществ: быстрое манипулирование данными на этапе их обработки, использование возможностей конкретной СУБД для облегчения труда исследователя и т.д.

Большинство методов идентификации однородности текста используют в качестве ключевых параметров характеристики уровня символов, слов и предложений. Основные операции, интересующие исследователей, связаны с получением наборов агрегированных параметров текста.

Для исследовательских целей необходимо обеспечить возможность извлечения информации из БД для выборок разного объема. На основе исследования была разработана концептуальная модель БД для хранения характеристик текста, представленная на рисунке 6.2, проведено физическое проектирование базы данных на основе СУБД MySQL.

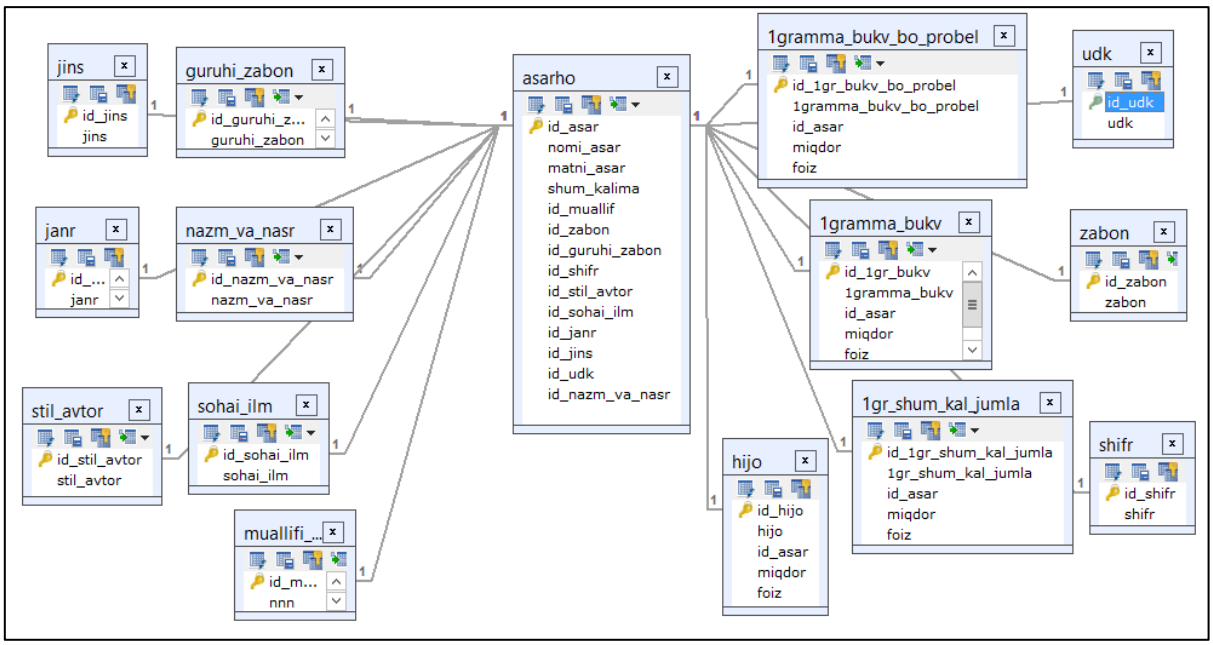


Рисунок 6.2. – Концептуальная модель базы данных для хранения характеристик текста

Текст в базе данных хранится как список распределения элементов алфавита. Рассмотрим таблицы и их поля подробнее.

1. Таблица «Авторы» (muallifi_asar) предназначена для хранения информации об авторах и содержит следующие поля: идентификатор записи (id_muallif), фамилию, имя, отчество автора (поля nnn соответственно), год рождения автора (soli_tavallud), год смерти (soli_vafot), пол автора (jins), возраст (sinnu_sol), профессию автора (vaz_muallif) и национальную принадлежность (millat).

2. Таблица «Произведения» (asarho) содержит следующие поля: идентификатор произведения автора (id_asar), название произведения (nomi_asar), текст произведения (matni_asar), количество слов (shum_kalima), идентификатор автора произведения (id_muallif), язык произведения (id_zabon), поэзия или проза (id_nazm_va_nasr), жанр произведения (id_janr) и т.д.

3. Таблица «Буквенные униграммы» (1gramma_bukv) содержит следующие поля: идентификатор униграмм произведения (id_1gr_bukv), сами униграммы (1gramma_bukv), идентификатор произведения автора (id_asar), абсолютные частоты униграмм в тексте (miqdor), относительные частоты униграмм (foiz).

4. Таблица «Униграммы с учетом пробела» (1gramma_bukv_bo_probel) содержит следующие поля: идентификатор униграмм произведения (id_1gr_bukv_bo_probel), сами униграммы (1gramma_bukv_bo_probel), идентификатор произведения автора (id_asar), абсолютные частоты униграмм в тексте (miqdor), относительные частоты униграмм (foiz).

5. Таблица «Слоги» (hijo) содержит следующие поля: идентификатор слогов произведения (id_hijo), сами слоги (hijo), идентификатор произведения автора (id_asar), абсолютные частоты слогов в тексте (miqdor), относительные частоты слогов (foiz).

6. Таблица «Длина предложений (в словах)» (1gr_shum_kal_jumla) содержит следующие поля: идентификатор длин предложений произведения (id_1gr_shum_kal_jumla), сами длины предложений (1gr_shum_kal_jumla), идентификатор произведения автора (id_asar), абсолютные частоты длин предложений в тексте (miqdor), относительные частоты длин предложений (foiz).

7. Таблицы zabon, guruhi_zabon, janr, jins, nazm_va_nasr, shifr, sohai_ilm, stil_avtor и udk содержат иформацию о языках, группах языков, жанре, поле автора, поэзии и прозе, шифре специальности научных работ, сфере науки, стиле автора и УДК.

В процессе разработки структуры БД было установлено, что хранить агрегированные характеристики текста непосредственно в таблицах нецелесообразно. Их можно получить с помощью запросов на языке SQL. Информацию о сочетаниях элементов текста можно также получить с помощью SQL-запросов.

Разработанная БД для хранения текстов позволяет значительно упростить

исследование характеристик текста в задачах идентификации однородности: за счет использования языка запросов SQL из базы данных можно извлекать наборы характеристик текста практически любой сложности и использовать эту информацию при дальнейшем анализе. Для принятия решения о признаках однородности могут использоваться любые методы классификации, кластеризации или проверки текстов на однородность и близость авторских стилей, использующие и хранящиеся в БД характеристики.

§ 6.3. Примеры SQL-запроса к БД

Приведём пример листинга запросов для работы с автором, произведением и распределениями элементов алфавита.

1. SELECT * from asarho where nomi_asar='" & sortedList(index).asarho & "';

2. SELECT zabon from zabon where id_zabon='" & id_zab & "';

3. SELECT b.id_asar, COUNT(a.id_asar) AS miq FROM " & chgramma1 & " AS b,asarho AS a WHERE b.id_asar=a.id_asar && a.namudi_asar='Матн' GROUP BY b.id_asar;

4. SELECT id_asar, count(id_asar) FROM " & chgramma1 & " group by id_asar;

5. SELECT muallifi_asar.nnn FROM muallifi_asar where id_muallif=(Select id_muallif from asarho where id_asar='" & id_asar2 & "');

6. SELECT id_asar, shum_kalima FROM asarho where id_muallif in (select id_muallif from muallifi_asar where nnn='" & nshoir2 & "') and nomi_asar='" & nasar2 & "';

7. SELECT nomi_asar FROM asarho where id_muallif in (select id_muallif from muallifi_asar where nnn='" & ComboBox1.Text & "');

8. SELECT id_guruhi_zabon from guruhi_zabon where guruhi_zabon='" & ComboBox4.Text & "';

9. SELECT id_shifr from shifr where shifr='" & ComboBox5.Text & "';

10. INSERT Into asarho SET nomi_asar='" & ComboBox2.Text & "', matni_asar ='" & TextBox1.Text & "', shum_kalima ='" & TextBox2.Text & "', id_muallif ='" & id_mualf & "', id_zabon ='" & id_zab & "', id_guruhi_zabon ='" & id_gur_zab & "', id_shifr ='" & id_shf & "', id_stil_avtor ='" & id_stil_avt & "', id_sohai_ilm ='" & id_sohai_il & "', id_janr ='" & id_jnr & "', id_jins ='" & id_jns & "', id_udk ='" & id_ud & "', id_nazm_va_nasr ='" & id_nzm_va_nsr & "';

11. DELETE FROM " & bigram_filename & " WHERE id_asar=" & id_as & ";

12. INSERT into asarho(nomi_asar, id_muallif) values('" & ComboBox2.Text & "', '" & id_sh & "');" & vbNewLine & "SELECT id_asar FROM asarho WHERE nomi_asar='" & ComboBox2.Text & "' and id_muallif='" & id_sh & "';

13. UPDATE asarho SET nomi_asar='" & TextBox1.Text & "', matni_asar ='" & RichTextBox1.Text & "', soli_navishti_asar ='" & TextBox2.Text & "', id_muallif ='" &

id_sh & "', joi_navishti_asar ='" & TextBox3.Text & "', id_kitob ='" & id_kb & "', zaboni_asar ='" & ComboBox3.Text & "', nazm_yo_nasr ='" & ComboBox4.Text & "', vazni_sherho ='" & ComboBox5.Text & "', janri_asar ='" & ComboBox6.Text & "', shaklhoi_asar ='" & ComboBox7.Text &"', navi_asar ='" & ComboBox8.Text &"', sanati_sherho ='" & ComboBox9.Text &"', mp3 ='" & TextBox4.Text &"', ejodiyot ='" & ComboBox10.Text &"', shum_kalima ='" & TextBox5.Text & "' WHERE id_asar = '" & TextBox6.Text & "';

14. SELECT 1gr_shum_kal_jumla from 1gr_shum_kal_jumla where id_asar='" & id_asr & "';

15. SELECT hijo from hijo where id_asar='" & id_asr & "';

16. SELECT 1gramma_bukv_bo_probel from 1gramma_bukv_bo_probel where id_asar='" & id_asr & "';

17. SELECT nazm_va_nasr from nazm_va_nasr where id_nazm_va_nasr='" & id_nzm_va_nsr & "';

18. SELECT udk from udk where id_udk='" & id_ud & "';

19. SELECT jins from jins where id_jins='" & id_jns & "';

20. INSERT into " & bigram_filename & " (" & bigram_filename & ", id_asar, miqdor, foiz) values ('" & mas1(i) & "', '" & id_as & "', '" & mas2(i) & "', '" & Math.Round((mas2(i) / sumstr), 15) & "');.

§ 6.4. Интерфейс программного продукта «THR»

Программная система «THR» содержит следующие подпрограммы:

1) корректировка текста;

2) соединение с базой данных (БД);

3) поиск, добавить, изменить, удалить или просмотреть информацию об авторах;

4) поиск, добавить, изменить, удалить или просмотреть информацию о произведении автора;

5) поиск, добавить, изменить, удалить или просмотреть информацию о произведениях автора;

6) поиск, добавить, изменить, удалить или просмотреть различные ЦПТ для фрагмента или полного произведения автора;

7) программа для вычисления значения $\gamma^{\text{опт}}$ и эффективности π;

8) программа для определения однородности по фрагменту или полному тексту.

Главное окно программы. Графический интерфейс программы выполнен по технологии SDI (однодокументный интерфейс). Главное окно программы (см. рисунок 6.3) панель инструментов с пиктограммами, соответствующими основным операциям, которые способны выполнять приложение.

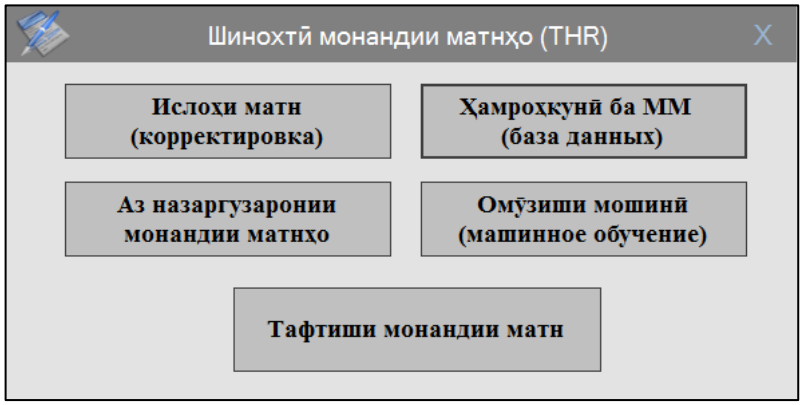


Рисунок 6.3. – Главное окно программы

Главное окно программы состоит из пяти основных пунктов:

1) «Ислоҳи матн» предназначено для работы корректировки текста;

2) «Ҳамроҳкунӣ ба ММ» предназначено для работы с БД;

3) «Аз назаргузаронии монандии матнҳо» – её основной функцией является вывод результатов в понятной и наглядной для исследователя форме;

4) «Омӯзиши мошинӣ» – предназначено для машинного обучения, компьютер связан с задачей обучения, находится эффективный порядок алфавита, и этот алфавит сохраняется;

5) «Тафтиши монандии матн» предназначено для определения однородности по фрагменту или полному тексту.

Корректировка текста. Для анализа текста и добавления его в базу данных первоначально произвести исправление ошибок. Для этого нажать кнопку «Ислоҳи матн (корректировка текста)», после чего появится окно (см. рисунок 6.4).

Пользователю предлагается выполнить следующие действия:

а) выбрать кодировку текста самостоятельно: Windows_1251, Unicode, Windows_1256, UTF-8 или определить её автоматически.

б) выбрать возможные варианты исправления текста. Если отметить поля «множественные пробелы (маҷмӯи пробелҳо)», «множественные тире (маҷмӯи тиреҳо)», «одинаковые слова подряд (калимаҳои якхела пайдарпай)», «переносы (аз сатр ба сатр калимаҳои гузаронидашуда)», «служебные символы (калимаҳои ёрирасон)» (или некоторые из этих полей в любой комбинации), то из текста будут удалены соответствующие структуры.

Дополнительно в поле «удалять также (несткунӣ боз)» можно ввести другие символы и их последовательности, которые необходимо удалять из текста.

Если отметить поле «замена латиницы таджикскими буквами (ивази символҳои лотинӣ ба тоҷикӣ)», то в тексте будет осуществлена замена символов латиницы таджикскими символами, схожими по начертанию.

Если отметить поле «замена всех кавычек универсальными (ивази ҳамаи намудӣ нохунакҳо ба намуди универсалӣ)», то в тексте будет осуществлена

замена всех кавычек универсальными.

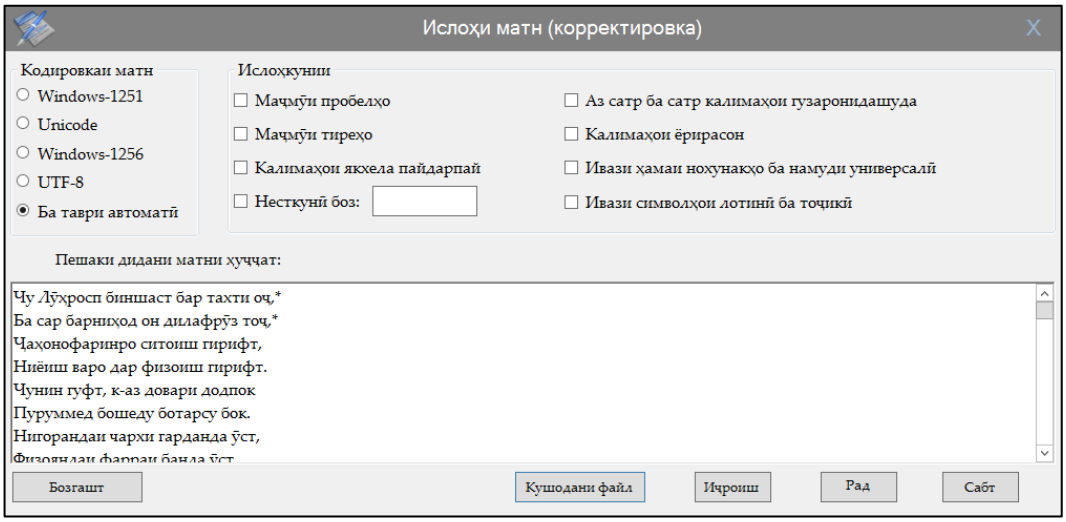


Рисунок 6.4. – Корректировка текста

Откорректированный текст можно сохранить в файл, нажав кнопку «Сохранить (Сабт)». Для дальнейшего анализа текста необходимо нажать кнопку «Принять (Иҷроиш)», для закрытия окна – «Отмена (Рад)» и для открытия нового файла необходимо нажать кнопку «Открыть (Кушодани файл)».

Для выполнения операции анализа текста и формирования файлов для исследований требуется подключение к базе данных. Пользователю предлагается ввести информацию о сервере (доменное имя или IP адрес), на котором установлена база данных; о порте, на котором работает СУБД; об имени и пароле пользователя, имеющего необходимые привилегии для работы с базой данных; о названии базы данных.

Программа для добавления в базу данных информации о фрагментах или полных произведениях автора и их ЦП, см. рисунок 6.5. С помощью кнопки «Интихоби файли асар» можно выбрать нужный нам файл, потом в пункте «Номи муаллиф» выбрать фамилию, имя и отчество автора текста. Если автор неизвестен, тогда с помощью правой кнопки можно добавить нового автора и его произведение. Потом выбрать следующие данные: количество слов, название файла, название произведения, язык, группы языков, шифр специальности, стиль автора, сферу науки, жанр, пол автора, поэзию или прозу, УДК и можно вычислить нужные элементы текста и, нажав кнопку «Ҳамроҳкунӣ», внести их в базу. На рисунке 6.5 представлена информация о поэме Абдурахмони Джоми «Лайлӣ ва Маҷнун».

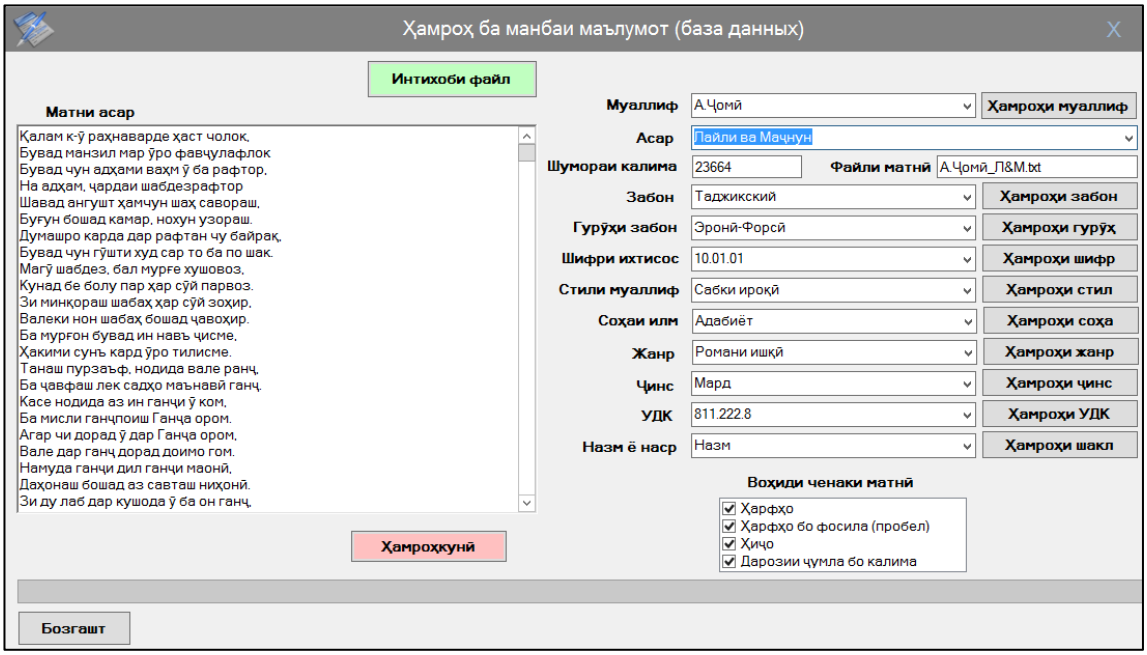


Рисунок 6.5. – Добавление произведения и ЦПТ в базу данных

При добавлении произведения и элементов текста первоначально проверяется их наличие в базе данных. Если они имеются, то программа запрашивает о необходимости их добавления. При согласии старые удаляются, их место замещается новыми.

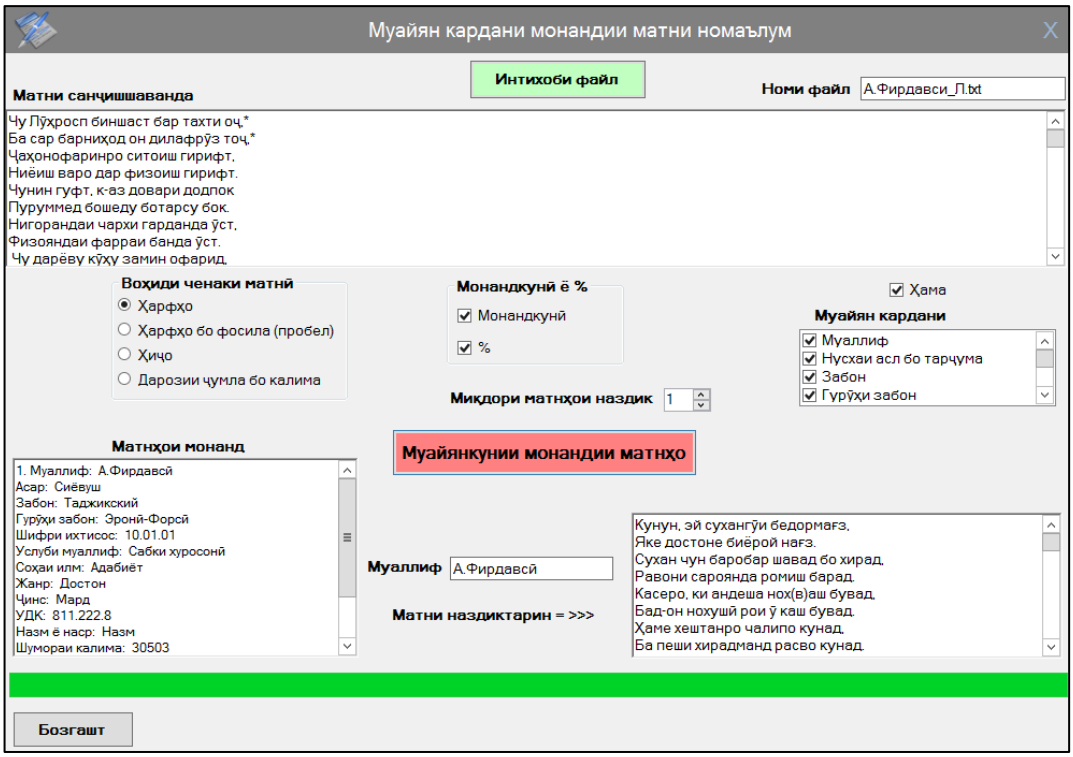


Рисунок 6.6. – Идентификация и определение однородности текста

Подсистема представления результатов (определение однородности по короткому-полному тексту), см. рисунок 6.6. Её основной функцией является

вывод результатов работы аналитического блока в понятной и наглядной для исследователя форме. В подсистеме можно установить тип проверки, количество ближайших текстов и т.д., для проверки произведения.

Затем, если выбраны какие-то конкретные пункты, то для обработки следует выбрать нужные нам элементы текста. После этого надо нажать на кнопку «Муайянкунии монандии матнҳо», внизу интерфейса программы показывается время сравнения.

К процессу алгоритма сравнения произведений авторов используется классификатор (1.1)-(1.7), он запрограммирован внутри программы.

И, наконец, итогом работы программы является определение однородности *T* текста. Если однородные произведения не найдены, то выдаётся «однородные произведения не найдены».

На проведенном примере видно, что при проверке произведения «Достони Подшоҳии Лӯҳросп» программа даёт результат: *T* текст принадлежит Абулкасиму Фирдоуси, он однороден с произведением «Достони Сиёвуш» автора, кроме этого язык – «Тоҷикӣ», группы языков – «Эронӣ-Форсӣ», шифр специальности – «10.01.01», стиль автора – «Сабки хуросонӣ», сфера науки – «Адабиёт», жанр – «Достон», пол автора – «Мард», поэзия или проза – «Назм» и УДК – «811.222.8».

Основным отличием данной программы от других является, во-первых, её полная завершённость (другие программы обычно являются демонстрационными), во-вторых, программа в большинстве случаев корректно распознаёт автора, язык, группы языков, шифр специальности, стиль автора, сферу науки, жанр, пол автора, поэзию или прозу, УДК и оригинал и перевод для полных текстов, даже для коротких, в-третьих, программа указывает пользователю оценку правильного распознавания, выраженную в процентах, в-четвёртых, программа полностью бесплатная.

§ 6.5. Контрольный пример для тестирование программного комплекса, вычисление τ, π и γ

Итак, мы имеем обучающую выборку $V = \bigcup_{k=1}^{n} V^{(k)}$, представленную в виде объединения некоторых непересекающихся подмножеств-классов $V^{(k)}$ с числом элементов $q^{(k)}$, $\sum_{k=1}^{n} q^{(k)} = Q$.

В нашем контрольном примере рассматриваем частный случай $n = 4$, то есть обучающая выборка V состоит из 4-х классов, причём в 3-х первых классах $V^{(k)}$ содержатся по 2 элемента ($q^{(k)} = 2,\ k = \overline{1,3}$), а в $V^{(4)}$ – три элемента, $q^{(4)} = 3$. Таким образом, общее число элементов будет равным $Q = \sum_{k=1}^{4} q^{(k)} = 9$, а число L всевозможных пар расстояний между ними, подсчитываемое по формуле § 1.4 п. 1.4.2, будет $L = Q(Q-1)/2 = 36$.

Теперь предположим, что таблица парных расстояний между элементами уже задана и имеет следующий вид (для простоты расстояниям приписаны

положительные целые значения):

Таблица 6.1. – Расстояния между элементами

	$V^{(11)}$	$V^{(12)}$	$V^{(21)}$	$V^{(22)}$	$V^{(31)}$	$V^{(32)}$	$V^{(41)}$	$V^{(42)}$	$V^{(43)}$
$V^{(11)}$									
$V^{(12)}$	1								
$V^{(21)}$	5	1							
$V^{(22)}$	4	5	2						
$V^{(31)}$	6	4	7	5					
$V^{(32)}$	7	4	3	3	4				
$V^{(41)}$	5	3	6	8	5	1			
$V^{(42)}$	7	5	6	7	2	3	2		
$V^{(43)}$	8	3	8	5	4	3	2	3	

В этой таблице в первой строке и первом столбце записаны обозначения номеров элементов, входящих в соответствующие классы $V^{(k)}$, Так, например, записи $V^{(32)}$ и $V^{(43)}$ обозначают 2-й элемент в классе $V^{(3)}$ и 3-й элемент в классе $V^{(4)}$. Следовательно, в ячейке таблицы на пересечении, например, столбца $V^{(31)}$ со строкой $V^{(42)}$ записано расстояние между 1-м элементом из класса $V^{(3)}$ и 2-м элементом из класса $V^{(4)}$.

Множество ячеек ниже главной диагонали разделены на две части: 6 ячеек жёлтого цвета - это расстояния между элементами из одних и тех же классов, и 30 непомеченных ячеек, характеризующих расстояния между элементами из разных классов. С двумя различными цветами связывается вполне приемлемая $\mathbb{H}$-гипотеза о том, что *элементы одного класса - «однородные», а разных классов - «неоднородные»*.

Пусть γ - некоторое положительное число. $\mathbb{H}$-гипотезе сопоставляется математическая модель в виде утверждения: *элементы* E_1, E_2 *называются* ***γ-однородными****, если*

$$\rho\,(E_1\,, E_2) \leq \gamma, \tag{6.1}$$

и ***γ- неоднородными****, если*

$$\rho\,(E_1\,, E_2) > \gamma. \tag{6.2}$$

Очевидно, что от значения γ зависит однородность или неоднородность любой пары элементов, следовательно, и степень выполнимости гипотезы. Однородность всех элементов одного класса в рамках математической модели означает справедливость неравенства (6.1), а неоднородность любых двух элементов разных классов – справедливость неравенства (6.2). Гипотеза $\mathbb{H}$ может нарушаться для каких-то пар элементов одного и того же класса в случае, когда вместо неравенства (6.1) имеет место неравенство (6.2), а также в случае, когда какие-то два элементов двух различных классов удовлетворяют неравенству (6.1) вместо того, чтобы выполнялось неравенство (6.2).

Пусть $\tau = \tau(\gamma)$ – суммарное количество нарушений гипотезы $\mathbb{H}$ одновременно в двух случаях: невыполнение неравенства «однородности» в случае двух элементов, принадлежащих одному классу, и невыполнение неравенства «неоднородности» в случае двух элементов, принадлежащих разным классам. Тогда для фиксированного γ *показатель выполнения гипотезы будем определять величиной* π, задаваемой формулой

$$\pi = 1 - \tau(\gamma)/L\,, \qquad (6.3)$$

где L – число взаимных расстояний между всеми парами элементов из обучающей выборки V. Из этой формулы следует, что π может принимать значения из отрезка [0, 1], причём $\pi = 0$, если $\tau = L$, и $\pi = 1$, если $\tau = 0$. В первом случае гипотезу $\mathbb{H}$ следует признать непригодной, а во втором – полностью согласованной с обучающей выборкой.

В связи с тем, что эффективность γ-классификатора зависит от значения параметра γ, представляет интерес найти такое его значение, при котором π принимает максимальное значение. *Именно в этом и заключается суть настройки γ-классификатора на данных обучающей выборки*. Если такая настройка будет приемлемой, то можно говорить о решении **задачи обучения γ-классификатора.**

Теперь перейдем к решению задачи. Оно, по существу, представляется алгоритмом вычисления $\tau = \tau(\gamma)$ для всевозможных значений γ из полуоси $\gamma \in (0, \infty)$. Существенную помощь в этом деле нам оказывают свойства функции $\tau(\gamma)$. По своему определению она – положительная, целочисленная и кусочно-гладкая, с конечным числом точек разрыва на полуоси $(0, \infty)$. Разрывы происходят в точках, координаты которых в нашем примере совпадают с L = 36 значениями парных расстояний между $Q = 9$ заданными элементами.

Действительно, обратимся к полуоси γ и на ней отметим все 36 значений парных расстояний. Нам будет удобно представить эту ситуацию в виде таблицы 6.2, которая по сути своей является переработкой таблицы 6.1, приспособленной к описанию алгоритма:

Таблица 6.2. – Частотность расстояний между элементами

	Значения парных расстояний (точек разрыва)								
Полуось γ	0	1	2	3	4	5	6	7	8
Частота пар расстояний между однородными элементами	0	1	3	1	1	0	0	0	0
Частота пар расстояний между неоднородными элементами	0	2	1	6	4	7	3	4	3
Общая суммарная частота	0	3	4	7	5	7	3	4	3

В этой таблице на пересечении полуоси γ со столбцом «Значения парных расстояний» цифрой 0 отмечено начало полуоси γ, последующие цифры от 1 до 8

указывают значения парных расстояний. Теперь, пользуясь таблицей, перейдем к вычислению числа $\tau = \tau(\gamma)$ нарушений $\mathbb{H}$-гипотезы. Для этих целей начнем изменять значения вещественного параметра γ пошагово, в начале в пределах интервала $(0, 1)$, затем в пределах полуинтервалов $[k, k + 1)$, k =1,…,7, и наконец, в $[8, \infty)$.

Для фиксированного γ всякий раз следует контролировать число пар расстояний, которые оказались левее, правее и равных γ, и в согласии с неравенствами (6.1) и (6.2) выявлять нарушения $\mathbb{H}$-гипотезы. Отметим сразу же, что при изменении γ строго внутри указанных интервала и полуинтервалов значение $\tau(\gamma)$ остаётся постоянным, равным значению в левом конце полуинтервала.

Итак, пусть $\gamma \in (0, 1)$. Тогда для всех без исключения парных расстояний таблицы 6.1 имеет место неравенство $\gamma < \rho\,(E_1\,, E_2)$. Но это значит, что $\tau(\gamma) = 6$ для $\gamma \in (0, 1)$. Действительно, для всех неоднородных пар элементов условие (6.2) выполняется, а вот условие *однородности* (6.1) для 6 пар элементов, которые по $\mathbb{H}$-гипотезе должны были быть однородными, не выполняется.

Теперь пусть $\gamma = 1$, то есть γ совпадает с первой точкой разрыва. Из таблицы 6.2 получаем, что для одной пары E_1 и E_2 *однородных* элементов (см. желтую 2-ю строку таблицы 6.2) и для двух каких-то определенных пар *неоднородных* элементов (см. 3-ю строку таблицы 6.2), действительно, имеют место равенства $\rho\,(E_1\,, E_2) = \gamma = 1$. Следовательно, одно условие, именно (6.1), выполняется (нарушение $\mathbb{H}$-гипотезы не происходит), а вот для 2-х *неоднородных* элементов условие (6.2) нарушается. Оба элемента вместо того, чтобы удовлетворять (6.2) удовлетворяют равенству (6.1). Итого, для всего полуинтервала $\gamma \in [1, 2)$ получим $\tau(\gamma)$ = 5 + 2 = 7. Результат вычисления τ, π и γ в программном продукте на основе примера показан на рисунке 6.7.

Предоставляя читателю самостоятельно произвести необходимые вычисления, приведем итоговые результаты для других полуинтервалов. В дальнейшем значение $\tau(\gamma)$ представляется в виде суммы двух слагаемых, из которых первое указывает на число нарушений $\mathbb{H}$-гипотезы для однородных элементов, а второе – для неоднородных элементов.

Если $\gamma \in [2, 3)$, то $\tau = 2 + 3 = 5$.

Если $\gamma \in [3, 4)$, то $\tau = 1 + 9 = 10$.

Если $\gamma \in [4, 5)$, то $\tau = 0 + 13 = 13$.

Если $\gamma \in [5, 6)$, то $\tau = 0 + 20 = 20$.

Если $\gamma \in [6, 7)$, то $\tau = 0 + 23 = 23$.

Если $\gamma \in [7, 8)$, то $\tau = 0 + 27 = 27$.

Если $\gamma \in [8, \infty)$, то $\tau = 0 + 30 = 30$.

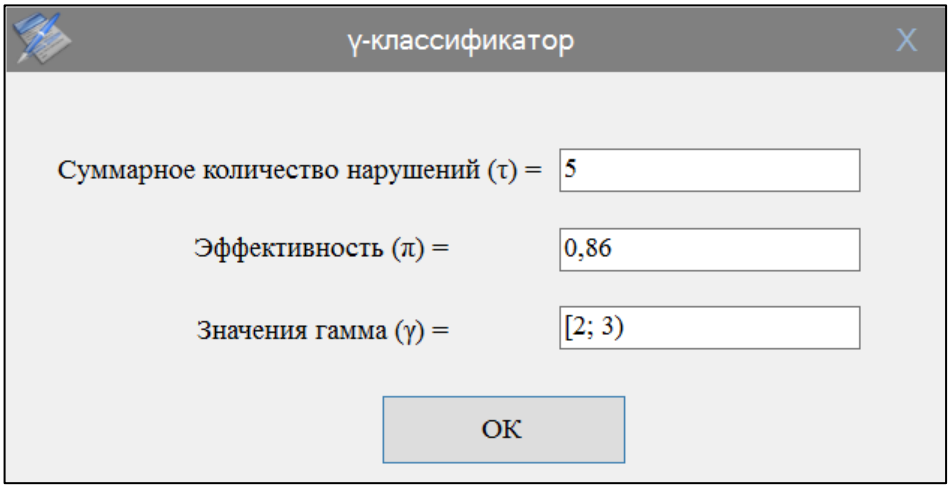


Рисунок 6.7. – Результат вычисления τ, π и γ в программном продукте

Минимальное значение $\tau = 5$ достигается на полуинтервале $\gamma \in [2, 3)$. Контрольный пример подсказывает новую (эквивалентную) запись смысла последовательных процедур алгоритма для определения минимального значения $\tau = \tau(\gamma)$, см. **1.4.3. Алгоритм настройки γ-классификатора** из § 1.4.

§ 6.6. Технические средства программного комплекса «THR»

Комплекс программ «THR» состоит из следующих компонентов:

– установочного пакета программы «THR»;

– базы данных;

– руководства пользователя.

Для обеспечения работоспособности программы «THR» предъявляется ряд системных требований:

к операционной системе –

установка программы «THR» возможна на компьютерах под управлением популярных операционных систем, таких как Microsoft Windows 98/ME/2000/XP/Vista/7/8/10.

к свободному пространству –

– жесткий диск должен иметь как минимум 89,1 Мб свободной памяти.

к процессору и оперативной памяти –

– оперативная память должна быть не менее 128 Мб,

– необходим процессор с тактовой частотой не ниже 500 МГц;

При увеличении мощности компьютера возрастает соответственно и производительность программы.

§ 6.7. Установка программного продукта

Программа совместима с операционными системами Windows 2000/XP/Vista/7/8/10.

Для работы приложения требуется установка СУБД MySQL 5.1, а также Mysql-ODBC версии 3.51.23.

1. Убедиться, что на компьютере установлены и настроены СУБД MySQL 5.1 и драйвер Mysql-ODBC версии 3.51.23. Установочные файлы этих программ можно скачать с официального сайта MySQL (www.mysql.com).

2. Распаковать файловый архив с программой в любой каталог. После распаковки каталог должен содержать следующие файлы и папки:

– **THR**.exe – исполняемый файл программы;

– THR.sql – сценарий восстановления базы данных, содержащий необходимые таблицы и хранимые процедуры для работы системы.

3. Выполнить сценарий THR.sql на сервере базы данных и настроить право доступа к базе данных.

ОБСУЖДЕНИЕ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

В диссертационной работе исследованы и решены научные проблемы определения однородности текстов на основе γ-классификатора.

Ранее лучшие результаты по этим задачам принадлежали английским исследователям Дж.Рудману, Т.Муколову, Д.Руззеллу, Б.Аллисону и русским математикам-программистам А.А.Шелупанову, Р.В.Мещерякову, А.С. Романову и А.В.Куртукову. В монографии [227] указано: известно очень много математических методов, применяемых для изучения проблемы. Из них выделяются два метода:

-машина опорных векторов;

-нейронные сети.

А также имеется 1000 количественных признаков для характеристики текста. Информативными признаками считаются буквенные униграммы, биграммы, триграммы, что для успешной идентификации текста необходимо 10000 слов для английского языка и 8000 слов для русского языка.

Мы рассмотрели различные варианты признаков однородности текстов:

- распознавание авторства, тематики текста, язык, группу языков, оригинал и его перевод, стиль произведений и шифры научных работ;

различное число текстов:

- 10, 20 и 40;

различные элементы текста:

- буквы алфавита естественного языка, буквенные *N*-граммы и слоги, знаки пунктуации, морфемы, словоформы, длины слов, предложений и абзацев (в символах и словах), анаграмм и др.;

различные классификаторы:

- нейронные сети, метод ближайшего соседа и γ-классификатор.

Этапы исследования и полученные результаты:

Этап 1. Первом этапе по научной теме диссертации использовалась и анализировалась существующая научная литература зарубежных стран, определялись другие этапы исследования.

Этап 2. На этом этапе проверялись различные методы определения автора текста, в ходе исследования был предложен новый метод, названный γ-классификатором. γ-классификатор по сравнению с современными методами дал высокую эффективность до 100% при определении автора текста.

Этап 3. В основе полученного метода, называемого γ-классификатором, были проверены различные единицы измерения текста, такие как буква, биграмма, триграмма, слово, длина предложения со словами, слогами и т.п. для определения автора текста, где наибольшую эффективность имела триграмма букв, которая давала от 95% до 100%.

Этап 4. В этом этапе мы обращались к модельной коллекции, составленной из трёх частей: произведений классиков таджикско-персидской литературы, произведений современных поэтов и произведений современных прозаиков. Каждая часть коллекции состоит из 10 произведений, по два произведения пяти авторов. Тестированы количественные признаки высокого уровня на предмет возможности их использования в качестве информативных признаков для распознавания автора на примере модельных коллекций художественных произведений таджикского языка, а также узбекского языка и в роли исследовательского аппарата применялись γ-классификатор З.Д. Усманова и метод ближайшего соседа. Наша цель заключалась не только в том, чтобы выявить различия в размерах и расположениях оптимальных полуинтервалов γ, но также и в определении числа нарушений гипотезы однородности, вычислении коэффициента эффективности распознавания авторов по их произведениям в целом и возможно минимальным фрагментам. Фрагменты извлекались из «начала», «середины» и «конца» произведения, «в пределах» которых бессистемно и случайным образом выбирались кусочки текста различных размеров. Путем применения метрического классификатора и метода ближайшего (по расстоянию) соседа удалось идентифицировать авторов убывающих по размерам последовательности текстовых фрагментов от величины в 7000 слов (40000 символов) вплоть до 20 слов (100 символов).

Этап 5. После того, как γ-классификатор З.Д. Усманова дал высокую эффективность для определения автора текста не только для таджикского, но и для разных языков, а также для фрагмента текстов, на этом этапе метод исследуются по другим вопросам, таким как определение тематики, языка, шифр специальности текста, авторского стиля, оригинала и его перевода и так далее. Для этих задач γ-классификатор З.Д. Усманова тоже дал высокую эффективность до 100% при использовании различных текстовых единиц измерения. Наконец, предложенная модель была протестирована с разными текстами, и почти все протестированные тексты дали близкое сходство с соответствующими текстами.

Этап 6. На предыдущих этапах применялся метод γ-классификатор на коллекции небольших текстов, но на этом этапе его исследовали на корпус текстов, что опять-таки дало высокий результат. На основании этих исследований в целом можно прийти к выводу, что метод γ-классификатора З.Д.Усманова, среди других известных методов, можно использовать для решения любой проблемы. Следует отметить, что для всех полученных результатов создан комплекс различных компьютерных программ.

Этап 7. На основе всех полученных общенаучных результатов создан комплекс программ для ЭВМ для определения сходства текстов, написанных на разных языках.

На наш взгляд, предложенная методология, обладающая определенной универсальностью, позволяет расширить круг решаемых задач как в теоретическом плане, так и оказать практическую помощь специалистам, занимающимся вопросами распознавания однородности текстов.

Комплекс программ под названием **«THR»** применён в следующих организациях:

1. Академия Министерства внутренних дел Республики Таджикистан.
2. Государственный комитет национальной безопасности Республики Таджикистан.
3. Институт языка и литературы имени Рудаки НАНТ.
4. Институт математики имени А.Джураева НАНТ.
5. ТТУ имени академика М.С. Осими (см. Приложение 1).

Построенный с широким использованием математических моделей и высокого уровня программирования комплекс, в частности, предназначен для развития таджикского языка с использованием возможностей информационных технологий.

Данный комплекс программы является важным как с точки зрения компьютерной лингвистики, так и с точки зрения литературоведения, и направлен на оказание практической помощи исследователям в области языка, литературы, математики и информационных технологий. Среди них призвано определить и распознать стиль каждого автора, особенности отдельных произведений разных авторов, частоту встречаемости букв, слогов, слов, словосочетаний, состав слов в отдельных произведениях, создание различных математических моделей.

В обозримом будущем предполагается развернуть исследования по приложению γ-классификатора к распознаванию код программ, формул, изображений, голоса и различные задачи, связанные с техникой, [68-А].

ЗАКЛЮЧЕНИЕ

Основные результаты диссертации:

1. Проанализированы имеющиеся в зарубежной научной литературе данные о количественных признаках текстов и алгоритмах, применяемых при распознавании однородности произведений. Определены перспективные направления исследований.

2. На расширенной коллекции произведений доказана эффективность применения γ-классификатора З.Д. Усманова для распознавания авторов полноценных произведений.

3. Установлена эффективность γ-классификатора, способного распознавать с точностью до 100% автора текстового фрагмента размером от величины в 7000 слов (40000 символов) вплоть до 20 слов (100 символов).

4. Установлена возможность существенного сокращения объёма вычислительных процедур за счёт использования не всех, а только высокоточных элементов ЦП текстов.

5. Установлена статистическая эффективность применения на основе распределения частотности различных алфавитных элементов текста и γ-классификатора (математической триады) для распознавания других признаков однородности, таких как тематики текста, язык, группа языков, оригинал и его перевод, стиль произведений и шифры научных работ.

6. Исследованы статистические закономерности распознавания авторов и языков произведений на корпусах художественных литературных произведений.

7. Путем применения метрического классификатора и методом ближайшего (по расстоянию) соседа удалось на тестируемых случайных выборках текстов идентифицировать с достаточно высокой точностью признаки однородности произведений различных модельных коллекций и корпусов.

8. Установлена эффективность применения γ-классификатора для атрибуции искусственно сгенерированных поэм «Шахнаме» А. Фирдоуси.

9. Исследованы особенности применения γ-классификатора при распознавании автора текста на примере модельной коллекции, количественные описания произведений которой основываются на различных вариантах упорядочения буквенных *N*-грамм (с учётом и без учёта пробелов).

10. Создан первый в Таджикистане объектно-ориентированный компьютерный программный комплекс для распознавания (идентификации) однородности текста на основе различных ЦП текста и γ-классификатора среди сколь угодно большого числа текстов, см. Приложение 1.

Рекомендации по практическому использованию результатов

Спроектированный комплекс рекомендуется для применения в автоматизации процесса обработки текстовой информации в государственной административной деятельности, для установления авторства анонимных текстов в сфере криминалистики, для обнаружения плагиата в курсовых и дипломных проектах, в представленных к защите кандидатских и докторских диссертациях в области образования и науки, а также для использования в изучении самых разнообразных научных проблем, связанных с вопросами распознавания «однородных» печатных текстов.

ЛИТЕРАТУРА

Список использованных источников:

1. Allison, B. Another Look at the Data Sparsity Problem [Text] / B. Allison, D. Guthrie, L. Guthrie // – 2006.

2. Bengio, Y. Learning Long-Term Dependencies with Gradient Descent is Difficult [Text] / Y. Bengio, P. Simard, P. Frasconi // IEEE Transactions on Neural Networks. – 1994.

3. Rudman, J. The state of authorship attribution studies: Some problems and solutions [Text] / J. Rudman // Computers and the Humanities. – 1998. – Vol. 31. – pp. 351-365.

4. Russell, D. Language-tree divergence times support the Anatolian theory of Indo-European origin [Text] / D. Russell, A. Gray, Q.D. Atkinson // Nature: журнал. – Великобритания: Nature Publishing Group, 2003. – Т.426. – №6965. – pp. 435-439.

5. Chang, W. «Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis (= Филогенетический анализ, связанный с предками, подтверждает гипотезу индоевропейских степей)» Language [Text] / W. Chang, Ch. Cathcart, D. Hall, A. Garrett // Volume 91. – Number 1. – March 2015. – pp. 194-244 (Article). – Published by Linguistic Society of America.

6. Kassian, A. Supplementary Information 2: Linguistics: Datasets; Methods; Results (в статье Kushniarevich A., Utevska O., Chuhryaeva M., Agdzhoyan A., Dibirova K., Uktveryte I. et al. (2015) Genetic Heritage of the Balto-Slavic Speaking Populations: A Synthesis of Autosomal, Mitochondrial and Y-Chromosomal Data [Text] / A. Kassian, A. Dybo // PLoS ONE 10(9): e0135820. https://doi.org/10.1371/journal.pone.0135820).

7. Calix, K. Stylometry for E-mail Author Identification and Authentication [Text] / K. Calix // [Electronic resource] Proceedings of CSIS Research Day, Pace University, May 2008. – URL: http://csis.pace.edu/~ctappert/srd2008/c2.pdf (дата обращения: 05.09.2023).

8. Hadi, W.M. A Comprehensive Comparative Study Using Vector Space Model with K-Nearest Neighbour on Text Categorization Data [Text] / W.M. Hadi // Asian Journal of Information Management. – 2008. – Vol. 2. – №1. – pp. 14-22.

9. Karr, J.R. Scientific Authorship, Collaboration, Interdisciplinarity, and Productivity [Text] / J.R. Karr, J.J. Hughey, T.K. Lee // [Electronic resource]. – 2008. – URL: http://covertlab.stanford.edu/projects/ScienceGenealogy (дата обращения: 14.02.2022).

10. Hochreiter, S. Long Short-Term Memory [Text] / S. Hochreiter, J. Schmidhuber // Neural Computation. – 1997. – № 8 (9). – С. 1735-1780.

11. Mikolov, T. Recurrent Neural Network based Language Model [Text] / T. Mikolov [и др.] // Proceedings of INTERSPEECH. – 2010.

12. Hochreiter, S. Untersuchungen zu dynamischen neuronalen Netzen [Text] / S. Hochreiter // Master's thesis, Institut für Informatik, Technische Universität, Munchen. – 1991.

13. Ioffe, S. Batch normalization: accelerating deep network training by reducing internal covariate shift [Text] / S. Ioffe, C. Szegedy // ICML'15 Proceedings of the 32nd International Conference on International Conference on Machine Learning. – Volume 37. – 2015. – №37. – pp. 448-456.

14. Abbasi, A. Identification and comparison of extremist-group Web forum messages using authorship analysis [Text] / A. Abbasi, H. Chen // IEEE Intelligent Systems. – 2005. – Vol. 20. – № 5. – pp. 67-75.

15. Abbasi, A. Visualizing Authorship for Identification [Text] / A. Abbasi, H. Chen // Proceedings of the 4th IEEE Symposium on Intelligence and Security Informatics. – 2006. – pp. 60-71.

16. Abbasi, A. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace [Text] / A. Abbasi, H. Chen // ACM Transactions on Information Systems. – NY : ACM. 2008. – Vol. 26. – № 2. Article 7. – 29 p.

17. Abbasi, A. Applying authorship analysis to extremist-group web forum messages [Text] / A. Abbasi, H. Chen // IEEE Intelligent Systems. – 2005. – Vol. 20. – № 6. – pp. 67-75.

18. Amasyah, M.F. Automatic Turkish Text Categorization in Terms of Author, Genre and Gender [Text] / M.F. Amasyah, B. Diri // NLDB 2006. – Berlin : Springer-Verlag. 2006. – Vol. LNCS 3999. – pp. 221-226.

19. Apte, C. Automated Learning of Decision Rules for Text Categorization [Text] / C. Apte, F. Damerau, S. Weiss // ACM Transactions on Information Systems. – NY : ACM. 1994. – Vol. 12. – Issue 3. – pp. 233-240.

20. Apte, C. Text mining with decision rules and decision trees [Text] // Proceedings of the Conference on Automated Learning and Discovery, June, 1998 / C. Apte, F. Damerau, S. Weiss. – 1998. [Электронный ресурс] – Режим доступа: http://www.research.ibm.com/dar/papers/pdf/cald98_with_cover.pdf (дата обращения: 10.04.2023).

21. Argamon, S. Measuring the usefulness of function words for authorship attribution [Text] / S. Argamon, S. Levitan // Proceedings of ACH / ALLC Conference. – 2005. [Электронный ресурс] – Режим доступа: http://mustard.tapor.uvic.ca/cocoon/ach_abstracts/xq/xhtml.xq?id=162 (дата обращения: 15.05.2023).

22. Argamon, S. Routing documents according to style [Text] / S. Argamon, M. Koppel, G. Avneri // Proceedings of the 1st International Workshop on Innovative Information. – 1998. [Электронный ресурс] – Режим доступа: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.52.688&rep=rep1&type=pdf

(дата обращения: 10.01.2021).

23. Argamon, S. Style mining of electronic messages for multiple authorship discrimination: first results [Text] / S. Argamon, M. Saric, S.S. Stein // Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – NY : ACM. 2003. – pp. 475-480.

24. Argamon, S. Stylistic text classification using functional lexical features [Text] / S. Argamon, C. Whitelaw, P. Chase et al. // Journal of the American Society of Information Science and Technology. – 2007. – Vol. 58. – №6. – pp. 802-822.

25. Baayen, R.H. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution [Text] / R.H. Baayen, H.V. Halteren, F.J. Tweedie // Literary and Linguistic Computing. – 1996. – Vol. 11. – pp. 121-131.

26. Baayen, R.H. An experiment in authorship attribution [Text] / R.H. Baayen, H.V. Halteren, A. Neijt et al. // Proceedings of JADT 2002. – Universit´e de Rennes, St. Malo. 2002. – pp. 29-37.

27. Burrows, J.F. «An ocean where each kind...»: Statistical analysis and some major determinants of literary style [Text] / J.F. Burrows // Computers and the Humanities. – 1989. – Vol. 23. – №4. – pp. 309-321.

28. Burrows, J.F. All the way through: Testing for authorship in different frequency data [Text] / J.F. Burrows // Literary and Linguistic Computing. – 2007. – Vol. 22. – №1. – pp. 27-47.

29. Chaski, C.E. Empirical evaluations of language-based author identification [Text] / C.E. Chaski // Forensic Linguistics. – 2001. – Vol. 8. – № 1. – pp. 1-65.

30. Chaski, C.E. Multilingual Forensic Author Identification through *N*-Gram Analysis [Text] / C.E. Chaski // Proceedings of the 8th Biennial Conference on Forensic Linguistics / Language and Law, July 2007, Seattle, WA. – 2007. [Электронный ресурс] – Режим доступа: http://www.allacademic.com/meta/p177064_index.html (дата обращения: 14.02.2022).

31. Chaski, C.E. Who's at the keyboard: Authorship attribution in digital evidence investigations [Text] / C.E. Chaski // International Journal of Digital Evidence. – 2005. – Vol. 4. – № 1. [Электронный ресурс] – Режим доступа: http://www.ijde.org (дата обращения: 09.05.2023).

32. Corney, M. Identifying the Authors of Suspect E-mail [Text] / M. Corney, A. Anderson, G. Mohay et al. // Computers and Security. – 2001. [Электронный ресурс] – Режим доступа: http://eprints.qut.edu.au/8021/1/CompSecurityPaper.pdf (дата обращения: 08.05.2023).

33. Corney, M. Gender-Preferential Text Mining of E-mail Discourse [Text] / M. Corney, O. de Vel, A. Anderson // Proceedings of 18th Annual Computer Security Applications Conference (ACSAC '02). – 2002. – p. 282.

34. De Vel, O. Mining e-mail content for author identification forensics [Text] / O.

De Vel, A. Anderson, M. Corney et al. // ACM SIGMOD. – NY : ACM. 2001. – Rec. 30. – № 4. – pp. 55-64.

35. Diederich, J. Authorship attribution with support vector machines [Text] / J. Diederich, J. Kindermann, E. Leopold // Applied Intelligence. – Springer Netherlands. 2003. – Vol. 19. – №1-2. – pp. 109-123.

36. Efron, B. Estimating the number of unseen species: How many words did Shakespeare know? [Text] / B. Efron, R. Thisted // Biometrika. – 1976. – Vol. 63. – № 3. – pp. 435-447.

37. Efron, B. Did Shakespeare write newly-discovered poem? [Text] / B. Efron, R. Thisted // Biometrika. – 1987. – Vol. 74. – № 3. – pp. 445-455.

38. Farringdon, J.M. Analyzing for Authorship [Text] / J.M. Farringdon with contributions by A.Q. Morton, M.G. Farringdon, M.D. Baker. – Cardiff : University of Wales Press. 1996. – 324 p.

39. Joachims, T. Text Categorization With Support Vector Machines: Learning With Many Relevant Features [Text] / T. Joachims // Proceedings of ECML-98, 10th European Conference on Machine Learning. – 1998. – № 1398. – pp. 137-142.

40. Juola, P. Cross-Entropy and Linguistic Typology [Text] / P. Juola // Proceedings of New Methods in Language Processing 3. – ACL. 1998. – pp. 141-149.

41. Juola, P. Measuring linguistic complexity: The morphological tier [Text] / P. Juola // Journal of Quantitative Linguistics. – 1998. – Vol. 5. – № 3. – pp. 206-213.

42. Juola, P. What can we do with small corpora? Document categorization via cross-entropy [Text] / P. Juola // Proceedings of an Interdisciplinary Workshop on Similarity and Categorization, Edinburgh, UK. – 1997. [Электронный ресурс] – Режим доступа: http://www.mathcs.duq.edu/~juola/papers.d/identification.ps (дата обращения: 14.02.2022).

43. Juola, P. A Controlled-Corpus Experiment in Authorship Identification by Cross-Entropy [Text] / P. Juola, H. Baayen // Literary and Linguistic Computing. – Oxford: Oxford University Press. 2005. – Vol. 20. – pp. 59-67.

44. Juola, P. A Prototype for Authorship Attribution Studies [Text] / P. Juola, J. Sofko, P. Brennan // Literary and Linguistic Computing. – 2006. – Vol. 21. – № 2. – pp. 169-178.

45. Kjell, B. Authorship attribution of text samples using neural networks and Bayesian classifiers [Text] / B. Kjell // IEEE International Conference on Systems, Man and Cybernetics, San Antonio. TX. – 1994.

46. Kjell, B. Authorship determination using letter pair frequencies with neural network classifiers [Text] / B. Kjell // Literary and Linguistic Computing. – 1994. – Vol. 9. – № 2. – pp. 119-124.

47. Kjell, B. Discrimination of authorship using visualization [Text] / B. Kjell, W.A. Woods, O. Frieder // Information Processing and Management. – 1994. – Vol. 30. –

№ 1. – pp. 141-150.

48. Koppel, M. Automatically categorizing written texts by author gender [Text] / M. Koppel, S. Argamon, A.R. Shimoni // Literary and Linguistic Computing. – 2002. – Vol. 17. – № 4. – pp. 401-412.

49. Koppel, M. Authorship verifcation as a one-class classification problem [Text] / M. Koppel, J. Schler // Proceedings of the 21st International Conference on Machine Learning. Banff, Canada. – NY : ACM Press. 2004. – pp. 489-495.

50. Koppel, M. Exploiting stylistic idiosyncrasies for authorship attribution [Text] / M. Koppel, J. Schler // Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, Acapulco, Mexico. – 2003. – pp. 69-72.

51. Lowe, D. Shakespeare vs. Fletcher: A stylometric analysis by Radial Basis Functions [Text] / D. Lowe, R. Matthews // Computers and the Humanities. – Springer Netherlands. 1995. – Vol. 29. – pp. 449-461.

52. Luyckx, K. Authorship Attribution and Verification with Many Authors and Limited Data [Text] / K. Luyckx, W. Daelemans // Proceedings of the 22nd International Conference on Computational Linguistics (COLING '08). – 2008. – pp. 513-520.

53. Matthews, R.A.J. Neural computation in stylometry I: An application to the works of Shakespeare and Fletcher [Text] / R.A.J. Matthews, T.V.N. Merriam // Literary and Linguistic Computing. – 1993. – Vol. 8. – № 4. – pp. 203-209.

54. Matthews, R.A.J. Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe [Text] / R.A.J. Matthews, T.V.N. Merriam // Literary and Linguistic Computing. – 1994. – Vol. 9. – № 1. – pp. 1-6.

55. Mendenhall, T.A. The characteristic curves of composition [Text] / T.A. Mendenhall // Science. – 1887. – № 11. – pp. 237-249.

56. Mendenhall, T.A. A mechanical solution to a literary problem [Text] / T.A. Mendenhall // Popular Science Monthly. – 1901. – № 60. – pp. 97-105.

57. Morton, A.Q. Literary Detection: How to Prove Authorship and Fraud In Literature and Documents [Text] / A.Q. Morton. – New York: Scribner's. 1978. – 221 p.

58. Morton, A.Q. The Authorship of Greek Prose [Text] / A.Q. Morton // Journal of the Royal Statistical Society (A). – 1965. – Series A. – № 128. – pp. 169-233.

59. Peng, F. Combining Naive Bayes and n-Gram Language Models for Text Classification [Text] / F. Peng, D. Schuurmans // Lecture Notes in Computer Science. – 2003. – Vol. 2633. – pp. 335-350.

60. Peng, F. Augumenting Naive Bayes Text Classifier with Statistical Language Models [Text] / F. Peng, D. Schuurmans, S. Wang // Information Retrieval. – 2004. – Vol. 7. – № 3-4. – pp. 317-345.

61. Peng, F. Language independent authorship attribution using character level language models [Text] / F. Peng, D. Schuurmans, S. Wang et al. // Proceedings of the

10th conference on European chapter of the ACL. – 2003. – Vol. 1. – pp. 267-274.

62. Peng, R.D. Quantitative analysis of literary styles [Text] / R.D. Peng, N.W. Hengartner // The American Statistician. – 2002. – Vol. 56. – № 3. – pp. 175-185.

63. Stamatatos, E. Author identification using imbalanced and limited training texts [Text] / E. Stamatatos // Porceedings of the 18th International Conference on Database and Expert Systems Applications. – 2007. – pp. 237-241.

64. Stamatatos, E. Computer-based authorship attribution without lexical measures [Text] / E. Stamatatos, N. Fakotakis, G. Kokkinakis // Computers and the Humanities. – 2001. – Vol. 35. – № 2. – pp. 193-214.

65. Teahan, W.J., Using compression-based language models for text categorization [Text] / W.J. Teahan, D.J. Harper, ed. J. Callan et al. // Workshop on Language Modeling and Information Retrieval, ARDA. – 2001. – pp. 83-88.

66. Teahan, W.J. A Compression-based Algorithm for Chinese Word Segmentation [Text] / W.J. Teahan, Y.Y. Wen, R. McNab et al. // Computational Linguistics. – 2000. – Vol. 26. – № 3. – pp. 375-393.

67. Tweedie, F.J. How Variable may a Constant be? Measures of Lexical Richness in Perspective [Text] / F.J. Tweedie, H. Baayen // Computers and the Humanities. – Springer. – 1998. – Vol. 32. – № 5. – pp. 323-352.

68. Tweedie, F.J. Neural network applications in stylometry: The Federalist Papers [Text] / F.J. Tweedie, S. Singh, D.I. Holmes // Computers and the Humanities. – 1996. – Vol. 30. – № 1. – pp. 1-10.

69. Waugh, S. Computational stylistics using Artificial Neural Networks [Text] / S. Waugh, A. Adams, F.J. Tweedie // Literary and Linguistic Computing. – 2000. – Vol. 15. – №2. – pp. 187-198.

70. Zheng, R. A framework for authorship analysis of online messages: Writing-style features and techniques [Text] / R. Zheng, J. Li, Z. Huang et al. // Journal of the American Society for Information Science and Technology. – 2006. – Vol. 57. – № 3. – pp. 378-393.

71. Ломакина, Л.С. Построение и исследование модели текста для его классификации по предметным категориям [Текст] / Л.С. Ломакина, А.В. Мордвинов, А.С. Суркова // Системы управления и информационные технологии. − Воронеж. − 2011. − №1(43). − С. 16-20.

72. Ломакин, Д.В. Золотая пропорция как инвариант структуры текста [Текст] / Д.В. Ломакин, А.З. Панкратова, А.С. Суркова // Журнал «Вестник Нижегородского университета им. Н.И. Лобачевского». − Н. Новгород. − 2011. − №4. − С. 196-199.

73. Ломакина, Л.С. Иерархическая кластеризация текстовых документов [Текст] / Л.С. Ломакина, В.Б. Родионов, А.С. Суркова // Системы управления и информационные технологии. − Воронеж. − 2012. − № 2(48). − С. 39-44.

74. Суркова, А.С. Построение модели и алгоритма кластеризации в интеллектуальном анализе данных [Текст] / А.С. Суркова, С.С. Буденков // Журнал «Вестник Нижегородского университета им. Н.И. Лобачевского». − Н. Новгород. − 2012. − №2(1). − С. 198-202.

75. Суркова, А.С. Алгоритм разбиения неструктурированного множества текстовых объектов [Текст] / А.С. Суркова, В.Б. Родионов // Научно-технический вестник Поволжья. − Казань. – 2013 г. − №5. – С .298-300

76. Суркова, А.С. Идентификация текстов на основе информационных портретов [Текст] / А.С. Суркова // Вестник Нижегородского университета им. Н.И. Лобачевского. − Н. Новгород. − 2014. − № 3 (1). − С. 145-149

77. Ломакина, Л.С. Кластеризация текстовых данных на основе нечеткой логики [Текст] / Л.С. Ломакина, А.С. Суркова, С.С. Буденков // Системы управления и информационные технологии. – Воронеж. − №1(55). – 2014. – С. 73-77.

78. Суркова, А.С. Анализ и моделирование текстовых данных в задачах обеспечения кибербезопасности [Текст] / А.С. Суркова // Системы управления и информационные технологии. − Воронеж. − №3.1(61). − 2015. – С. 178-182

79. Семенцов, М.С. Энтропийные характеристики символьного разнообразия в текстах исходных кодов программ [Текст] / М.С. Семенцов, А.С. Суркова // Системы управления и информационные технологии. − Воронеж. − №1.1(59). − 2015. – С. 173-176.

80. Ломакина, Л.С. Теоретические аспекты концептуального анализа и моделирования текстовых структур [Текст] / Л.С. Ломакина, А.С. Суркова // Фундаментальные исследования. − Москва: Издательский Дом «Академия Естествознания». − 2015. − № 2 (часть 17). − С. 3713-3717.

81. Ломакина, Л.С. Методологические аспекты концептуального анализа и моделирования текстовых структур [Текст] / Л.С. Ломакина, А.С. Суркова // Фундаментальные исследования. − Москва: Издательский Дом «Академия Естествознания». − 2015. − № 6 (часть 3). − С. 497-501.

82. Ломакина, Л.С. Прикладные аспекты концептуального анализа и моделирования текстовых структур [Текст] / Л.С. Ломакина, А.С. Суркова // Фундаментальные исследования. − Москва: Издательский Дом «Академия Естествознания». − 2015. − № 7 (часть 3). − С. 540-544.

83. Ломакин, Д.В. Методология формирования системоорганизующих характеристик текстовых данных [Текст] / Д.В. Ломакин, М.Д. Ломакина, А.С. Суркова // Фундаментальные исследования. − Москва: Издательский Дом «Академия Естествознания». − 2015. − № 11 (часть 3). − С. 480-483.

84. Суркова, А.С. Применение нейронных сетей для определения авторства текстов исходных кодов программ [Текст] / А.С. Суркова, А.А. Царев // Системы

управления и информационные технологии. − Воронеж. − №1(63). − 2016. – С. 78-82.

85. Суркова, А.С. Моделирование текстов на основе энтропийных характеристик в задачах классификации [Текст] / А.С. Суркова, С.С. Скорынин // Вестник ВГАВТ. − Н. Новгород. − 2016. − №4. – С. 54-61.

86. Surkova, A.S. Hierarchical Clustering of Text Documents [Text] / L.S. Lomakina, V.B. Rodionov, A.S. Surkova // Automation and Remote Control, 2014. – Vol. 75. – No. 7. – pp. 1309-1315.

87. Ломакина, Л.С. Информационные технологии анализа и моделирования текстовых данных: Монография / Л.С. Ломакина, А.С. Суркова – Воронеж: Издательство «Научная книга». – 2015. – 208 с.

88. Ломакина, Л.С. Автоматизированные информационно-поисковые системы. Задачи. Принципы. Методология: учеб. пособие / Л.С. Ломакина, А.С. Суркова // Нижегор. гос. техн. ун-т им. Р.Е. Алексеева. – Н. Новгород. – 2011. – 109 с.

89. Ломакина, Л.С. Системный подход в лингвистических исследованиях [Текст] / Л.С. Ломакина, А.С. Суркова // Материалы 6-ой международной конференции «НТИ-2002. Информационное общество. Интеллектуальная обработка информации. Информационные технологии» - М.: Изд-во ВИНИТИ, 2002. – С. 224-225.

90. Панкратова, А.З. Структурно-статистические методы обработки текста [Текст] / А.З. Панкратова, А.С. Суркова // Проблемы прикладной лингвистики: Сборник статей Международной научно-практической конференции. – Пенза. – 2004. – С. 246-248.

91. Ломакина, Л.С. Идентификация автора и языка текста на основании использования его структурно-вероятностных закономерностей [Текст] / Л.С. Ломакина, А.С. Суркова // Системы обработки информации и управления. Тр. НГТУ. – Т. 57, вып. 13. – Н.Новгород. – 2006. – С. 97-101.

92. Ломакина, Л.С. Статистические возможности идентификации текста [Текст] / Л.С. Ломакина, А.З. Панкратова, А.С. Суркова // Интеллектуальные системы INTELS’2008: Труды Восьмого международного симпозиума // под ред. К.А. Пупкова. – М.: РУСАКИ. – 2008. – С. 234-238

93. Ломакина, Л.С. Развитие методов автоматизированной классификации текста по тематическим категориям [Текст] / Л.С. Ломакина, А.В. Мордвинов, А.С. Суркова // Нижегородский государственный технический университет им. Р.Е. Алексеева. – Нижний Новгород, 2011. – 61 с.

94. Родионов, В.Б. Иерархическая кластеризация текстовых документов с использованием алгоритмов сжатия [Текст] / В.Б. Родионов, А.С. Суркова // Современные проблемы информатизации в моделировании и социальных

технологиях. – Сб. трудов. Вып. 17. – Воронеж: Издательство «Научная книга». – 2012. – С. 214-216.

95. Ломакина, Л.С. Классификация и кластеризация в интеллектуальном анализе данных на основе принципа сжатия [Текст] / Л.С. Ломакина, А.С. Суркова // Интеллектуальные системы: Труды Десятого международного симпозиума / Под ред. К.А. Пупкова. – М.: РУСАКИ. – 2012. – С. 291-295.

96. Ломакина, Л.С. Моделирование текстов в задачах кластеризации, классификации, идентификации [Электронный ресурс] / Л.С. Ломакина, А.С. Суркова, В.Б. Родионов, С.С. Буденков // Материалы 8-ой международной конференции «НТИ-2012. Актуальные проблемы информационного обеспечения науки, аналитической и инновационной деятельности» – М., ВИНИТИ. – 2012. – С. 118-120. [Электронный ресурс] – Режим доступа: http://www.viniti.ru/download/russian/konf2012.pdf (дата обращения: 14.02.2022).

97. Суркова, А.С. Иерархическая кластеризация текстовых объектов на основе принципов сжатия и нечеткой логики (научный руководитель Л.С. Ломакина) / А.С. Суркова, В.Б. Родионов, С.С. Буденков и др. // Отчет по НИР No государственной регистрации 02201364023 от 19.12.2013. – Н. Новгород: НГТУ. – 121 с.

98. Ломакина, Л.С. Основные принципы интеллектуального анализа текстовых данных [Текст] / Л.С. Ломакина, А.С. Суркова // Труды Конгресса по интеллектуальным системам и информационным технологиям «IS&IT'14». – М.: Физматлит. – 2014. – С. 311-318.

99. Суркова, А.С. Концептуальные модели текстовых структур в интеллектуальном анализе данных [Текст] / А.С. Суркова // Системный анализ в проектировании и управлении: Сб. науч. тр. XVIII Междунар. науч.-прак. конф. Ч.1. – СПб.: Изд-во Политехн. ун-та. – 2014. – С. 154-156.

100. Ломакина, Л.С. Использование оценки Колмогоровской сложности для анализа текстовых структур [Текст] / Л.С. Ломакина, В.Б. Родионов, А.С. Суркова // Системный анализ в проектировании и управлении: Сб. науч. тр. XVIII Междунар. науч.-прак. конф. Ч.1. – СПб.: Изд-во Политехн. ун-та. – 2014. – С. 156-158.

101. Ломакина, Л.С. Методы интеллектуального анализа данных на основе принципов нечеткой логики [Текст] / Л.С. Ломакина, А.С. Суркова // Интеллектуальные системы: труды одиннадцатого международного симпозиума (INTELS'2014) / Под ред. К.А. Пупкова. – М.: РУДН. – 2014. – С. 229-233.

102. Ломакина, Л.С. Концептуальный анализ, принципы моделирования и оптимизация алгоритмов синтеза текстовых структур [Текст] / Л.С. Ломакина, А.С. Суркова // Современные методы прикладной математики, теории управления и компьютерных технологий: сб. тр. VIII междунар. конф. «ПМТУКТ-2015». –

Воронеж. – 2015. – С. 208-210.

103. Ломакина, Л.С. Моделирование и технологии обработки текстов с целью их идентификации [Текст] / Л.С. Ломакина, А.С. Суркова // Труды X Международной конференции «Идентификация систем и задачи управления» SICPRO ‘15. – Москва. – 26-29 января 2015. – С. 1202-1210.

104. Ломакина, Л.С. Анализ и моделирование в прикладных задачах обработки текстов [Текст] / Л.С. Ломакина, А.С. Суркова // Сборник трудов конференции Международна научна школа "ПАРАДИГМА". – ЛЯТО-2015 Варна. – 20-23 августа 2015. – С. 90-95.

105. Суркова, А.С. Модификация моделей текста на основе выбора подстрок различной длины в задачах классификации [Текст] / А.С. Суркова, С.С. Скорынин // Материалы XXII Междунар. научно-технической конференции «Информационные системы и технологии» ИСТ-2016. – С. 351-352.

106. Суркова, А.С. Реализация алгоритмов LDA И LSA для обработки текстов на естественном языке [Текст] / А.С. Суркова, И.Д. Чернобаев, Е.В. Цыбульская // Материалы XXII Междунар. научно-технической конференции «Информационные системы и технологии» ИСТ-2016. – С. 353-354.

107. Суркова, А.С. Классификация текстовых данных на основе энтропийных характеристик символьного разнообразия [Текст] / А.С. Суркова, С.С. Скорынин // Сборник научных трудов XX Международной научно-практической конференции «Системный анализ в проектировании и управлении». – 2016. – С. 321-327.

108. Surkova, A.S. Neural networks and decision trees algorithms – the base of automated text classification and clustering [Text] / A.S. Surkova, A.A. Domnin, I.V. Bulatov, A.A. Tsarev // American Journal of Control Systems and Information Technology. Science Book Publishing House. – LLC. – 2013. – №2. – pp. 33-35.

109. Surkova, A.S. Modified classification algorithm with fuzzy interpretation of clusters [Text] / A.S. Surkova, S.S. Skorynin // American Journal of Control Systems and Information Technology. – Vol. 4. – No. 2. – 2014. – pp. 27-30.

110. Surkova, A.S. Automated text classification and clustering using neural networks and decision trees algorithms [Text] / A.S. Surkova, A.A. Domnin, I.V. Bulatov, A.A. Tsarev // Modern informatization problems: Proceedings of the XIX-th International Open Science Conference. – 2014. – pp.141-144.

111. Surkova, A.S. Algorithm of text classification based on the principles of fuzzy logic [Text] / A.S. Surkova, S.S. Skorynin // Modern informatization problems in simulation and social technologies: Proceedings of the XX-th International Open Science Conference. – 2015. – pp. 213-217.

112. Surkova, A.S. Comparative analysis fuzzy algorithms of text classification [Text]/ A.S. Surkova, S.S. Skorynin, I.D. Chernobaev // Modern informatization

problems in simulation and social technologies: Proceedings of the XXI-th International Open Science Conference. – 2016. – pp. 213-218.

113. Ломакина, Л.С. Визуализация структурно-статистических методов идентификации текстов / Л.С. Ломакина, А.С. Суркова // Свидетельство об официальной регистрации программы для ЭВМ № 2005611470. Зарегистрировано в реестре программ для ЭВМ Федеральной службы по интеллектуальной собственности, патентам и товарным знакам от 17.06.2005

114. Ломакина, Л.С. Система иерархической кластеризации текстов / Л.С. Ломакина, А.С. Суркова, В.Б. Родионов // Свидетельство о государственной регистрации программы для ЭВМ №2013611004, зарегистрировано в Реестре программ для ЭВМ Федеральной службы по интеллектуальной собственности от 09.01.2013.

115. Суркова, А.С. Многоуровневый нечеткий кластеризатор потоковых текстовых данных / А.С. Суркова, А.А. Домнин // Свидетельство о государственной регистрации программы для ЭВМ №2015662510, зарегистрировано в Реестре программ для ЭВМ Федеральной службы по интеллектуальной собственности от 26.11.2015.

116. Суркова, А.С. Нейросетевой классификатор исходных кодов программ / А.С. Суркова, А.А. Царев // Свидетельство о государственной регистрации программы для ЭВМ №2016612679, зарегистрировано в Реестре программ для ЭВМ Федеральной службы по интеллектуальной собственности от 11.01.2016.

117. Ломакина, Л.С. Система идентификации авторства исходных кодов программ / Л.С. Ломакина, М.С. Семенцов, А.С. Суркова// Свидетельство о государственной регистрации программы для ЭВМ №2016612838, зарегистрировано в Реестре программ для ЭВМ Федеральной службы по интеллектуальной собственности от 12.01.2016.

118. Большакова, Е.И. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика [Текст] / Е.И. Большакова, Э.С. Клышинский, Д.В. Ландэ, А.А. Носков, О.В. Пескова, Е.В. Ягунова // Учеб. Пособие. М.: МИЭМ. – 2011. – 272 с.

119. Александров, В.В. Алгоритмы и программы структурного метода обработки данных [Текст] / В.В. Александров, Н.Д. Горский. – Л., «Наука». – 1983. – 208 с.

120. Апресян, Ю.Д. Лингвистическое обеспечение системы ЭТАП-2 [Текст] / Ю.Д. Апресян, И.М. Богуславский, Л.Л. Иомдин и др. – М.: Наука. – 1989.

121. Арапов, М.В. Квантитативная лингвистика [Текст] / М.В. Арапов. – М.: Наука. – 1988. – 184 с.

122. Арский, Ю.М. Принципы конструирования интеллектуальных систем [Текст] / Ю.М. Арский, В.К. Финн // Информационные технологии и

вычислительные системы. – №4. – 2008. – С. 4-37.

123. Барсегян, А.А. Методы и модели анализа данных: OLAP и Data Mining [Текст] / А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод. – СПб.: БХВ-Петербург. – 2004. – 336 с.

124. Батура, Т.В. Формальные методы установления авторства текстов и их реализация в программных продуктах [Текст] / Т.В. Батура // Программные продукты и системы. – 2013. – №4. – С. 286-295.

125. Башмаков, А.И. Интеллектуальные информационные технологии [Текст] / А.И. Башмаков // Учеб. пос. – М.: Изд-во МГТУ им. Н.Э. Баумана. – 2005. – 304 с.

126. Белов, В.С. Информационно-аналитические системы [Текст] / В.С. Белов // Основы проектирования и применения. – М. – 2005. – 111 с.

127. Белоногов, Г.Г. Компьютерная лингвистика и перспективные информационные технологии [Текст] / Г.Г. Белоногов. – М.: Русский мир. – 2004. – 248 с.

128. Белоногов, Г.Г. Проблемы автоматической смысловой обработки текстовой информации [Текст] / Г.Г. Белоногов, Р.С. Гиляревский, А.А. Хорошилов // Научно-техническая информация. – Сер. 2. – 2012. – № 11. – С. 24-28.

129. Белоногов, Г.Г. Метод аналогии в компьютерной лингвистике [Текст] / Г.Г. Белоногов, Ю.Г. Зеленков, А.П. Новоселов, Ал-др.А. Хорошилов, Ал-сей.А. Хорошилов // Научно-техническая информация. – Сер. 2. – 2000. – № 1. – С. 21-30.

130. Белоногов, Г.Г. Языковые средства автоматизированных информационных систем [Текст] / Г.Г. Белоногов, Б.А. Кузнецов. – М. – 1983.

131. Болдин, М.Б. Знаковый статистический анализ линейных моделей [Текст] / М.Б. Болдин, Г.И. Симонова, Ю.Н. Тюрин. – М.: Наука. Физматлит. – 1997. – 288 с.

132. Большаков, А.А. Методы обработки многомерных данных и временных рядов [Текст] / А.А. Большаков, Р.Н. Каримов // Учебное пособие для вузов. – М. – 2007. – 522 с.

133. Боровков, А.А. Математическая статистика: оценка параметров, проверка гипотез [Текст] / А.А. Боровков. – 2007. – 472 с.

134. Бурбаки, Н. Архитектура математики [Текст] / Н. Бурбаки // Математическое просвещение. – Вып. 5. – 1960. – С. 99-112

135. Быстров, И.И. Основы применения онтологии и компьютерной лингвистики при проектировании перспективных автоматизированных информационных систем [Текст] / И.И. Быстров, Б.В. Тарасов, А.А. Хорошилов, С.И. Радоманов // Системы и средства информатики. – 2015. – Том 25. – Выпуск

4. – С. 128-149.

136. Вапник, В.Н. Теория распознавания образов (статистические проблемы обучения) [Текст] / В.Н. Вапник, А.Я. Червоненкис. – М.: Наука, 1974. – 416 стр.

137. Верещагин, Н.К. Колмогоровская сложность и алгоритмическая случайность [Текст] / Н.К. Верещагин, В.А. Успенский, А. Шень. – М.: МЦНМО, 2013.

138. Волкова, В.Н. Теория систем [Текст] / В.Н. Волкова, А.А. Денисов // Учеб. пос. – М. – 2006. – 511 с.

139. Гаврилова, Т.А. Базы знаний интеллектуальных систем [Текст] / Т.А. Гаврилова, В.Ф. Хорошевский // Учеб. пособие. – СПб.: Питер. – 2000.

140. Гальперин, И.Р. Текст как объект лингвистического исследования [Текст] / И.Р. Гальперин. – М.: Наука, 1981. – 140 с.

141. Гладких, А.В. Синтаксические структуры естественного языка в автоматизированных системах сообщений [Текст] / А.В. Гладких. – М.: Наука, 1985.

142. Городецкий, Б.Ю. Компьютерная лингвистика: моделирование языкового общения [Текст] / Б.Ю. Городецкий // Новое в зарубежной лингвистике. – Вып. 24. – М., 1989.

143. Дмитриев, А.С. Хаос, фракталы и информация [Текст] / А.С. Дмитриев // Наука и жизнь. – №5. – 2001.

144. Дюран, Б. Кластерный анализ [Текст] / Б. Дюран, П. Одел // М., Статистика. – 1977. – 128 с.

145. Егорушкин, А. У каждого свой язык [Текст] / А. Егорушкин // Компьютера. – №21. – 2002.

146. Еремеев, А.П. Построение решающих функций на базе тернарной логики в системах принятия решений в условиях неопределенности [Текст] / А.П. Еремеев // Изв. РАН. Теория и системы управления. – 1997. – №5. – С. 138-143.

147. Загоруйко, Н.Г. Методы распознавания и их применение [Текст] / Н.Г. Загоруйко. – М., 2012. – 211 с.

148. Заде, Л.А. Понятие лингвистической переменной и его применение к принятию приближенных решений [Текст] / Л.А. Заде. – М.: Мир, – 1976. – 164 с.

149. Заде, Л.А. Размытые множества и их применение в распознавании образов и кластер-анализе [Текст] / Л.А. Заде // Классификация и кластер. – М.: Мир, – 1980. – С. 208-247.

150. Иомдин, Л.Л. Автоматическая обработка текста на естественном языке: модель согласования [Текст] / Л.Л. Иомдин. – М.: Наука, 1990.

151. Кендалл, М. Многомерный статистический анализ и временные ряды [Текст] / М. Кендалл, А. Стьюарт. – М., Наука. – 1976. – 736 с.

152. Кирдин, А.Н. Скрытые параметры и транспонированная регрессия

[Текст] / А.Н. Кирдин, А.Ю. Новоходько, В.Г. Царегородцев // Нейроинформатика – Новосибирск: Наука. – Сибирское предприятие РАН. – 1998. – 296 с.

153. Кобзарь, А.И. Прикладная математическая статистика [Текст] / А.И. Кобзарь. – М.: Физматлит, 2006. – 816 с.

154. Колмогоров, А.Н. Три подхода к определению понятия «Количество информации» [Текст] / А.Н. Колмогоров // Новое в жизни, науке, технике. Сер. "Математика, кибернетика". – 1991. – №1. – С. 24-29.

155. Костышин, А.С. О применимости некоторых формальных методов для исследования полных строений текстов [Текст] / А.С. Костышин // Материалы конференции КВАЛИСЕМ-2000. – Новосибирск. – Изд-во Новосибир. гос. пед. ун-та. – 2000.

156. Кофман, А. Введение в теорию нечетких множеств [Текст] / А. Кофман. – М., Радо и связь. – 1982. – 432 с.

157. Курейчик, В.М. Обработка информации на основе онтологий [Текст] / В.М. Курейчик // Труды конгресса "IS&IT15" по интеллектуальным системам и информационным технологиям. – Изд-во ЮФУ. – 2015 г. – Т.2. – С. 63-75.

158. Кучуганов, В.Н. Анализ многозначностей в естественно-языковых текстах [Текст] / В.Н. Кучуганов // Десятая национальная конференция по искусственному интеллекту с международным участием КИИ-2006. – Труды конференции. – В 3-т. – М: Физматлит. – 2006. [Электронный ресурс] – Режим доступа: http://www.raai.org/resurs/papers/kii-2006/ (дата обращения: 14.02.2022).

159. Кучуганов, В.Н. Вербализация реальности и виртуальности. Ассоциативная семантика [Текст] / В.Н. Кучуганов // Искусственный интеллект и принятие решений. – 2011. – № 1. – С. 55-66.

160. Кучуганов, В.Н., Элементы теории ассоциативной семантики [Текст] / В.Н. Кучуганов // Управление большими системами. – 2012. – Выпуск 40. – С. 30-48.

161. Ландэ, Д.В. Интернетика. Навигация в сложных сетях: модели и алгоритмы [Текст] / Д.В. Ландэ, А.А. Снарский, И.В. Безсуднов. – М.: Либроком. – 2009. – 264 с.

162. Леман, Э. Проверка статистических гипотез [Текст] / Э. Леман. – М.: Наука. – Главная редакция физико-математической литературы. – 1979. – 408 с.

163. Леоненков, А.В. Нечеткое моделирование в среде Matlab и fuzzyTECH [Текст] / А.В. Леоненков // С.Пб.: BHV-Санкт-Петербург. – 2003. – 736 с.

164. Леонтьева, Н.Н. Автоматическое понимание текста: системы, модели, ресурсы: учебное пособие [Текст] / Н.Н. Леонтьева. – М.: Издательский центр «Академия». – 2006.

165. Леонтьева, Н.Н. Динамика единиц в семантических структурах [Текст] / Н.Н. Леонтьева // Труды Международного семинара Диалог-2002 по

компьютерной лингвистике и ее приложениям. – Том 1. – Теоретические проблемы. – М. – 2002.

166. Лукашевич, Н.В. Тезаурусы в задачах информационного поиска [Текст] / Н.В. Лукашевич. – М.: Изд-во Московского университета. – 2011.

167. Маевский, Д.А. Определение авторства программного обеспечения по исходному коду программ [Текст] / Д.А. Маевский, Ю.П. Чербаджи // Радиоэлектронные и компьютерные системы. – 2014. – № 6. – С. 64-68.

168. Мартыненко, Г.Я. Основы стилеметрии [Текст] / Г.Я. Мартыненко. – Л.: Изд-во ЛГУ. – 1988. – 176 с.

169. Мельничук, А.С. Понятие системы и структуры языка [Текст] / А.С. Мельничук // Вопросы языкознания. – 1970. – №1. – С. 27.

170. Москальчук, Г.Г. Структура текста как синергетический процесс [Текст] / Г.Г. Москальчук. – М.: УРСС. – 2003. – 296 с.

171. Мурзин, Л.Н. Текст и его восприятие [Текст] / Л.Н. Мурзин, А.С. Штерн // Свердловск. – 1991. – 172 с.

172. Напреенко, Г.В. Идентификация текста по его авторской принадлежности на лексическом уровне (формально-количественная модель) [Текст] / Г.В. Напреенко // Вестник Томского гос. университета. – 2014. – №379. – С.17-23.

173. Нариньяни, А.С. Автоматическое понимание текста – новая перспектива [Текст] / А.С. Нариньяни // Труды международного семинара Диалог-97 по компьютерной лингвистике и ее приложениям. – Москва. – 1997. – С. 203-208.

174. Негуляев, В.А. Исследование коммуникативных микроструктур патентного текста и их роли для автоматической обработки информации [Текст] / В.А. Негуляев // Вычислительная лингвистика. – М.: Наука. – 1976.

175. Орлов, А.А. Технические аспекты создания автоматизированных информационных систем многоцелевого применения [Текст] / А.А. Орлов, А.А. Тельных, Е.А. Степанов, А.Д. Сорокин, Ю.Е. Аксенова // Наукоемкие технологии в космических исследованиях Земли. – 2013. – Т.5. – № 4. – С. 40-44.

176. Орлов, А.И. Анализ нечисловой информации в социологических исследованиях [Текст] / А.И. Орлов. – М.: Наука. – 1985. – 224 с.

177. Пескова, О.В. Методы автоматической классификации текстовых электронных документов [Текст] / О.В. Пескова // Научно-техническая информация. – Сер. 2. – 2006. – №3. – С. 13-20.

178. Пескова, О.В. Методы автоматической классификации электронных текстовых документов без обучения [Текст] / О.В. Пескова // Научно-техническая информация. – Сер. 2. – 2006. – № 12. – С. 21-32.

179. Поликарпов, А.А. Циклические процессы в становлении лексической системы языка: моделирование и эксперимент [Текст] / А.А. Поликарпов. – М. –

2001.

180. Пономаренко, И.Н. Фрактал в структуре художественного текста [Текст] / И.Н. Пономаренко // Русский язык: исторические судьбы и современность. II Международный конгресс русистов-исследователей. – М. – 2004.

181. Прангишвили, И.В. Системный подход и общесистемные закономерности [Текст] / И.В. Прангишвили. – М.: СИНТЕГ. – 2000. – 528 с.

182. Пруцков, А.В. Интернет-приложение метода обработки количественных числительных естественных языков [Текст] / А.В. Пруцков, Д.М. Цыбулько // Вестник Рязанского государственного радиотехнического университета. – №3 (Выпуск 41). – 2012. – С. 70-74.

183. Рассел, С. Искусственный интеллект: современный подход [Текст] / С. Рассел, П. Норвиг // 2-е изд. – М.: Вильямс. – 2006. – 1408 с.

184. Руспини, Э.Г. Последние достижения в нечетком кластер-анализе [Текст] / Э.Г. Руспини // Нечеткие множества и теория возможностей: Последние достижения. Под ред. Р.Р. Ягера. – М: Радио и связь. – 1986. – С. 114-132.

185. Рутковская, Д. Нейронные сети, генетические алгоритмы и нечеткие системы [Текст] / Д. Рутковская, М. Пилиньский, Л. Рутковский. – М.: Горячая линия – Телеком. – 2006. – 452 с.

186. Рыжов, А.П. Элементы теории нечетких множеств и измерения нечеткости [Текст] / А.П. Рыжов. – М., Диалог-МГУ. – 1998.

187. Сафронова, Ю.Б. Некоторые системно-количественные характеристики лексико-семантических парадигм разных видов [Текст] / Ю.Б. Сафронова // Уч. зап. ТГУ. – Вып. 745. – 1986. – С. 129-138.

188. Севбо, И.П. Графическое представление синтаксических структур и стилистическая диагностика [Текст] / И.П. Севбо. – Киев: Наук. Думка. – 1981. – 192 с.

189. Сегаран, Т. Программируем коллективный разум [Текст] / Т. Сегаран. – М., Символ-Плюс. – 2008. – 368 с.

190. Скороходько, Э.Ф. Семантические сети и автоматическая обработка текста [Текст] / Э.Ф. Скороходько – Киев, Наукова думка. – 1983. – 218 с.

191. Сметанин, Ю.Г. Мера символьного разнообразия – характеристика временных рядов [Текст] / Ю.Г. Сметанин, М.В. Ульянов // Materials of the III International Scientific Conference «Information-Management Systems and Technologies». – Odessa. – 2014. – С. 19-21.

192. Сметанин, Ю.Г. Мера символьного разнообразия: подход комбинаторики слов к определению обобщенных характеристик временных рядов [Текст] / Ю.Г. Сметанин, М.В. Ульянов // Бизнес-информатика. – №3 (29) . – 2014. – С. 40-48.

193. Сметанин, Ю.Г. Энтропийный подход к построению меры символьного

разнообразия слов и его применение к кластеризации геномов растений [Текст] / Ю.Г. Сметанин, М.В. Ульянов, А.С. Пестова // Математическая биология и биоинформатика. – 2016. – Т.11. – №1. – С. 114-126.

194. Солганик, Г.Я. Стилистика текста [Текст] / Г.Я. Солганик // Учебное пособие. – М.: Флинта: Наука. – 1997. – 256 с.

195. Солнцев, В.М. Язык как системно-структурное образование [Текст] / В.М. Солнцев. – М.: Наука. – 1971.

196. Турыгина, Л.А. Моделирование языковых структур средствами вычислительной техники [Текст] / Л.А. Турыгина. – М., 1988.

197. Финн, В.К. Об интеллектуальном анализе данных [Текст] / В.К. Финн // Новости искусственного интеллекта. – 2004. – №3. – С. 3-18.

198. Харкевич, А.А. О ценности информации [Текст] / А.А. Харкевич // Проблемы кибернетики: сб. – Вып 4. – М.: Физматгиз. – 1960. – С. 33-41.

199. Хартли, Р. Передача информации [Текст] / Р. Хартли // Теория информации и ее приложения. – М. – 1959. – С. 5-35.

200. Хорошевский, В.Ф. Пространства знаний в сети Интернет и Semantic Web (Часть 1) [Текст] / В.Ф. Хорошевский // Искусственный интеллект и принятие решений. – 2008. – № 1. – С. 80-97.

201. Хорошилов, А.А. Системы обнаружения плагиата нового поколения, базирующиеся на методах концептуального анализа текстов и использовании предметно ориентированных концептуальных словарей [Текст] / А.А. Хорошилов // Информатизация и связь. – 2013. – №3. – С. 112-118.

202. Хьетсо, Г. Кто написал «Тихий Дон»? (Проблема авторства «Тихого Дона») [Текст] / Г. Хьетсо и др.– М. – 1989.

203. Цвиркун, А.Д. Структура сложных систем [Текст] / А.Д. Цвиркун. – М. – 1975. – 200 с.

204. Цыпкин, Я.З. Основы теории обучающихся систем [Текст] / Я.З. Цыпкин. – М. Наука. – 1970. – 252 с.

205. Чекунов, И.Г. Киберпреступность: понятие и классификация [Текст] / И.Г. Чекунов // Российский следователь. – 2012. – №2. – С. 37-43.

206. Чекунов, И.Г. Современные киберугрозы. Уголовно-правовая и криминологическая квалификация киберпреступлений [Текст] / И.Г. Чекунов // Право и кибербезопасность. – 2012. – №1. – С. 9-23.

207. Шрейдер, Ю.А. Системы и модели [Текст] / Ю.А. Шрейдер, А.А. Шаров. – М.: Радио и связь. – 1982. – 152 с.

208. Штовба, С.Д. Введение в теорию нечетких множеств и нечеткую логику [Текст] / С.Д. Штовба // [Электронный ресурс] – Режим доступа: http://matlab.exponenta.ru/fuzzylogic/book1/index.php (дата обращения: 14.02.2022).

209. Комиссаров, А.Ю. Криминалистическое исследование письменной речи с

использованием ЭВМ [Текст]: дис. … канд. юрид. наук: 12.00.09 / А.Ю. Комиссаров. – М., 2001. – 225 с.

210. Кузьмина, Н.Н. Информационно-коммуникационная система региона [Текст] / Н.Н. Кузьмина, Н.А. Попов, А.А. Шелупанов. Томск: Изд-во В-Спектр. – 2010. – 220 с.

211. Марков, А.А. Пример статистического исследования над текстом «Евгения Онегина», иллюстрирующий связь испытаний в цепь [Текст] / А.А. Марков // Известия Имп. Акад. наук. – 1913. – Серия VI. Т. X. – №3. – С. 153.

212. Марков, А.А. Об одном применении статистического метода [Текст] / А.А. Марков // Известия Имп. Акад. наук. – 1916. – Серия VI. – Том X. – №4. – 239 с.

213. Марусенко, М.А. Атрибуция анонимных и псевдонимных литературных произведений методами теории распознавания образов [Текст] / М.А. Марусенко. – Л.: ЛГУ. – 1990. – 164 с.

214. Мещеряков, Р.В. Специальные вопросы информационной безопасности [Текст] / Р.В. Мещеряков, А.А. Шелупанов. Томск: Изд-во Института Оптики Атмосферы ТНЦ СО РАН. – 2003. – 250 с.

215. Мещеряков, Р.В. Основы информационной безопасности [Текст] / Р.В. Мещеряков, А.А. Шелупанов, Е.Б. Белов, В.П. Лось. М.: Горячая линия – Телеком. – 2006. – 544 с.

216. Морозов, Н.А. Лингвистические спектры: средство для отличения плагиатов от истинных произведений того или иного известного автора. Стилеметрический этюд [Текст] / Н.А. Морозов // Известия отд. русского языка и словесности Имп. Акад. наук. – 1915. – Том XX. – Кн. 4. – С. 93-127.

217. Павлов, Ю.Н. Оценка устойчивости во времени частотных словарей авторов в задачах идентификации текстов [Текст] / Ю.Н. Павлов, Е.А. Тихомирова // Наука и Образование. – Декабрь 2011. № 12. – С. 10. [Электронный ресурс] – Режим доступа: http://www.technomag.edu.ru/doc/274006.html (дата обращения: 14.02.2022).

218. Поддубный, В.В. Сравнительный анализ стилей текстов по частотным признакам на основе гипергеометрического критерия [Текст] / В.В. Поддубный, О.Г. Шевелев // Информационные технологии и математическое моделирование: Матер. III Всерос. науч.-практ. конф. (11-12 декабря 2004 г.). – Ч. 2. – Томск: Изд-во Том. ун-та. – 2004. – С. 48-51.

219. Поддубный, В.В. Сравнение стилей текстовых произведений по частному признаку на основе гипергеометрического критерия [Текст] / В.В. Поддубный, О.Г. Шевелев // Теоретическая и прикладная информатика / Под ред. проф. А.Ф. Терпугова. – Томск: Изд-во Том. ун-та. – 2004. – Вып. 1. – С. 101-109.

220. Поддубный, В.В. Сравнение качества подходов к кластеризации текстов

на основе гипергеометрического критерия [Текст] / В.В. Поддубный, О.Г. Шевелев, Д.А. Бормашов // Вестник Том. гос. ун-та. – 2006. – № 293. – С. 120-125.

221. Поддубный, В.В. Классификация текстов по авторству с помощью метода Хмелева и его модификаций [Текст] / В.В. Поддубный, О.Г. Шевелев // Научное творчество молодежи. Матер. X Всерос. науч.-практ. конф. – Ч.1. – 2006. – С. 175-177.

222. Поддубный, В.В. Образует ли последовательность символов текста простую цепь Маркова? [Текст] / В.В. Поддубный, О.Г. Шевелев // Информационные технологии и математическое моделирование (ИТММ-2005): Матер. IV Всерос. науч.-практ. конф. (18-19 ноября 2005 г.). – Ч. 2. – Томск: Изд-во Том. ун-та. – 2005. – С. 14-16.

223. Рогов, А.А. Компьютерная обработка текстов при помощи ИС «СМАЛТ» [Текст] / А.А. Рогов, Ю.В. Сидоров, А.В. Король, А.И. Солопова // Проблемы развития гуманитарной науки на Северо-западе России: опыт, традиции, инновации: Матер. науч. конф. – Петрозаводск. – 2004. – Том 1. – С. 122-124.

224. Романов, А.С. Анализ характеристик текста для целей выявления плагиата [Текст] / А.С. Романов // IX Всероссийская научная конференция «Техническая кибернетика, радиоэлектроника и системы управления»: Тезисы докладов. – Таганрог : Изд-во ТТИ ЮФУ. – 2008. – С. 126-127.

225. Романов, А.С. Модель базы данных для хранения текстов и их характеристик [Текст] / А.С. Романов // Доклады Том. гос. ун-та систем управления и радиоэлектроники. – 2008. – № 1(17). – С. 70-73.

226. Романов, А.С. Методика и программный комплекс для идентификации автора неизвестного текста [Текст] / А.С. Романов // Автореферат. Томск. – 2010. – 26 с.

227. Романов, А.С. Разработка и исследование математических моделей, методик и программных средств информационных процессов при идентификации автора текста [Текст] / А.С. Романов, А.А. Шелупанов, Р.В. Мещеряков. -В-Спектр, Томск. – 2011. – 188 с.

228. Сидоров, Ю.В. Математическая и информационная поддержка методов обработки литературных текстов на основе формально-грамматических параметров [Текст]: автореф. дис. канд. тех. наук: 05.13.18 / Ю.В. Сидоров. – СПб. – 2002. – 19 с.

229. Сидоров, Ю.В. Компьютерная автоматизированная система для лингвистического разбора литературных текстов [Текст] / Ю.В. Сидоров, А.А. Леонтьев, А.А. Рогов, В.Н. Захаров // IV-я Санкт-Петербургская Ассамблея молодых ученых и специалистов. Тезисы докладов. – СПб. – 1999. – С. 66.

230. Фоменко, В.П. Авторский инвариант русских литературных текстов

[Текст] / В.П. Фоменко, Т.Г. Фоменко // Новая хронология Греции: Античность в средневековье. – М. : Изд-во МГУб. – 1996. – Том 2. – С. 768-820.

231. Хьетсо, Г. Принадлежность Достоевскому: к вопросу об атрибуции Ф.М. Достоевскому анонимных статей в журналах «Время» и «Эпоха» [Текст] / Г. Хьетсо. – Oslo: Solum Forlag A.S. – 1986. – 86 с.

232. Хмелев, Д.В. Классификация и разметка текстов с использованием методов сжатия данных. Краткое введение [Электронный ресурс] [Текст] / Д.В. Хмелев. – Дата обновления: 11.03.2003. [Электронный ресурс] – Режим доступа: http://compression.graphicon.ru/download/articles/classif/intro.html (дата обращения: 14.02.2022).

233. Хмелев, Д.В. Распознавание автора текста с использованием цепей А.А. Маркова [Текст] / Д.В. Хмелев // Вестник МГУ, Сер. 9. Филология. – 2000. – № 2. – С. 115-126.

234. Шевелев, О.Г. Методы автоматической классификации текстов на естественном языке [Текст]: учебное пособие / О.Г. Шевелев. – Томск: ТМЛ-Пресс. – 2007. – 144 с.

235. Шевелев, О.Г. Разработка и исследование алгоритмов сравнения стилей текстовых произведений [Текст]: дис. ... канд. техн. наук: 05.13.18 / О.Г. Шевелев. – Томск. – 2006. – 176 с.

236. Шевелев, О.Г. Классификация текстов с помощью деревьев решений и сетей прямого распространения [Текст] / О.Г. Шевелев, А.В. Петраков // Вестник Том. гос. ун-та. – 2006. – № 290. – С. 300-307.

237. Романов, А.С. Идентификация авторства коротких текстов методами машинного обучения [Текст] / А.С. Романов, Р.В. Мещеряков // Компьютерная лингвистика и интеллектуальные технологии: матер. ежегод. междунар. конф. «Диалог» (Бекасово, 26-30 мая 2010 г.). М.: Изд-во РГГУ. – 2010. – Вып. 9 (16). – С. 407-413.

238. Романов, А.С. Определение пола автора короткого электронного сообщения [Текст] / А.С. Романов, Р.В. Мещеряков // Компьютерная лингвистика и интеллектуальные технологии: матер.ежегод. междунар. конф. «Диалог» (Бекасово, 25-29 мая 2011 г.). М.: Изд-во РГГУ. – 2011. – Вып. 10 (17). – С. 620-626.

239. Дягилев, В.В. Архитектура сервиса определения плагиата, исключающая возможность нарушения авторских прав [Текст] / В.В. Дягилев, А.А. Цхай, С.В. Бутаков // Вестник НГУ. Сер.: Информационные технологии. – 2011. – Том 9. вып. 3. – С. 23-29.

240. Романов, А.С. Модификация метода накопительных сумм для проверки однородности текста и выявления плагиата [Текст] / А.С. Романов // Электронные средства и системы управления: матер. докл. IX Междунар. науч.-практ. конф.

(30-31 октября 2013 г.): в 2 ч. – Ч. 2. – Томск: В-Спектр. – 2013. – С. 30-38.

241. Шумская, А.О. Выбор параметров для идентификации искусственно созданных текстов [Текст] / А.О. Шумская // Доклады Том. гос. ун-та систем управления и радиоэлектроники. – 2013. – № 2 (28). – С. 126-128.

242. Седов, А.В. Анализ неоднородностей в тексте на основе последовательностей частей речи [Текст] / А.В. Седов, А.А. Рогов // Современные проблемы науки и образования. – 2013. – № 1. [Электронный ресурс] – Режим доступа: www.science-education.ru/107-8339 (дата обращения: 14.02.2022).

243. Романов, А.С. Методика проверки однородности текста и выявления плагиата на основе метода опорных векторов и фильтра быстрой корреляции [Текст] / А.С. Романов, З.И. Резанова, Р.В. Мещеряков // Доклады томского государственного университета систем управления и радиоэлектроники. Томск: Издательство Томского государственного университета систем. – 2014. – № 2 (32). – С. 264-269.

244. Резанова, З.И. Задачи авторской атрибуции текста в аспекте гендерной принадлежности (к проблеме междисциплинарного взаимодействия лингвистики и информатики) [Текст] / З.И. Резанова, А.С. Романов, Р.В. Мещеряков // Вестник Томского государственного университета. – 2013. – № 370. – С. 24-28.

245. Резанова, З.И. О выборе признаков текста, релевантных в автороведческой экспертной деятельности [Текст] / З.И. Резанова, А.С. Романов, Р.В. Мещеряков // Вестник Томского государственного университета. Филология. – 2013. – № 6 (26). – С. 38-52.

246. Романов, А.С. Модификация метода накопительных сумм для проверки однородности текста и выявления плагиата [Текст] / А.С. Романов // Электронные средства и системы управления. Томск. – 2013. – Том 2. – С. 30-38.

247. Шарапова, Е.В. Универсальная система проверки текстов на плагиат «Автор.net» [Текст] / Е.В. Шарапова, Р.В. Шарапов // Информатика и её применение. – 2012. – Том 6. – вып. 3. – С. 52-58.

248. Воронцов, К.В. Математические методы обучения по прецедентам [Текст] / К.В. Воронцов // [Электронный ресурс] – Режим доступа: http://www.ccas.ru/voron (дата обращения: 14.02.2022).

249. Дьяконов, А.Г. Анализ данных, обучение по прецедентам, логические игры, системы WEKA, RapidMiner и MatLab (Практикум на ЭВМ кафедры математических методов прогнозирования) [Текст] / А.Г. Дьяконов // Учебное пособие. – М.: Издательский отдел факультета ВМК МГУ имени М.В. Ломоносова, 2010. – 278 с.

250. Проблема авторства текстов М.А. Шолохова. Материал из Википедии [Текст] // [Электронный ресурс] – Режим доступа: https://ru.wikipedia.org/wiki/Проблема_авторства_текстов_М.А._Шолохова –

свободной энциклопедии (дата обращения: 14.02.2022).

251. Макаров, А.Г. Цветок-Татарник. В поисках автора «Тихого Дона»: от Михаила Шолохова к Федору Крюкову [Текст] / А.Г. Макаров, С.Э. Макарова. – М.: «АИРО–XX», 2001. – 504 с.

252. Гмурман, В.Е. Теория вероятностей и математическая статистика [Текст] / В.Е. Гмурман. – Москва: «Высшая школа». – 2005. – 480 с.

253. Пушкина, А.С. Лаборатория общей компьютерной лексикологии и лексикографии. КИИСа Корпусная информационно-исследовательская система Электронная энциклопедия языка А.С. Пушкина (1-я очередь): стихи и драмы Пушкина [Текст] / А.С. Пушкина. [М., – 2010] [Электронный ресурс] – Режим доступа: http://www.philol.msu.ru/-lex/kiisa.html (дата обращения: 14.02.2022).

254. Большев, Л.Н. Таблицы математической статистики [Текст] / Л.Н. Большев, Н.В. Смирнов. – Москва: Наука, Гл. ред. физ-мат. литературы. – 1983. – 416 с.

255. Усманов, З.Д. Проблема раскладки символов на компьютерной клавиатуре [Текст] / З.Д. Усманов, О.М. Солиев // – Душанбе: Ирфон, – 2010. – 104 с.

256. Каримов, А.А. О цифровом портрете текстовой информации [Текст] / А.А. Каримов // Политехнический вестник, 2019. – 1 (45). – Серия: интеллект, инновации, инвестиции. – С. 7-10.

257. Каюмов, М.М. О цифровом портрете текстовой информации, основанном на частотности знаков пунктуации [Текст] / М.М. Каюмов // Политехнический вестник, 2019. – 1 (45). – Серия: интеллект, инновации, инвестиции. – С. 20-23.

258. Усманов, З.Д. Об одном цифровом портрете текста и его приложении [Текст] / З.Д. Усманов // Политехнический вестник, 2019. – 3 (47). – Серия: интеллект, инновации, инвестиции.

259. Усманов, З.Д. Классификатор дискретных случайных величин [Текст] / З.Д. Усманов // Доклады Академии наук Республики Таджикистан, 2017. – Т.60. – № 7-8. – С. 291-300.

260. Усманов, З.Д. Алгоритм настройки кластеризатора дискретных случайных величин [Текст] / З.Д. Усманов // Доклады Академии наук Республики Таджикистан, 2017. – Т.60. – № 9. – С. 392-397.

261. Каюмов, М.М. О распознавании автора текста на основе частотности $\alpha\beta$-кодов словоформ [Текст] / М.М. Каюмов // Политехнический вестник, Серия: интеллект, инновации, инвестиции. – 2020. – 2(50). – С. 29-36.

262. Ашурова, Ш.Н. Оценка эффективности использования словесных биграмм при идентификации текста [Текст] / Ш.Н. Ашурова // Материалы международной научно-практической конференции ТУТ «Роль ИКТ в

инновационном развитии экономики Республики Таджикистан». – Душанбе: Бахманруд, 2017. – С. 292-297.

263. Ашурова, Ш.Н. Оценка эффективности использования словесных триграмм при идентификации текста [Текст] / Ш.Н. Ашурова // Вестник Технологического университета Таджикистана. – 2017. – № 4 (31). – С. 51-58.

264. Ашурова, Ш.Н. О распознавании автора текста на основе частотности словесных биграмм [Текст] / Ш.Н. Ашурова, Х.А. Тошхуджаев // Политехнический вестник, Серия: интеллект, инновации, инвестиции. – 2020. – 2(50). – С. 57-61.

265. Бахтеев, К.С. О применимости укороченных цифровых портретов для идентификации автора текста [Текст] / К.С. Бахтеев // Политехнический вестник, Серия Интеллект, Инновация, Инвестиция. – 2020. – 2(50). – С. 25-28.

266. Бахтеев, К.С. О распознавании авторства по усечённым цифровым портретам текста [Текст] / К.С. Бахтеев // Известия АН Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2018. – № 4(173). – С. 82-92.

267. Усманов, З.Д. *N*-граммы в распознавании однородных текстов [Текст] / З.Д. Усманов // Материалы 20 научно-практического семинара «Новые информационные технологии в автоматизированных системах». – Москва, 2017. – С. 52-54.

268. O‘zbek adabiyoti [Текст] // [Электронный ресурс] – Режим доступа: http://kutubxona.com/wiki/index.php?page=Bosh_sahifa (дата обращения: 14.02.2022).

269. Косимов, О.А. Идентификация авторов экономика политической произведений с помощью символьных униграмм [Текст] / О.А. Косимов // Всероссийская научно-практическая конференция «Состояние и перспективы развития ИТ-образования», Чувашская Республика. – 2019.

270. Солиев, О.М. Муайянкунии муаллифи асарҳои сиёсӣ-иқтисодӣ бо воситаи биграммаҳои рамзӣ [Текст] / О.М. Солиев, О.А. Косимов // Вестник ПИТТУ имени академика М.С. Осими «Научно-технический журнал». – Худжанд, 2019. – №1 (10). – С. 19-26.

271. Гамма-классификатор. Материал из Википедии [Текст] // [Электронный ресурс] – Режим доступа: https://ru.wikipedia.org/wiki/Гамма-классификатор – свободной энциклопедии (дата обращения: 14.02.2022).

272. Список букв кириллицы. Материал из Википедии [Текст] // [Электронный ресурс] – Режим доступа: https://ru.wikipedia.org/wiki/Список_букв_кириллицы – свободной энциклопедии (дата обращения: 14.02.2022).

273. Список латинских букв. Материал из Википедии [Текст] //

[Электронный ресурс] – Режим доступа: https://ru.wikipedia.org/wiki/Список_латинских_букв – свободной энциклопедии (дата обращения: 14.02.2022).

274. Генеалогическое дерево славянских языков. Материал из Википедии [Текст] // [Электронный ресурс] – Режим доступа: http://900igr.net/kartinka/biologija/tema-proiskhozhdenie-jazykov-127785/5.-genealogicheskoe-drevo-slavjanskikh-jazykov-2.html – свободной энциклопедии (дата обращения: 14.02.2022).

275. Усманов, З.Д. Автоматический поиск и статистические закономерности множества анаграмм [Текст] / З.Д. Усманов // Издательство «Дониш». – 2020. – 81с.

276. Усманов, З.Д. Оценка эффективности применения γ-классификатора для атрибуции печатного текста [Текст] / З.Д. Усманов // ДАН РТ. – 2020. – Т.63. – № 3-4. – С.172-179.

277. Фирдавсӣ, А. – Шоҳнома [Текст] / А. Фирдавсӣ. – Душанбе: Адиб. – 2007/2008/2009/2010. – Ҷилд 1-10. – 4736 с.

278. Усманов, З.Д. Об одном обобщении формулы золотого сечения [Текст] / З.Д. Усманов // Доклады Академии наук Республики Таджикистан. – 2014. – Т.57. – № 1. – С. 5-8.

279. Фирдоуси, А. Шахнаме [Текст] / А. Фирдоуси. – М.: Издательство «Академии наук СССР», «Наука». – 1957/1960/1965/1969/1984/1989. – Том I-VI. 3472 с.

280. Мухсинзода, М.Ё. Генерация новых национальных таджикских имён с помощью искусственных нейронных сетей [Текст] / М.Ё. Мухсинзода, О.М. Солиев // Политехнический Вестник. Серия: Интеллект. Инновации. Инвестиции. – 2019. – № 4 (48). – С. 18-23.

281. Унитарный код. Материал из Википедии [Текст] // [Электронный ресурс] – Режим доступа: https://ru.wikipedia.org/wiki/Унитарный_код – свободной энциклопедии (дата обращения: 14.02.2022).

282. Искусственно сгенерированная поэма Шахнаме [Текст] // [Электронный ресурс] – Режим доступа: https://mirzodaler.page.link/shohnoma-5002 (дата обращения: 14.02.2022).

283. **Косимов, А.А.** Разработка основ автоматической системы распознавания автора незнакомого текста (на примере художественных произведений на таджикском языке) [Текст] / **А.А. Косимов** // дис. … канд. тех. наук. ТТУ имени академика М.С. Осими. – Душанбе, 2018. – 107 с.

284. **Косимов, А.А.** Разработка основ автоматической системы распознавания автора незнакомого текста (на примере художественных произведений на таджикском языке) [Текст] / **А.А. Косимов** // автореф. дис. …

канд. тех. наук. ТТУ имени академика М.С. Осими. – Душанбе, 2018. – 20 с.

285. **Косимов, А.А.** Коркарди асосҳои системаи автоматии муайянкунии муаллифи матни номаълум (дар мисоли асарҳои бадеӣ бо забони тоҷикӣ) [Матн] / **А.А. Косимов** // автореф. дис. … канд. тех. наук. ТТУ имени академика М.С. Осими. – Душанбе, 2018. – 21 с.

286. Усманов, З.Д. Особенности применения γ-классификатора для распознавания однородных объектов [Текст] / З.Д. Усманов // Вестник Филиала Московского государственного университета имени М.В. Ломоносова в городе Душанбе. – 2021. – № 1 (17). – С. 20-22.

287. Усманов, З.Д. Обзор результатов по применению γ-классификатора [Текст] / З.Д. Усманов // Известия Национальной академии наук Таджикистана. Отделение физико-математических, химических, геологических и технических наук. – 2021. – № 3 (184). – С. 62-73.

288. Хисрав, Н. Саодатнома (Мунтахаби осор, Куллиёт, Ҷилди 1) [Матн] / Н. Хисрав. – Душанбе: Ирфон. – 1991. – С. 551-568.

289. Хисрав, Н. Аз маснавиҳо (Гулшани Адаб, Ҷилди 1) [Матн] / Н. Хисрав. – Душанбе: Ирфон. – 1975. – С. 168-175.

290. Турсунзода, М. Мунтахаби осор [Матн] / М. Турсунзода. – Душанбе. – 2011. – 145 с.

291. Каноат, М. Масъуднома [Матн] / М. Каноат. [Электронный ресурс] – Режим доступа: http://www.cit.tj/mumin (дата обращения: 08.05.2023).

292. Каноат, М. Сурӯши якум [Матн] / М. Каноат. [Электронный ресурс] – Режим доступа: http://www.cit.tj/mumin (дата обращения: 08.05.2023).

293. Шерали, Л. Куллиёт, Ҷилди 1 [Матн] / Л. Шерали. – Душанбе: Адиб. – 2008. – 564 с.

294. Айнӣ, С. Аҳмади Девбанд (Куллиёт) [Матн] / С. Айнӣ. – Душанбе. – 1963. – С. 5-36.

295. Айнӣ, С. Ғуломон [Матн] / С. Айнӣ. – Сталинобод: Нашриёти давлатии Тоҷикистон. – 1950. – 493 с.

296. Икроми, Дж. Ман гунаҳкорам (Асарҳои мунтахаб, Ҷилди 1) [Матн] / Дж. Икроми. – Душанбе: Адиб. – 1987. – С. 161-348.

297. Хуҷандӣ, К. Девон [Матн]. / К. Хуҷандӣ. – Хуҷанд: Андеша. – 2011. – 413 с.

298. Румӣ, Ҷ. Маснавии Маънавӣ (Дафтари Аввал) [Матн] / Ҷ. Румӣ. – Душанбе. – 2015. – 233 с.

299. Румӣ, Ҷ. Маснавии Маънавӣ (Дафтари Дуввум) [Матн] / Ҷ. Румӣ. – Душанбе. – 2015. – 216 с.

300. Айнӣ, С. Одина (Асарҳои мунтахаб) [Матн] / С. Айнӣ. – Сталинобод: Нашриёти давлатии Тоҷикистон. – 1949. – С. 277-422.

301. Айнӣ, С. Шеърҳо. Қисми 1-2-3 [Матн] / С. Айнӣ. [Электронный ресурс] – Режим доступа: http://misol.tj/2017/03/29/Садриддин_Айни (дата обращения: 08.05.2023).

302. Турсун, С. Нисфирӯзӣ [Матн] / С. Турсун. – Душанбе. – 1973. – 25 с.

303. Хайём, У. Рубоиёт [Матн] / У. Хайём. – Душанбе: Дониш. – 1983. – 100 с.

304. **Косимов, А.А.** О множестве анаграмм в произведениях Л. Шерали [Текст] / **А.А. Косимов**, О.А. Косимов // Конференсияи илмию амалии минтақавӣ. – Исфара, 2018. – 8 с.

305. Солиев, О.М. Идентификация авторов экономика политической произведений с помощью символьных триграмм [Текст] / О.М. Солиев, О.А. Косимов // Вестник ПИТТУ имени академика М.С. Осими «Научно-технический журнал». – Худжанд, 2019. – №2 (11). – С. 22-29.

306. Косимов, О.А. Идентификация авторов экономико-политических произведений с помощью символьных униграмм [Текст] / О.А. Косимов // В сборнике: Состояние и перспективы развития ИТ-образования Сборник докладов и научных статей Всероссийской научно-практической конференции, Чувашская Республика. – 2019. – С. 131-138.

307. Солиев, О.М. Муайянкунии муаллифи асарҳои сиёсӣ-иқтисодӣ бо воситаи триграммаҳои рамзӣ [Матн] / О.М. Солиев, О.А. Косимов // Конференсияи илмӣ-амалии байналмилалӣ дар мавзӯи «Масъалаҳои муосири математика ва методикаи таълими он», Донишгоҳи давлатии Бохтар ба номи Н. Хусрав, Бохтар. – 2019.

308. Солиев, О.М. Муайянкунии муаллифи асарҳои сиёсӣ-иқтисодӣ бо воситаи униграммаҳои рамзӣ [Матн] / О.М. Солиев, О.А. Косимов // Донишгоҳи давлатии омӯзгории Тоҷикистон ба номи С. Айнӣ, конференсияи илмӣ-амалии ҷумҳуриявӣ дар мавзӯи «Мушкилотҳои муосири таҳқиқотии илмҳои табиӣ-риёзӣ». – Душанбе, 30.05.2019.

309. Косимов, О.А. Идентификация авторов экономики политических произведений с помощью символьных биграмм [Текст] / О.А. Косимов // Ежегодная межвузовская научно-техническая конференция студентов, аспирантов и молодых специалистов имени Е.В. Арменского, МИЭМ НИУ ВШЭ, Секция №1 «Математика и компьютерное моделирование». – 2020.

310. Усманов, З.Д. Муайянкунии шифри ихтисос дар асарҳои илмӣ бо воситаи биграммаҳои ҳарфӣ [Матн] / З.Д. Усманов, О.А. Косимов // Вестник ПИТТУ имени академика М.С. Осими «Научно-технический журнал». – Худжанд, 2020. – №1 (14). – С. 7-16.

311. Косимов, О.А. Муайянкунии шифри ихтисос дар асарҳои илмӣ бо воситаи триграммаҳои ҳарфӣ [Матн] / О.А. Косимов // Политехнический

вестник, Серия: интеллект, инновации, инвестиции. – 2020. – 2(50). – С. 36-40.

312. Косимов, О.А. Идентификация авторов экономики политических произведений с помощью буквенных биграмм [Текст] / О.А. Косимов // Материалы республиканской научно-теоретической конференции на тему «Цифровая экономика и необходимость внедрения новой системы национальных счетов», 17 февраля 2021 года. – Душанбе, типография ТНУ. – С. 76-82.

313. **Косимов, А.А.** Об анаграммах в произведениях Л. Шерали [Текст] / А. Набави, **А.А. Косимов**, О.А. Косимов // Машъалдори шеър (Ҷойгоҳи Лоиқ Шералӣ дар шеъри муосир). – Душанбе, 2021. – С. 376-383.

314. Бахтеев, К.С. Компьютерные методы идентификации авторов текста [Текст] / К.С. Бахтеев // Материалы науч.-практ. конф. профессорско-преподавательского состава и магистрантов факультета Управления и информационных технологий РТСУ. «XXIII Славянские чтения». – 2019. – РТСУ. – С.113-116.

315. Бахтеев, К.С. Автоматическая символьная предобработка текстов таджикского языка [Текст] / К.С. Бахтеев // Известия АН Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2012. – №4(149). – С.37-40.

316. Бахтеев, К.С. О перспективности консонантного письма [Текст] / М.А. Умаров, К.С. Бахтеев // Известия АН Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2019. – №1(174) – С. 50-56.

317. Каюмов, М.М. О распознавании авторов произведений на основе частотности однозначных и многозначных $\alpha\beta$-кодов словоформ [Текст] / М.М. Каюмов // Известия АН РТ отделения: физико-математических, химических, геологических и технических наук. – Душанбе. – 4(181). – 2020. – С. 30-40.

318. Каюмов, М.М. О распознавании авторов произведений на основе частотности N-значных $\alpha\beta$-кодов словоформ [Текст] / М.М. Каюмов // ДНАНТ. – 2020. – Т.81. – №1-2.

319. Каюмов, М.М. Кластеризация произведений модельной коллекции текстов методом ближайшего соседа [Текст] / М.М. Каюмов // Вестник технологического университета Таджикистана. – Душанбе. – 1 (44). – 2021.

320. Каюмов, М.М. О распознавании жанров произведений на основе частотности $\alpha\beta$-кодов словоформ [Текст] / М.М. Каюмов // Политехнический Вестник ТТУ имени ак. М.С. Осими, Серия: интеллект, инновации, инвестиции. – Душанбе: ТТУ. – 2021. – 1(53). – С. 31-39.

321. Каюмов, М.М. Исследование эффективности применения различных цифровых портретов печатного текста для распознавания авторов

произведений [Текст] / М.М. Каюмов, Ш. Сафаров, Г. Гуломов, Ш. Ширинов // Материалы международной научно-практической конференции «Перспектива развития науки и образования» (27-28-го ноября 2019) Таджикский технический университет имени акад. М.С. Осими, часть 1. – Душанбе, 2019. – С. 113-114.

322. Каюмов, М.М. Применения «*αβ*-код»-а в качестве цифрового портрета печатного текста для распознавания авторов произведений [Текст] / М.М. Каюмов // Материалы международной научно-практической конференции «Применение информационно-телекоммуникационных технологий в создании электронного правительства и индустриализации страны» Таджикский технический университет имени академика М.С. Осими. – Душанбе, 2020. – С. 168-171.

323. Каюмов, М.М. Распознавание автора фрагмента текста на основе *αβ*-кодов [Текст] / М.М. Каюмов // Материалы VI Республиканской научно-практической конференции «Наука-основа инновационного развития» Таджикский технический университет имени акад. М.С. Осими. Материалы, часть 1. – Душанбе, 2021. – С. 127-135.

324. Каюмов, М.М. О распознавании автора текста на основе частотности *αβ*-кодов словоформ с учетом лексикографического порядка [Текст] / М.М. Каюмов // Материалы VI Республиканской научно-практической конференции «Наука-основа инновационного развития» Таджикский технический университет имени акад. М.С. Осими. Материалы, часть 1. – Душанбе, 2021. – С. 135-143.

Авторские публикации.

Список публикаций соискателя ученой степени по теме диссертации в изданиях из перечня ВАК РТ:

[1-А]. Косимов, А.А. Цифровой образ “Шахнаме” (“Книги царей”) А.Фирдоуси [Текст]. / З.Д. Усманов, **А.А. Косимов** // Доклады Академии наук Республики Таджикистан. – 2014. – Том 57. – № 6. – С. 471-476.

[2-А]. Косимов, А.А. Частотность букв таджикской литературы [Текст] / З.Д. Усманов, **А.А. Косимов** // Доклады Академии наук Республики Таджикистан. – 2015. – Том 58. – № 2. – С. 112-115.

[3-А]. Косимов, А.А. Частотность биграмм в таджикской литературе [Текст] / З.Д. Усманов, **А.А. Косимов** // Доклады Академии наук Республики Таджикистан. – 2016. – Том 59. – № 1-2. – С. 28-32.

[4-А]. Косимов, А.А. О распознавании авторства таджикского текста [Текст] / З.Д. Усманов, **А.А. Косимов** // Доклады Академии наук Республики Таджикистан. – 2016. – Том 59. – № 3-4. – С. 114-119.

[5-А]. Косимов, А.А. О множестве анаграмм в поэме А.Фирдауси “Шахнаме” [Текст] / **А.А. Косимов** // Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2016. – № 1 (162). – С. 48-53.

[6-А]. Косимов, А.А. Оценка эффективности использования униграмм при идентификации текста [Текст] / **А.А. Косимов** // Доклады Академии наук Республики Таджикистан. – 2017. – Т.60. – № 3-4. – С. 132-137.

[7-А]. Косимов, А.А. Оценка эффективности использования биграмм при идентификации текста [Текст] / **А.А. Косимов** // Доклады Академии наук Республики Таджикистан. – 2017. – Т.60. – № 5-6. – С. 224-229.

[8-А]. Косимов, А.А. Оценка эффективности использования триграмм при идентификации текста [Текст] / **А.А. Косимов** // Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2017. – №1(166). – С. 51-57.

[9-А]. Косимов, А.А. Определение минимального объёма выборки слов для идентификации текста [Текст] / **А.А. Косимов** // Вестник Таджикского национального университета, Серия естественных наук, Душанбе. – 2017. – №1/5. – С. 178-180.

[10-А]. Косимов, А.А. О минимальном объёме текста, необходимого для распознавания его автора [Текст] / **А.А. Косимов** // Доклады Академии наук Республики Таджикистан. – 2017. – Т.60. – № 9. – С. 398-401.

[11-А]. Косимов, А.А. Об идентификации текста с помощью символьных триграмм [Текст] / **А.А. Косимов,** О.А. Косимов // Вестник Технологического Университета Таджикистана, Душанбе. – 2018. – С. 37-42.

[12-А]. Косимов, А.А. Программный комплекс Tajik_Text_Author [Текст] / **А.А. Косимов** // Политехнический вестник, Серия: интеллект, инновации, инвестиции. – 2019. – 3(47). – С. 22-28.

[13-А]. Косимов, А.А. Применение специфичного цифрового портрета для идентификации авторов произведений [Текст] / **А.А. Косимов**, К.С. Бахтеев // Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2019. – №3(176). – С. 7-11.

[14-А]. Косимов, А.А. О распознавании автора текста на основе частотности слогов [Текст] / Х.А. Худойбердиев, **А.А. Косимов** // Доклады Академии наук Республики Таджикистан. – 2019. – Т.62. – № 11-12. – С. 641-645.

[15-А]. Косимов, А.А. О распознавании автора текстового фрагмента [Текст] / **А.А. Косимов**, К.С. Бахтеев // Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2019. – №4(177). – С. 18-25.

[16-А]. Косимов, А.А. К вопросу об автоматическом распознавании авторства и стилей произведений таджикско-персидской художественной литературы [Текст] / З.Д. Усманов, **А.А. Косимов** // Доклады Академии наук Республики Таджикистан. – 2020. – Т.63. – № 1-2. – С. 49-54.

[17-А]. Косимов, А.А. О распознавании автора текста на основе частотности длин предложений [Текст] / **А.А. Косимов,** К.С. Бахтеев // Доклады Академии наук Республики Таджикистан. – 2020. – Т.63. – №3-4. – С. 180-186.

[18-А]. Косимов, А.А. Автоматический поиск анаграмм словоформных N-грамм [Текст] / З.Д. Усманов, **А.А. Косимов** // Доклады Академии наук Республики Таджикистан. – 2020. – Т.63. – № 5-6. – С. 316-321.

[19-А]. Косимов, А.А. О влиянии цифрового портрета текста на распознавание автора произведения [Текст] / З.Д. Усманов, **А.А. Косимов** // Известия Академии наук Республики Таджикистан. Отделение физико-математических, химических, геологических и технических наук. – 2020. – №3(180). – С. 36-42.

[20-А]. Косимов, А.А. Об идентификации текста на основе частотности слогов [Текст] / Х.А. Худойбердиев, **А.А. Косимов**, Х.А. Тошхуджаев // Политехнический вестник, Серия: интеллект, инновации, инвестиции. – 2020. – 2(50). – С. 52-56.

[21-А]. Косимов, А.А. Об автоматическом распознавании языка произведений [Текст] / З.Д. Усманов, **А.А. Косимов** // Доклады Академии наук Республики Таджикистан. – 2020. – Т.63. – № 7-8. – С. 461-466.

[22-А]. Косимов, А.А. Оценка эффективности применения γ-классификатора для атрибуции искусственно сгенерированных поэм «Шахнаме» А. Фирдоуси [Текст] / М.Ё. Мухсинзода, **А.А. Косимов** // Политехнический вестник, Серия: интеллект, инновации, инвестиции. – 2020. – 4(52). – С. 35-39.

[23-А]. Косимов, А.А. Тестирование γ-классификатора, настроенного на распознавание языков произведений на основе латинского алфавита [Текст] / З.Д. Усманов, **А.А. Косимов** // Научный Вестник НГТУ «Системы анализа и обработки данных». – Том 82. – № 2. – 2021. – С. 83-94.

[24-А]. Косимов, А.А. Об автоматическом распознавании языков произведений на основе кириллического алфавита [Текст] / М.Л. Мирзохасанов, **А.А. Косимов** // Вестник Технологического Университета Таджикистана, Душанбе. – 2021. – 1(44). – С. 101-107.

[25-А]. Косимов, А.А. Структура однородностей поэм произведения А. Фирдоуси «Шахнаме» [Текст] / **А.А. Косимов**, Н.М. Курбонов // Политехнический вестник, Серия: интеллект, инновации, инвестиции. – 2021. – 2(54). – С. 35-38.

[26-А]. Косимов, А.А. Об однородности оригинала и его перевода [Текст] /

А.А. Косимов // Доклады Национальной академии наук Таджикистана. – 2021. – Т.64. – № 11-12. – С. 660-665.

[27-А]. Косимов, А.А. О распознавании автора текстового фрагмента на основе частотности слогов [Текст] / **А.А. Косимов** // Политехнический вестник, Серия: интеллект, инновации, инвестиции. – 2021. – 4(56). – С. 59-64.

[28-А]. Косимов, А.А. О распознавании автора текстового фрагмента на основе частотности буквенных биграмм [Текст] / **А.А. Косимов** // Системы анализа и обработки данных. – Том 85. – № 1. – 2022. – С. 73-82. DOI: 10.17212/2782-2001-2022-1-73-82.

[29-А]. Косимов, А.А. О распознавании автора текстового фрагмента на основе частотности буквенных триграмм [Текст] / **А.А. Косимов, Н.А. Шокирова** // Вестник Технологического Университета Таджикистана, Душанбе. – 2022. – 2(49). – С. 35-43.

[30-А]. Косимов, А.А. О влиянии порядка буквенных униграмм на распознавание автора произведения [Текст] / **А.А. Косимов** // Доклады Национальной академии наук Таджикистана. – 2022. – Т.65. – № 5-6. – С. 324-330.

[31-А]. Косимов, А.А. О влиянии порядка буквенных триграмм на распознавание автора произведения [Текст] / **А.А. Косимов** // Вестник Филиала МГУ имени М.В. Ломоносова в городе Душанбе, Серия естественных наук. – 2022. – № 1. – С. 14-21.

[32-А]. Косимов, А.А. Определение шифр специальности с помощью символьных униграмм [Текст] / **А.А. Косимов** // Вестник Филиала МГУ имени М.В. Ломоносова в городе Душанбе, Серия естественных наук. – 2023. – Том 1. – №1 (29). – С. 16-24.

[33-А]. Косимов, А.А. О влиянии порядка символьных триграмм на определение языка произведения [Текст] / **А.А. Косимов** // Политехнический вестник, Серия: интеллект, инновации, инвестиции. – 2023. – 1(61). – С. 34-37.

[34-А]. Косимов, А.А. О влиянии порядка буквенных биграмм на определение языка произведения [Текст] / **И.К. Каландарбеков, А.А. Косимов** // Вестник Филиала МГУ имени М.В. Ломоносова в городе Душанбе, Серия естественных наук. – 2023. – Том 1. – №2 (31). – С. 26-32.

Монографии и учебные пособия:

[35-А]. Косимов, А.А. Барномарезии ба объект нигаронидашуда (БОН) [Матн] / **А.А. Косимов** // ДПДТТ ба номи ак. М.С. Осимӣ, Хуҷанд: «Меҳвари дониш». – 2019. – 138 с.

[36-А]. Косимов, А.А. Амалияи барномасозӣ дар забони Python [Матн] / **А.А. Косимов** // ДТТ ба номи ак. М.С. Осимӣ, Душанбе. – 2023. – 163 с.

[37-А]. Косимов, А.А. Становление компьютерной лингвистики

Таджикистана: монография [Текст] / **А.А. Косимов** // ТТУ имени академика М.С. Осими, – 05.05.2021 (№34), Душанбе: «Ирфон». – 2021. – 102 с.

[38-А]. Косимов, А.А. Разработка программного комплекса для распознавания автора незнакомого текста: монография [Текст] / З.Д. Усманов, **А.А. Косимов** // Институт математики имени А. Джураева НАНТ. – 12.01.2022 (№1), Душанбе: «Дониш». – 2022. – 105 с.

Публикации в других изданиях, трудах и материалах конференций:

[39-А]. Косимов, А.А. О минимальном числе высокоточных *N*-грамм, необходимых для распознавания автора текста [Текст] / **А.А. Косимов** // Российско-китайский научный журнал «Содружество», Ежемесячный научный журнал, научно-практической конференции. – 2017. – Часть 1. – № 17. – С. 58-59.

[40-А]. Косимов, А.А. Оиди муносибати шаклҳои калима ва калимаҳо дар ҳуруфоти форсии китоби «Шоҳнома»-и А. Фирдавсӣ [Матн] / **А.А. Косимов** // Роль ИКТ в инновационном развитии экономики Республики Таджикистан, Материалы международной научно-практической конференции, Бахшида ба 80-солагии академик Усмонов Зафар Ҷӯраевич, Душанбе: Баҳманрӯд. – 2017. – С. 321-328.

[41-А]. Косимов, А.А. О метризации произведений художественной литературы [Текст] / З.Д. Усманов, **А.А. Косимов** // Материалы двадцать первого научно-практического семинара «Новые информационные технологии в автоматизированных системах», Москва. – 2018. – С. 183-186.

[42-А]. Косимов, А.А. Об идентификации текста с помощью символьных биграмм [Текст] / Х.А. Худойбердиев, **А.А. Косимов**, О.А. Косимов // Саромади маорифчиёни асил, Конференсияи илмию амалии минтақавӣ бахшида ба 90-солагии устод Темурхон Мақсудов, Исфара. – 2018. – С. 175-179.

[43-А]. Косимов, А.А. Машинный анализ соотношений словоформ и словоупотреблений персидского языка в произведении А. Фирдоуси «Шахнаме» [Текст] / Х.А. Худойбердиев, **А.А. Косимов** // Вестник ПИТТУ имени академика М.С. Осими «Научно-технический журнал», Худжанд. – 2018. – №1 (6). – С. 7-14.

[44-А]. Косимов, А.А. О применимости γ-классификатора к распознаванию авторства и тематики художественных произведений [Текст] / З.Д. Усманов, **А.А. Косимов** // Материалы двадцать второго научно-практического семинара «Новые информационные технологии в автоматизированных системах», Москва. – 2019. – С. 174-178.

[45-А]. Косимов, А.А. О соотношении словоформ и словоупотреблений в творчестве А. Навои [Текст] / **А.А. Косимов** // В сборнике: Состояние и перспективы развития ИТ-образования Сборник докладов и научных статей Всероссийской научно-практической конференции, Чувашская Республика. –

2019. – С. 125-131.

[46-А]. Косимов, А.А. О распознавании автора текста на основе частотности буквенных триграмм [Текст] / **А.А. Косимов** // Вестник ПИТТУ имени академика М.С. Осими «Научно-технический журнал», Худжанд. – 2019. – №4 (13). – С. 28-37.

[47-А]. Kosimov, A.A. About the automatic recognition of the languages of works based on the latin alphabet [Text] / Z.J. Usmanov, **A.A. Kosimov** // Proceedings of the 8th International Scientific and Practical Conference science and practice: implementation to modern society, Manchester, Great Britain. – 26-28.12.2020. – №3 (39). – pp. 834-840.

[48-А]. Косимов, А.А. О распознавании автора текста на узбекском языке с помощью символьных биграмм [Текст] / **А.А. Косимов**, П.Э. Зульфикарова // Ежегодная межвузовская научно-техническая конференция студентов, аспирантов и молодых специалистов имени Е.В. Арменского, МИЭМ НИУ ВШЭ, Секция №1 «Математика и компьютерное моделирование». – 2020. – С. 50-51.

[49-А]. Косимов, А.А. О распознавании автора текста на основе частотности буквенных биграмм [Текст] / **А.А. Косимов**, Ф.А. Рахмонов // Конференсияи илмӣ-амалии омӯзгорон, муҳаққиқони ҷавон, докторантон PhD, магистрантон ва донишҷӯён бахшида ба эълон гардидани солҳои 2019-2021 «Солҳои рушди деҳот, сайёҳӣ ва ҳунарҳои мардумӣ», солҳои 2020-2040 «Бистсолаи омӯзиш ва рушди фанҳои табиатшиносӣ, дақиқ ва риёзӣ дар соҳаи илму маориф», Рӯзи илми тоҷик ва 30-солагии Истиқлолияти давлатии Ҷумҳурии Тоҷикистон, ДПДТТХ ба номи М.С. Осимӣ, Хуҷанд. – 30 апрели соли 2020. – 11 с.

[50-А]. Kosimov, A.A. About the position of the culmination point in art works [Text] / Z.J. Usmanov, **A.A. Kosimov** // Труды международной конференции по компьютерной и когнитивной лингвистике TEL-2020, Казань: Изд-во Академии наук РТ. – 2020. – С. 70-74.

[51-А]. Косимов, А.А. О распознавании автора текста на узбекском языке с помощью символьных униграмм [Текст] / Х.А. Худойбердиев, **А.А. Косимов**, П.Э. Зульфикарова // Проблемы вычислительной и прикладной математики, Ташкент. – 2020. – №6(30). – С. 49-55.

[52-А]. Косимов, А.А. К вопросу о распознавании однородных пар произведений художественной литературы [Текст] / З.Д. Усманов, **А.А. Косимов** // Труды международной конференции по компьютерной и когнитивной лингвистике TEL-2020, Казань: Изд-во Академии наук РТ. – 2020. – С. 137-153.

[53-А]. Косимов, А.А. Распознавание языка произведения с помощью γ-классификатора [Текст] / З.Д. Усманов, **А.А. Косимов** // Труды международной конференции по компьютерной и когнитивной лингвистике TEL-2020, Казань: Изд-во Академии наук РТ. – 2020. – С. 174-179.

[54-А]. Косимов, А.А. Определение авторства таджикских литературных текстов на основе частотности слогов [Текст] / Х.А. Худойбердиев, **А.А. Косимов** // Вестник ПИТТУ имени академика М.С. Осими «Научно-технический журнал», Худжанд. – 2020. – №2 (15). – С. 7-16.

[55-А]. Косимов, А.А. О распознавании автора текста на узбекском языке с помощью символьных триграмм [Текст] / **А.А. Косимов**, П.Э. Зульфикарова // Вестник ПИТТУ имени академика М.С. Осими «Научно-технический журнал», Худжанд. – 2020. – №2 (15). – С. 24-31.

[56-А]. Косимов, А.А. Тестирование γ-классификатора, настроенного на распознавание языков произведений на основе кириллического алфавита [Текст] / **А.А. Косимов**, Х.А. Шарипов // Конференсияи ҷумҳуриявии илмӣ-амалии «Илм – асоси рушди инноватсионӣ», Донишгоҳи техникии Тоҷикистон ба номи академик М.С. Осимӣ, шаҳри Душанбе. – 27-28 апрели соли 2021. – С. 314-318.

[57-А]. Косимов, А.А. Барномаи зидди асардуздӣ (ANTIPLAGIAT_TJ) [Матн] / **А.А. Косимов**, Р.Р. Булбулов, А.А. Хасанов, Ш.Г. Мерганзода // Конференсияи ҷумҳуриявии илмӣ-амалии «Илм – асоси рушди инноватсионӣ», Донишгоҳи техникии Тоҷикистон ба номи академик М.С. Осимӣ, шаҳри Душанбе. – 27-28 апрели соли 2021. – С. 318-321.

[58-А]. Косимов, А.А. О распознавании автора текста на основе частотности буквенных униграмм [Текст] / **А.А. Косимов**, Р.Ш. Умарализода, А.А. Хасанов, Ш.С. Саидов // Конференсияи ҷумҳуриявии илмӣ-амалии «Илм – асоси рушди инноватсионӣ», Донишгоҳи техникии Тоҷикистон ба номи академик М.С. Осимӣ, шаҳри Душанбе. – 27-28 апрели соли 2021. – С. 322-326.

[59-А]. Kosimov, A.A. About of the metric homogeneity of texts in Slavic languages [Text] / Z.J. Usmanov, **A.A. Kosimov** // XI международная научно-техническая конференция «Открытые семантические технологии проектирования интеллектуальных систем», Open Semantic Technologies for Intelligent Systems (OSTIS-2021), г. Минск, Республика Беларусь. – 16-18 сентября 2021. – С. 313-316.

[60-А]. Косимов, А.А. Об автоматическом распознавании языков произведений на основе латинского алфавита [Текст] / **А.А. Косимов** // Материалы международной научно-практической конференции «Технические науки и инженерное образование для устойчивого развития», Таджикский технический университет имени академика М.С. Осими, Душанбе. – Часть 2. – 12-13 ноября 2021 г. – С. 104-108.

[61-А]. Косимов, А.А. О применимости γ-классификатора к распознаванию однородности текстов на славянских языках [Текст] / **А.А. Косимов** // XXII Международная конференция «Информатика: проблемы, методы, технологии» (IPMT-2022), Воронежский государственный университет, Воронеж. – 10-12

февраля 2022 г. – С. 1136-1145.

[62-А]. Косимов, А.А. Исследование статистических закономерностей распознавания языка текстов на основе латинского алфавита в корпусах произведений художественной литературы [Текст] / **А.А. Косимов** // VI Международной научно-практической конференции «Global and regional aspects of sustainable development», Копенгаген, Дания. – 26-28 февраля 2022 года. – №100. – С. 814-828.

[63-А]. Косимов, А.А. О распознавании автора текстового фрагмента на основе частотности буквенных униграмм [Текст] / **А.А. Косимов,** К.А. Бобозода // Современные проблемы естествознания в науке и образовательном процессе: сборник материалов Республиканской научно-практической конференции, посвященной двадцатилетию изучения и развития естественных, точных и математических наук, РТСУ, Душанбе. – 2022. – С. 239-244.

[64-А]. Косимов, А.А. Муайянкунии шифри ихтисос дар асарҳои илмӣ бо воситаи униграммаҳои ҳарфӣ [Матн] / **А.А. Косимов,** М.С. Саидова, И.А. Ҷумаева, М.Б. Ғаниева // Конференсияи Ҷумҳуриявии VI илмӣ-амалии донишҷӯён, магистрантҳо ва аспирантону унвонҷӯён таҳти унвони “Илм – асоси рушди инноватсионӣ”, Донишгоҳи техникии Тоҷикистон ба номи академик М.С. Осимӣ, шаҳри Душанбе. – 27-28 апрели соли 2022. – С. 46-50.

[65-А]. Косимов, А.А. Исследование статистических закономерностей распознавания языка текстов на основе кириллического алфавита в корпусах произведений художественной литературы [Текст] / С.М. Пиров, **А.А. Косимов** // Материалы международной научно-практической конференции «XII Ломоносовские чтения», посвященной 30-летию установления дипломатических отношений между Республикой Таджикистан и Российской Федерацией, Филиал МГУ имени М.В. Ломоносова в г. Душанбе. – 29-30 апреля 2022 года. – С. 49-58.

[66-А]. Косимов, А.А. О влиянии порядка буквенных биграмм на распознавание автора произведения [Текст] / **А.А. Косимов** // Материалы международной научно-практической конференции «XII Ломоносовские чтения», посвященной 30-летию установления дипломатических отношений между Республикой Таджикистан и Российской Федерацией, Филиал МГУ имени М.В. Ломоносова в г. Душанбе. – 29-30 апреля 2022 года. – С. 20-27.

[67-А]. Косимов, А.А. Исследование статистических закономерностей распознавания автора текстов в корпусах произведений художественной литературы [Текст] / **А.А. Косимов** // Сборник международной конференции, посвящённой памяти профессора А.А. Тарасова и О.В. Казарина, по теме «Взаимодействие вузов, научных организаций и учреждений культуры в сфере защиты информации и технологий безопасности», г. Москва. – 19 и 20 апреля 2022 года. – С. 155-167.

[68-А]. Косимов, А.А. О распознавании автора отсканированного рукописного текста на основе частотности значения каналов RGB в пикселях [Текст] / **З.Х. Рахмонов, А.А. Косимов, С. Хочиабдурахим** // В сборнике: Современные проблемы математики. Материалы международной конференции, посвящённой 50-летию Института математики им. А.Джураева Национальной академии наук Таджикистана, г. Душанбе. – 2023. – С. 104-108.

Свидетельства о государственной регистрации программы для ЭВМ:

[69-А]. Косимов, А.А. База данных $\alpha\beta$-кодирования для распознавания анаграмм / З.Д. Усманов, О.М. Солиев, Х.А. Худойбердиев, Г.М. Довудов, **А.А. Косимов** // Свидетельство о государственной регистрации информационного ресурса, Республика Таджикистан. – 16.05.2018. – №4201800377.

[70-А]. Косимов, А.А. Web-приложение проверки уникальности текста на таджикском языке Taj_Text_Plagiat / З.Д. Усманов, О.М. Солиев, Х.А. Худойбердиев, П.А. Солиев, **А.А. Косимов** // Свидетельство о государственной регистрации информационного ресурса, Республика Таджикистан. – 16.05.2018. – №4201800378.

[71-А]. Косимов, А.А. База данных «Единица измерений текстов произведений таджикских классических поэтов и писателей» / З.Д. Усманов, Х.А. Худойбердиев, **А.А. Косимов** // Свидетельство о государственной регистрации информационного ресурса, Республика Таджикистан. – 16.05.2018. – №4201800380.

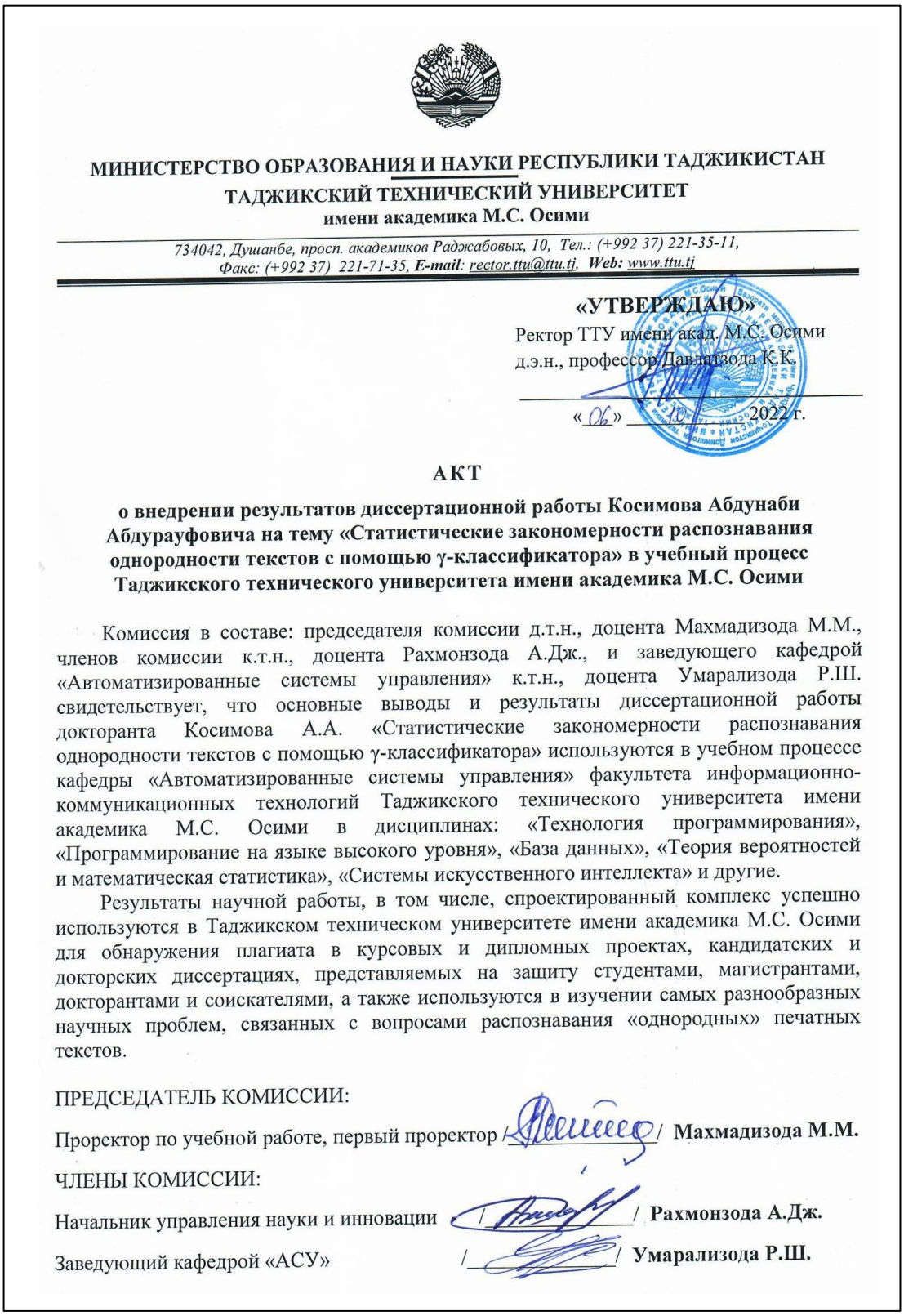
[72-А]. Косимов, А.А. База данных «Единица измерений текстов произведений таджикских современных поэтов и писателей» / З.Д. Усманов, Х.А. Худойбердиев, **А.А. Косимов** // Свидетельство о государственной регистрации информационного ресурса, Республика Таджикистан. – 16.05.2018. – №4201800381.

[73-А]. Косимов, А.А. База данных $\alpha\beta$-кодов словоформ для определения автора незнакомого текста / З.Д. Усманов, **А.А. Косимов**, М.М. Каюмов // Свидетельство о государственной регистрации информационного ресурса, Республика Таджикистан. – 07.06.2021. – №1202100478.

ПРИЛОЖЕНИЯ

Приложение 1.

Практическое использование результатов исследований



МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РЕСПУБЛИКИ ТАДЖИКИСТАН
ТАДЖИКСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
имени академика М.С. Осими

734042, Душанбе, просп. академиков Раджабовых, 10, Тел.: (+992 37) 221-35-11,
Факс: (+992 37) 221-71-35, E-mail: rector.ttu@ttu.tj, Web: www.ttu.tj

«УТВЕРЖДАЮ»
Ректор ТТУ имени акад. М.С. Осими
д.э.н., профессор Давлатзода К.К.
«06» 10 2022 г.

АКТ

о внедрении результатов диссертационной работы Косимова Абдунаби Абдурауфовича на тему «Статистические закономерности распознавания однородности текстов с помощью γ-классификатора» в учебный процесс Таджикского технического университета имени академика М.С. Осими

Комиссия в составе: председателя комиссии д.т.н., доцента Махмадизода М.М., членов комиссии к.т.н., доцента Рахмонзода А.Дж., и заведующего кафедрой «Автоматизированные системы управления» к.т.н., доцента Умарализода Р.Ш. свидетельствует, что основные выводы и результаты диссертационной работы докторанта Косимова А.А. «Статистические закономерности распознавания однородности текстов с помощью γ-классификатора» используются в учебном процессе кафедры «Автоматизированные системы управления» факультета информационно-коммуникационных технологий Таджикского технического университета имени академика М.С. Осими в дисциплинах: «Технология программирования», «Программирование на языке высокого уровня», «База данных», «Теория вероятностей и математическая статистика», «Системы искусственного интеллекта» и другие.

Результаты научной работы, в том числе, спроектированный комплекс успешно используются в Таджикском техническом университете имени академика М.С. Осими для обнаружения плагиата в курсовых и дипломных проектах, кандидатских и докторских диссертациях, представляемых на защиту студентами, магистрантами, докторантами и соискателями, а также используются в изучении самых разнообразных научных проблем, связанных с вопросами распознавания «однородных» печатных текстов.

ПРЕДСЕДАТЕЛЬ КОМИССИИ:
Проректор по учебной работе, первый проректор / / **Махмадизода М.М.**

ЧЛЕНЫ КОМИССИИ:
Начальник управления науки и инновации / / **Рахмонзода А.Дж.**
Заведующий кафедрой «АСУ» / / **Умарализода Р.Ш.**

Рисунок П1.1. – Акт о внедрении результатов исследования (ТТУ имени академика М.С. Осими)

ВАЗОРАТИ РУШДИ ИҚТИСОД ВА САВДОИ ҶУМҲУРИИ ТОҶИКИСТОН
МУАССИСАИ ДАВЛАТИИ «МАРКАЗИ МИЛЛИИ ПАТЕНТУ ИТТИЛООТ»
МИНИСТЕРСТВО ЭКОНОМИЧЕСКОГО РАЗВИТИЯ И ТОРГОВЛИ РЕСПУБЛИКИ ТАДЖИКИСТАН
ГОСУДАРСТВЕННОЕ УЧРЕЖДЕНИЕ «НАЦИОНАЛЬНЫЙ ПАТЕНТНО- ИНФОРМАЦИОННЫЙ ЦЕНТР»

ШАҲОДАТНОМА
дар бораи бақайдгирии давлатии захираи иттилоотӣ

СВИДЕТЕЛЬСТВО
о государственной регистрации информационного ресурса

Номгуй / Наименование: Web-приложение проверки уникальности текста на таджикском языке Taj_Text_Plagiat
Сарзамин / Страна: Республика Таджикистан
Доранда / Владелец: Усманов З.Дж., Солиев О.М., Худойбердиев Х.А., Солиев П.А, Косимов А.А
Таҳиягар / Разработчик: Усманов З.Дж., Солиев О.М., Худойбердиев Х.А., Солиев П.А., Косимов А.А
№ қайди давлатӣ / № государственной регистрации: № 4201800378
Ба Феҳристи давлатии захираҳои иттилоотии Ҷумҳурии Тоҷикистон дохил карда шудааст
Внесен в Государственный реестр информационных ресурсов Республики Таджикистан: 16 мая 2018 г.

Директор: Ҷ. Ҷумъахонзода

Рисунок П1.2. – Web-приложение проверки уникальности текста на таджикском языке Taj_Text_Plagiat

ВАЗОРАТИ РУШДИ ИҚТИСОД ВА САВДОИ ҶУМҲУРИИ ТОҶИКИСТОН
МУАССИСАИ ДАВЛАТИИ «МАРКАЗИ МИЛЛИИ ПАТЕНТУ ИТТИЛООТ»
МИНИСТЕРСТВО ЭКОНОМИЧЕСКОГО РАЗВИТИЯ И ТОРГОВЛИ РЕСПУБЛИКИ ТАДЖИКИСТАН
ГОСУДАРСТВЕННОЕ УЧРЕЖДЕНИЕ «НАЦИОНАЛЬНЫЙ ПАТЕНТНО- ИНФОРМАЦИОННЫЙ ЦЕНТР»

ШАҲОДАТНОМА
дар бораи бақайдгирии давлатии захираи иттилоотӣ

СВИДЕТЕЛЬСТВО
о государственной регистрации информационного ресурса

Номгуй / Наименование: База данных «Единица измерений текстов произведений таджикских классических поэтов и писателей»
Сарзамин / Страна: Республика Таджикистан
Доранда / Владелец: Усманов З.Дж., Худойбердиев Х.А., Косимов А.А.
Таҳиягар / Разработчик: УсмановЗ.Дж., Худойбердиев Х.А., Косимов А.А.
№ қайди давлатӣ / № государственной регистрации: № 4201800380
Ба Феҳристи давлатии захираҳои иттилоотии Ҷумҳурии Тоҷикистон дохил карда шудааст
Внесен в Государственный реестр информационных ресурсов Республики Таджикистан: 16 мая 2018 г.

Директор: Ҷ. Ҷумъахонзода

Рисунок П1.3. – База данных «Единица измерений текстов произведений таджикских классических поэтов и писателей»

Рисунок П1.4. – Участие в вебинаре на тему: «Экспертная оценка оригинальности научных работ с помощью системы «Антиплагиат»»

Рисунок П1.5. – База данных *αβ*-кодирования для распознавания анаграмм

ВАЗОРАТИ РУШДИ ИҚТИСОД ВА САВДОИ ҶУМҲУРИИ ТОҶИКИСТОН
МУАССИСАИ ДАВЛАТИИ «МАРКАЗИ МИЛЛИИ ПАТЕНТУ ИТТИЛООТ»
МИНИСТЕРСТВО ЭКОНОМИЧЕСКОГО РАЗВИТИЯ И ТОРГОВЛИ РЕСПУБЛИКИ ТАДЖИКИСТАН
ГОСУДАРСТВЕННОЕ УЧРЕЖДЕНИЕ «НАЦИОНАЛЬНЫЙ ПАТЕНТНО- ИНФОРМАЦИОННЫЙ ЦЕНТР»

ШАҲОДАТНОМА
дар бораи бақайдгирии давлатии захираи иттилоотӣ

СВИДЕТЕЛЬСТВО
о государственной регистрации информационного ресурса

Номгуй / Наименование: База данных «Единица измерений текстов произведений таджикских современных поэтов и писателей
Сарзамин / Страна: Республика Таджикистан
Доранда / Владелец: Усманов З.Дж., Худойбердиев Х.А., Косимов А.А.
Таҳиягар / Разработчик: УсмановЗ.Дж., Худойбердиев Х.А., Косимов А.А.
№ кайди давлатӣ / № государственной регистрации: № 4201800381
Ба Феҳристи давлатии захираҳои иттилоотии Ҷумҳурии Тоҷикистон дохил карда шудааст
Внесен в Государственный реестр информационных ресурсов Республики Таджикистан: 16 мая 2018 г.

Директор: Ч. Ҷумъахонзода

Рисунок П1.6. – База данных «Единица измерений текстов произведений таджикских современных поэтов и писателей»

ВАЗОРАТИ РУШДИ ИҚТИСОД ВА САВДОИ ҶУМҲУРИИ ТОҶИКИСТОН
МУАССИСАИ ДАВЛАТИИ «МАРКАЗИ МИЛЛИИ ПАТЕНТУ ИТТИЛООТ»
МИНИСТЕРСТВО ЭКОНОМИЧЕСКОГО РАЗВИТИЯ И ТОРГОВЛИ РЕСПУБЛИКИ ТАДЖИКИСТАН
ГОСУДАРСТВЕННОЕ УЧРЕЖДЕНИЕ «НАЦИОНАЛЬНЫЙ ПАТЕНТНО- ИНФОРМАЦИОННЫЙ ЦЕНТР»

ШАҲОДАТНОМА
дар бораи бақайдгирии давлатии захираи иттилоотӣ

СВИДЕТЕЛЬСТВО
о государственной регистрации информационного ресурса

Номгуй / Наименование: База данных αβ-кодов словоформ для определения автора незнакомого текста
Сарзамин / Страна: Республика Таджикистан
Доранда / Владелец: Усманов Зафар Дж., Косимов Абдунаби А., Каюмов Махмадзоир М.
Таҳиягар / Разработчик: Усманов Зафар Дж., Косимов Абдунаби А., Каюмов Махмадзоир М.
№ кайди давлатӣ / № государственной регистрации: № 1202100478
Ба Феҳристи давлатии захираҳои иттилоотии Ҷумҳурии Тоҷикистон дохил карда шудааст
Внесен в Государственный реестр информационных ресурсов Республики Таджикистан: 07 июня 2021 г.

Директор: Исмоилзода М.Х

Рисунок П1.7. – База данных *αβ*-кодов словоформ для определения автора незнакомого текста

МАҚОМОТИ ИҶРОИЯИ ҲОКИМИЯТИ ДАВЛАТИИ
ВИЛОЯТИ СУҒД
РАЁСАТИ КОР БО ҶАВОНОН ВА ВАРЗИШ

ИФТИХОРНОМА

барои сазовор шудан ба ҷойи I дар озмуни вилоятии «Барномасози бехтарин» байни ҷавонон бахшида ба «Соли ҷавонон» сарфароз гардонида шуд.

Сардори Раёсат О.Зоҳидзода

ш.Хуҷанд,
25 ноябри соли 2017

Рисунок П1.8. – Занял первое место и стал обладателем конкурса «Лучший программист» Согдийской области

АКАДЕМИЯИ МИЛЛИИ
ИЛМҲОИ ТОҶИКИСТОН
ИНСТИТУТИ ЗАБОН ВА АДАБИЁТИ
БА НОМИ РӮДАКӢ

НАЦИОНАЛЬНАЯ АКАДЕМИЯ
НАУК ТАДЖИКИСТАН
ИНСТИТУТ ЯЗЫКА И ЛИТЕРАТУРЫ
ИМЕНИ РУДАКИ

734025, ш.Душанбе, хиёбони Рӯдакӣ -21
тел.:227-29-07, 227-75-50 www. iza. tj

734025, г.Душанбе, пр. Рудаки -21
тел.:227-29-07, 227-75-50 www. iza. tj

№ 1038/22 «06» 06 2022

"Тасдиқ мекунам"
директори Институти
забон ва адабиёти ба
номи Рудакии АМИТ
Шарифзода Б.Б.
"06" 06 2022

АКТ

оид ба татбиқи комплекси барномаҳо ва натиҷаҳои кори диссертатсияи доктории
Қосимов Абдунаби Абдурауфович
дар мавзуи "Қонуниятҳои омории шинохти ҳамгунии матн бо истифода аз γ-таснифгар
(Статистические закономерности распознавания однородности текстов с помощью
γ-классификатора)"

Акти зерин дар он хусус тартиб дода мешавад, ки дар ҳақиқат комплекси барномаҳо ва натиҷаҳои кори диссертатсияи Қосимов Абдунаби Абдурауфович дар мавзуи "Қонуниятҳои омории шинохти ҳамгунии матн бо истифода аз γ-таснифгар" дар Институти забон ва адабиёти ба номи Рӯдакии АМИТ татбиқ ва истифода бурда мешавад. Комплекси барномаҳои мазкур ҳам аз нигоҳи забоншиносии компютерӣ, ҳам аз нигоҳи адабиётшиносӣ хеле муҳим аст. Инчунин барои ёрии амалӣ расонидан ба муҳаққиқон ва пажӯҳишгарони соҳаи забону адабиёт нигаронида шуда барои муайян кардан ва мушаххас намудани сабки нигориши ҳар як муаллиф, хусусиятҳои хоси асарҳои алоҳидаи муаллифони гуногун, истифодаи чандомади ҳарф, ҳичо, калима, ибора, таркиби калима дар асарҳои алоҳида пешбинӣ гардида, дар ҳалли масъалаҳои мазкур ба мутахассисон ба хусус донишҷӯён, магистрантон, докторантон ва унвонҷӯён дар навиштани рисолаҳои илмӣ ва таҳқиқотҳои гуногун кӯмак расонида истодааст.

Ҳамчунин татбиқи натиҷаҳои кори илмии мазкур барои шоирону нависандагон дар амри омӯзиши осори адибони тоҷик ва шинохти ҳунару неруи сухандонӣ ва суханофариниии онҳо ёрии амалӣ расонида метавонад.

Барои омӯзгорони мактабҳои олӣ ва таҳсилоти умумии кишвар низ натиҷаҳои барнома дастуру воситаи хубе барои таълими адабиёт ва забони асарҳои адибон метавонад бошад.

Аз ин рӯ, комплекси барномаҳои тартиб дода шуда, барои шахсоне, ки дар соҳаи забоншиносии компютерӣ ва таҳлили асарҳои шоирону нависандагони тоҷик сару кор доранд, хеле зарур аст. Лоиҳа ба аспирантон, унвонҷӯён, докторантони PhD ва мутахассисони забону адабиёти тоҷик судманд буда, коркардӣ бисёр аз мавзуъҳои матншиносӣ ва забоншиносии миллиро барои истифодабарандагон осон мегардонад.

Ҳамчунин бояд қайд кард, ки комплекси барномаҳои тартиб дода шуда, дар ичрои амалӣ гардидани амри Пешвои миллат, Асосгузори сулҳу вахдати миллӣ, Президенти Ҷумҳурии Тоҷикистон муҳтарам Эмомалӣ Раҳмон эълон гардидани солҳои 2020-2040 "Бистсолаи омӯзиш ва рушди илмҳои табиатшиносӣ, дақиқ ва риёзӣ дар соҳаи илм ва маориф" аз 31.01.2020 № 1445 қадами хубе ба ҳисоб меравад.

Муовини директор
оид ба илм ва таълим
номзади илмҳои филологӣ **Муҳаммадиев Ш.**

Рисунок П1.9. – Акт о внедрении результатов исследования (Институт языка и литературы имени Рудаки)

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
РЯЗАНСКИЙ ГОСУДАРСТВЕННЫЙ РАДИОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

ДИПЛОМ

НАГРАЖДАЕТСЯ

Косимов
Абдунаби Абдурауфович

за участие во
Всероссийской научно-практической
конференции с международным участием
«Информационный обмен
в междисциплинарных исследованиях»

18–20 октября 2022 года
Россия · Рязань · РГРТУ

Проректор по научной работе
и инновациям РГРТУ С.И. Гусев

Рисунок П1.10. – Получение диплома за участие во Всероссийской научно-практической конференции

АКАДЕМИЯИ МИЛЛИИ
ИЛМҲОИ ТОҶИКИСТОН
Институти математикаи
ба номи А.Ҷӯраев

НАЦИОНАЛЬНАЯ АКАДЕМИЯ
НАУК ТАДЖИКИСТАНА
Институт математики
им. А. Джураева

734063, ш. Душанбе, к. Айнӣ 299/4, тел: (99237)2258089, URL: http://www.mintas.tj, e-mail: info@mitas.tj

№ 31004/23-№13 аз 18.01.2023 с.

САНАД

оид ба татбиқи комплекси барномаҳо ва натиҷаҳои кори диссертатсияи доктории Қосимов Абдунаби Абдурауфович
дар мавзӯи "Қонуниятҳои омории шинохти якҷинсагии матн бо истифода аз γ-таснифгар (Статистические закономерности распознавания однородности текстов с помощью γ-классификатора)"

Дар Институти математикаи ба номи А.Ҷӯраеви Академияи милли илмҳои Тоҷикистон дар самти лингвистикаи компютерӣ коркарди комплекси барномаҳои компютерӣ пайгирӣ карда шуданд. Комплекси барномаҳо барои рушди забони тоҷикӣ бо истифодаи имкониятҳои технологияҳои иттилоотӣ равона шуда, ба амал даровардани онҳо бо истифодаи васеи моделҳои математикӣ ва сатҳи баланди барномасозӣ ба даст оварда шудаанд. Санади зерин дар он хусус тартиб дода мешавад, ки дар ҳақиқат комплекси барномаҳо ва натиҷаҳои кори диссертатсияи Қосимов А.А. дар мавзӯи "Қонуниятҳои омории шинохти якҷинсагии матн бо истифода аз γ-таснифгар" дар Институти математикаи ба номи А.Ҷӯраеви Академияи милли илмҳои Тоҷикистон татбиқ ва истифода бурда мешавад. Комплекси барномаҳои мазкур ҳам аз нигоҳи забоншиносии компютерӣ, ҳам аз нигоҳи адабиётшиносӣ хеле муҳим аст. Инчунин барои расонидани ёрии амалӣ ба муҳаққиқон ва пажӯҳишгарони соҳаҳои забону адабиёт, математика ва технологияи иттилоотӣ нигаронида шуда, барои муайян ва мушаххас намудани сабки нигориши ҳар як муаллиф, хусусиятҳои хоси асарҳои алоҳидаи муаллифони гуногун, истифодаи чандомади ҳарф, ҳиҷо, калима, ибора, таркиби калима дар асарҳои алоҳида, сохтани моделҳои математикии гуногун пешбинӣ шудааст, инчунин бевосита масъалаҳои зикршуда ба мутахассисон ба хусус донишҷӯён, магистрантон, докторантон ва унвонҷӯён дар навиштани рисолаҳои илмӣ ва таҳқиқоти гуногун кӯмак расонида истодааст.

Институти математикаи ба номи А.Ҷӯраеви Академияи милли илмҳои Тоҷикистон гуфтаҳои дар боло зикр шударо тасдиқ намуда, ҳиссаи бевоситаи Қосимов А.А.-ро дар раванди пешбарии технологияҳои иттилоотӣ ва рушди забони тоҷикӣ баҳои баланд медиҳад.

Директор, академик Раҳмонов З.Ҳ.

Рисунок П1.11. – Акт о внедрении результатов исследования (Институт математики имени А.Джураева)

Рисунок П1.12. – Получение удостоверения и диплома 100-успешных людей в сфере IT-технологий и искусственного интеллекта.

Рисунок П1.13. – “Лучший педагог” 2023 года среди стран СНГ.

«УТВЕРЖДАЮ»
Начальник Академии МВД
Республики Таджикистан,
доктор юридических наук, доцент,
генерал-майор милиции
Ф. Р. Шарифзода
«13» 06 2023 г.

АКТ

о внедрении результатов докторского диссертационного исследования Косимова Абдунаби Абдурауфовича на тему «Статистические закономерности распознавания однородности текстов с помощью γ-классификатора (Қонуниятҳои омории шинохти якҷинсагии матн бо истифода аз γ-таснифгар)» в образовательную и научно-исследовательскую деятельность Академии МВД Республики Таджикистан

Экспертная комиссия в составе: председателя – начальника учебного отдела Академии МВД Республики Таджикистан, полковника милиции С.С. Саидзода; заместителя председателя – начальника организационно-научного и редакционно-издательского отдела Академии МВД Республики Таджикистан, кандидата юридических наук, майора милиции С.З. Заробидинзода, и членов комиссии: начальника факультета № 2 Академии МВД Республики Таджикистан, кандидата юридических наук, доцента, подполковника милиции А.М. Мансурзода и начальника кафедры информационной безопасности и цифровых технологий, кандидата юридических наук, полковника милиции А.Л. Арипова изучив содержание результатов докторского диссертационного исследования Косимова Абдунаби Абдурауфовича на тему: «Статистические закономерности распознавания однородности текстов с помощью γ-классификатора (Қонуниятҳои омории шинохти якҷинсагии матн бо истифода аз γ-таснифгар)» составила настоящий акт о том, что исследования проведенные автором и его результаты являются актуальными и представляют интерес для правоохранительной практики.

Задача распознавания текста возникла, по-существу, одновременно с зарождением письменности. В течение длительного времени она экспонировалась лишь одной своей гранью – необходимостью определения исполнителя письменного произведения.

Рисунок П1.14. – Акт о внедрении результатов исследования (Академия Министерства внутренних дел Республики Таджикистан)

Различные способы идентификации авторства текста являются на сегодняшний день эффективными инструментами в криминалистике для разрешения вопросов о спорном авторстве, плагиате, установления авторства анонимных текстов, пола автора, психологического портрета и т.д.

Программный комплекс может быть успешно использован в Академии МВД Республики Таджикистан для выявления автора анонимных текстов, заимствований, а также применятся для решения ряда смежных задач: идентификации пола и гендера, профессии, национальности, уровня образования автора и подобных вопросов.

В итоге, экспертная комиссия пришла к выводу о целесообразности использования результатов докторского диссертационного исследования Косимова Абдунаби Абдурауфовича на тему «Статистические закономерности распознавания однородности текстов с помощью γ-классификатора (Қонуниятҳои омории шинохти якҷинсагии матн бо истифода аз γ-таснифгар)» в образовательную и научно-исследовательскую деятельность Академии МВД Республики Таджикистан.

Председатель комиссии С. С. Саидзода

Заместитель председателя комиссии С. З. Заробидинзода

Члены комиссии А. М. Мансурзода

А. Л. Арипов

Рисунок П1.15. – Акт о внедрении результатов исследования (Академия Министерства внутренних дел Республики Таджикистан)