**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

**А В Т О Р Е Ф Е Р А Т**

диссертации на соискание ученой степени доктора технических наук по специальности **05.13.11** – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»

Душанбе – 2024

Диссертация выполнена в отделе математического моделирования Института математики имени А. Джураева Национальной академии наук Таджикистана и на кафедре автоматизированных систем управления Таджикского технического университета имени академика М.С. Осими.

**Научный консультант:** **Усманов Зафар Джураевич**, доктор физико-математических наук, академик НАНТ, профессор,

**Официальные оппоненты:** **Пруцков Александр Викторович**, доктор технических наук, Федеральное государственное бюджетное образовательное учреждение высшего образования «Рязанский государственный радиотехнический университет», профессор кафедры «Вычислительная и прикладная математика»

**Одинаев Раим Назарович**, доктор физико-математических наук, профессор, заведующего кафедрой математического и компьютерного моделирования механико-математического факультета Таджикского национального университета

**Рахимов Нодир Одилович,** доктор технических наук, профессор, заведующего кафедрой программного обеспечения информационных технологий Ташкентского университета информационных технологий имени Мухаммада ал-Хоразми Республики Узбекистан

**Ведущая организация:** Межгосударственное образовательное учреждение высшего профессионального образования «Российско-Таджикский (Славянский) университет»

Защита состоится 20 сентября 2024 г. в 14:00 часов на заседании разового диссертационного совета 6D.KOA-049 при Таджикском техническом университете имени академика М.С. Осими, г. Душанбе, проспект академиков Раджабовых, 10.

С диссертацией можно ознакомиться в библиотеке Таджикского технического университета имени академика М.С. Осими и на официальном сайте университета: https://web.ttu.tj/tj/elonho/78

Автореферат разослан «___» _____ 2024 года

Отзывы на автореферат в двух экземплярах, подписанные и заверенные печатью учреждения, просим направлять по адресу: 734042, г. Душанбе, проспект академиков Раджабовых, 10, тел.: (+992 37) 227-37-81, e-mail: sultonzoda.sh@mail.ru

Ученый секретарь
диссертационного совета,
кандидат технических наук, доцент **Султонзода Ш.М.**

## ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ[1]

**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 современной цивилизации контролировать, упорядочивать, осмысливать и использовать лавинообразный приток знаний, порождаемый её собственной деятельностью.

Одной из граней этой проблемы является проектирование автоматических систем распознавания новизны и адресности информации, охватывающих такие вопросы, как компиляция, плагиат, заимствование, идентификация авторства, сходство 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 (цифровой портрет) на основе распределения частотности различных алфавитных элементов текста для распознавания новизны, компиляции, плагиата, заимствования, идентификации

[1] В автореферате используется нумерация параграфов, формул, таблиц, рисунков и т.п. в соответствии с обозначениями, принятыми в диссертации.

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.

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.

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.

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.

**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];

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,

[2] Усманов, З.Д. Классификатор дискретных случайных величин // ДАН РТ, 2017, Т.60, №7-8, С. 291-300 и Алгоритм настройки кластеризатора дискретных случайных величин // ДАН РТ, 2017, Т.60, №9, С. 392-397.

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.

**Научная новизна** диссертации состоит в следующем:

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

2) установлена статистическая эффективность $\pi$ математической модели опознавания авторов произведений таджикской классической поэзии ($\pi = 1.00$) на основе триграмм, современной поэзии ($\pi = 0.98$) с помощью униграмм и современной прозы ($\pi = 0.96$) 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.

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 научно-исследовательских базах Института математики имени А. Джураева НАНТ и в Таджикском техническом университете имени академика М.С. Осими. 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.

## ОСНОВНАЯ ЧАСТЬ ИССЛЕДОВАНИЯ

**Материал и методы исследования.** Материал исследования посвящен изучению проблемы распознавания однородности текстовых фрагментов. Признаки однородности, так как проблема связана с текстом, рассматриваются следующие: распознавание авторства, тематики текста, язык, группа языков, оригинал и его перевод, стиль произведений и шифры научных работ.

В работе для распознавания однородных текстов используются математические модели принятия решений, среди которых особо успешными являются нейронные сети, машина опорных векторов, метод ближайшего (по расстоянию) соседа и недавно разработанный в Институте математики имени А. Джураева НАНТ γ-классификатор.

**Результаты исследования.** Приводим краткое изложение результатов глав диссертационной работы.

Во **Введении** обосновывается актуальность темы диссертационной работы, формулируются цели и задачи диссертации, научная проблема, определяются объект, предмет, методы исследования, излагаются основные положения, выносимые на защиту, научная новизна, теоретическая и практическая значимость исследования, приводятся сведения об апробации работы.

В **главе 1** «*Основные понятия и определения*» 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.

В **§1.1** сообщается обзор литературы (статей и публикаций) по автоматическому распознаванию однородности текста. 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[3] используется около 1000 групп характеристик 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 [4] и Р.В. Мещерякова представлен обширный обзор работ по распознаванию однородности текста на основе разнообразных ЦП текстов и

[3] Rudman, J. The state of authorship attribution studies: Some problems and solutions // Computers and the Humanities, 1998, Vol. 31, pp. 351-365.

[4] Романов, А.С., Шелупанов, А.А., Мещеряков, Р.В. Разработка и исследование математических моделей, методик и программных средств информационных процессов при идентификации автора текста // -В-Спектр, Томск, 2011, 188 с.

применяемых методов классификации. В следующей подглаве формулируется постановка задачи, решаемой в настоящей диссертации.

В **§ 1.2** описана постановка задач, решение которых формирует полное представление об эффективности применения γ-классификатора для распознавания однородности произведения. 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.

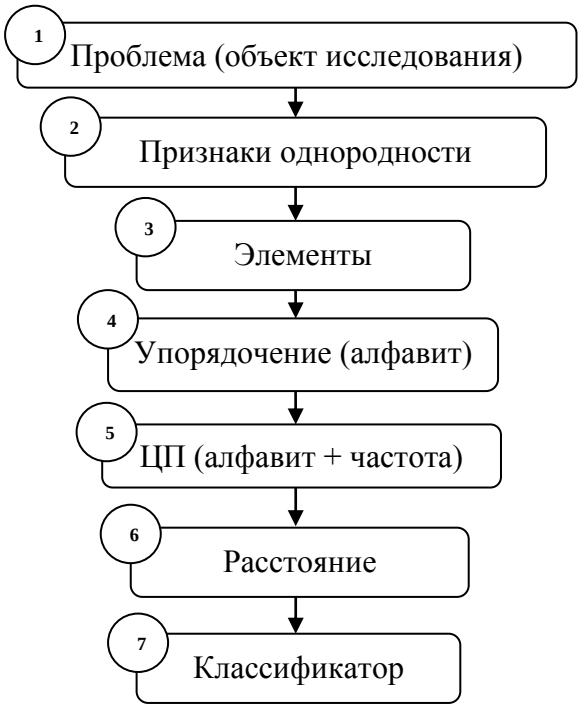



**Рисунок 1.1. – Описание проблемы**

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. В следующих **§§ 1.3** и **1.4** дается описание существа γ-классификатора.

В **§ 1.3** вводятся терминология и понятие, которые используются при описании математической модели текста.

**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 – необозримое множество. 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:

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

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

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

**1.3.3. Расстояния между ЦП текстов.**

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

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

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*

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

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.

В **§ 1.4** дается описание используемого в работе метода принятия решения с помощью так называемого γ-классификатора, см. сноску 2 на стр. 3.

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 $\gamma$ – некоторое положительное число.

**Определение 1.4.1.** *Тексты* $T_1$, $T_2$ *называются* ***$\gamma$-однородными****, если*

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

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

Очевидно, что от значения $\gamma$ 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 $\gamma$ *показатель выполнения гипотезы будет определяться величиной* $\pi$, 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 $\pi$ 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 $\gamma$, представляет интерес найти такое его значение, при котором $\pi$ 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**.**

**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 $\gamma$-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 ЦП текста и γ-классификатора среди сколь угодно большого числа текстов.

В **главе** 2 «*Исследование эффективности распознавания однородности текстов на примерах модельных коллекций художественных произведений*» мы обращались к модельной коллекции, составленной их 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 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).

В **главе 3** «*Распознавание признаков однородности*» 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.

В **§ 3.1** 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. 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, Н. Островского, Б. Полевой, К. Симонова, А. Фадеева, Д. Фурманова, А.С. Серафимовича и Ф.Д. Крюкова, представляемые девятью различными ЦП.

В **§ 3.2** на примере модельной коллекции текстов устанавливается применимость γ-классификатора для автоматического распознавания языка произведения на основе частотности алфавитных букв. В **§ 3.2.5** 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 ЦПТ – распределение в тексте частотности буквенных униграмм; вторым компонентом служит формула для вычисления расстояний между ЦП текстов 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 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*).

Вычисления по формулам (1.1) – (1.4) 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** | |

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 $\gamma$ оказывается в пределах

$$\gamma^{onm} \in [0.0833;\ 0.1353); \qquad (3.47)$$

в соответствии с определением **1.3.4** это значит, что если расстояние $\rho(T_1, T_2)$ между двумя текстами не превосходит значение $\gamma^{onm}$ из указанного полуинтервала, то пара текстов принадлежит одному и тому же языку (соответствующие расстояния в таблице помечены серым цветом); если же превосходит, то принадлежит разным языкам (соответствующие расстояния оставлены непомеченными);

– отметим, что для всех (без исключения) произведений коллекции ***C*** полностью подтвердилась гипотеза $\mathrm{H}$ и её математическая интерпретация в виде определения **1.3.4**, и потому получено

$$\tau = \tau_{\min} = 0,$$

то есть, ни одно из неравенств (1.5) и (1.6) не было нарушено;

– вследствие чего показатель эффективности предложенной в настоящей работе математической модели распознавания языка произведений оказался равным

$$\pi = \pi_{\max} = 1.$$

**Тестирование.** Итак, настройка (обучение) $\gamma$ -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*).

Для шести произведений, предназначенных для тестирования, построены цифровые портреты (1.1) и затем по формулам (1.2), (1.3), (1.4) 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[5] 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 $\gamma^{onm}$, см. (3.47). Следовательно, как и ожидалось, для Text_Ro в коллекции не оказалось ни одного однородного объекта. Интересно, однако, отметить, что $\gamma$-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, $\gamma$-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 (1.1) – (1.4) 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 $\gamma$-классификатора не только для распознавания языков, но также и для самых разнообразных однородностей текстовых документов. Аналогичный результат получен в **§§ 3.2.1 – 3.2.5** для языков в кириллической графике, а также латинском шрифте, но для других модельной коллекции текстов.

[5] Воронцов, К.В. Математические методы обучения по прецедентам // [Электронный ресурс] – Режим доступа: http://www.ccas.ru/voron и Дьяконов, А.Г. Анализ данных, обучение по прецедентам, логические игры, системы WEKA, RapidMiner и MatLab (Практикум на ЭВМ кафедры математических методов прогнозирования) // Учебное пособие, М.: Издательский отдел факультета ВМК МГУ имени М.В. Ломоносова, 2010, 278 с.

В **§ 3.2.6** 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.

В **§ 3.3** 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.

В **§ 3.4** определяется 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. 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 ***A-коллекцией (модельной).***

Расчет по формулам (1.1) – (1.4) 45 парных расстояний $\rho\,(T_1, T_2)$ между авторефератами коллекции $\boldsymbol{A}$ (результаты расчетов приведены в следующей таблице 3.31).

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»;

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** | |

– половина оптимального интервала значений $\gamma$ находится в пределах

$$\gamma^{\text{опт}} \in [0.0610;\ 0.0614);$$

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

$$\pi = \pi_{max} = 0.87$$

**Тестирование.** Итак, результаты предыдущего раздела показывают, что настройка (обучение) γ-классификатора на 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 (1.4), была получена следующая таблица расстояний от каждого протестированного автореферата 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** |
|---|---|---|---|---|---|---|
| | | З | З | З | З | З |
| **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.** $\gamma$-классификатор с фиксированным значением $\gamma = \gamma^{\text{опт}}$ был протестирован на случайных выборках авторефератов и подтвердил 87%-ую способность к распознаванию шифра специальности авторефератов.

В **§ 3.5** 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. В **заключение** приведем следующую таблицу, в которой показано всего лишь 4 стиля из 7. 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).

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** | |

Для такого сочетания авторов со своими произведениями получено значение $\gamma \in [2.8324; 2.8385)$, 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.

В **главе 4** «*Исследование статистических закономерностей распознавания однородных текстов в корпусах художественных литературных произведений*» исследуется подобная задача не в модельной коллекции текстов, а в корпусах произведений художественной литературы, для которой удастся ли получить удовлетворительный результат решения рассматриваемой задачи.

В **§ 4.1** 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. 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 на тех же языках, что и тексты коллекции. Методом ближайшего (по расстоянию) соседа три случайных текста проверяются

на однородность с соответствующими парами одноязычных произведений. В процессе настройки выполняются следующие операции:

– предобработка экспериментальных данных путём удаления из всех произведений коллекции дополнительных сверх кириллического алфавита букв;

– вычисление ЦП (1.1) (частотности 26 кириллических букв) для всех 70 произведений коллекции ***C***;

– вычисление по формулам (1.2), (1.3) и (1.4) 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 |

По данным таблицы 4.1 получены следующие результаты:

– оптимальный полуинтервал значений $\gamma$ оказывается в пределах

$$\gamma^{onm} \in [0.1001; 0.1638); \qquad (4.7)$$

в соответствии с определением **1.4.1** это значит, что если расстояние $\rho(T_1, T_2)$ между двумя текстами не превосходит значение $\gamma^{onm}$ из указанного полуинтервала, то пара текстов принадлежит к одному и тому же языку; если же превосходит, то принадлежит к разным языкам;

– наивысшее значение $\pi$=100% коэффициента эффективности распознавания языка текста реализуется на корпусах 5 языков с 10 текстами;

– коэффициент $\pi$ 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 $\pi$ эффективности равен 98% и 99% при выборе корпуса текстов 10, 20 языков с 20, 40 текстами.

Таким образом, математическая триада в составе ЦП текстов, представляемых распределениями частотности 26 кириллических букв, формул (1.1)-(1.4) 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 $\gamma$ – классификатора не только для распознавания языков, но также и для самых разнообразных однородностей текстовых документов.

В **§ 4.2** 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.

В параграфе **4.3** 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 10, 20, 40 текстов, 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.

В **§ 4.4** 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.

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 поэмы «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 $\rho((\text{АН}),(\text{ШШ})) = 0.4021$.

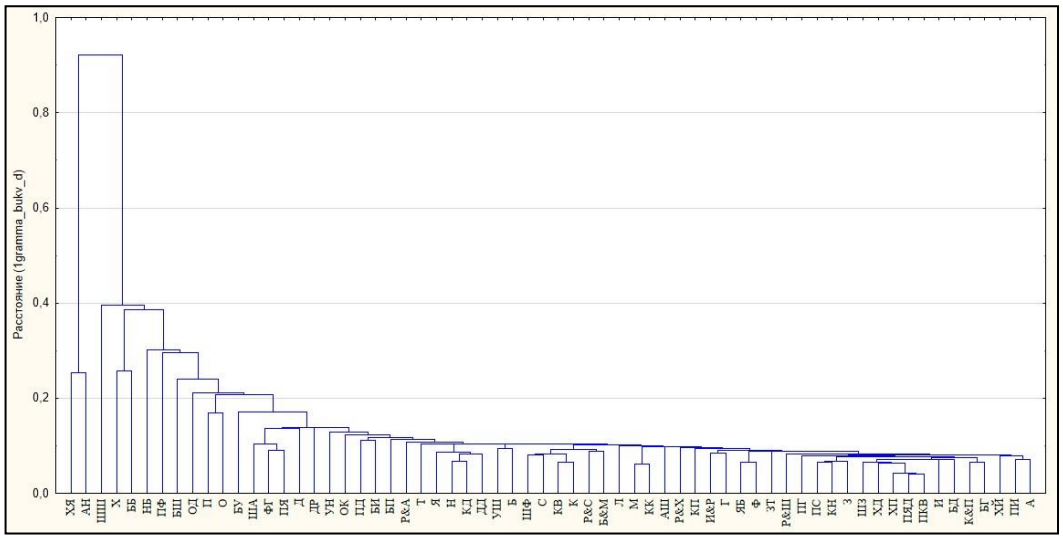



Рисунок 4.1. – Результаты иерархической классификации поэм в виде дендрограммы

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 $\rho = 0.0851$.

В **§ 4.5** 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 **«TTA»** (tajik text author) для определения автора текста устанавливается. 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 слов) | 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 слов) | 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 |

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 **«ТТА»** с большой вероятностью смог предсказать А. Фирдоуси как автора этих текстов.

В **главе 5** «*Исследование влияния порядка ЦП текста на распознавание однородности произведения*» на примере модельной коллекции, количественные описания произведений которой основываются на различных вариантах упорядочения элементов алфавита, выявляются особенности применения γ-классификатора при распознавании автора текста.

В **§ 5.1** 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.

**5.1.4.** Множества -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 $ABC$)[6];

2) элементы располагаются в порядке, обратном алфавитному с пробелом в качестве первого элемента алфавита (обозначается как $CBA$)[7];

3) элементы располагаются в порядке убывания их частотности в тексте (обозначается символом « ↘»);

4) элементы располагаются в порядке возрастания их частотности в тексте (обозначается символом « ↗»).

Итоги автоматической обработки модельной коллекции текстов показаны в таблицах 5.1-5.3.

Таблица 5.1. – Значения $\pi$ и $\gamma$ для произведений классической поэзии

| Элементы текста | Число элементов алфавита | Порядок элементов алфавита | $\pi$-эффективность распознавания автора | Оптимальный $\gamma$-полуинтервал |
|---|---|---|---|---|
| униграммы | 36 | A B C | 0.98 | [0.0354; 0.0447) |
| | | C B A | 0.98 | [0.0354; 0.0447) |
| | | по ↘ | 0.98 | [0.0337; 0.0342) |
| | | по ↗ | 0.98 | [0.0337; 0.0342) |
| биграммы | 1296 | A B C | 0.98 | [0.2987; 0.3551) |
| | | C B A | 0.98 | [0.2987; 0.3551) |
| | | по ↘ | 0.96 | [0.2065; 0.2212) |
| | | по ↗ | 0.96 | [0.2065; 0.2212) |
| триграммы | 46656 | A B C | 1.00 | [2.1630; 2.1648) |
| | | C B A | 1.00 | [2.1630; 2.1648) |
| | | по ↘ | 0.96 | [1.2426; 1.4051) |
| | | по ↗ | 0.96 | [1.2426; 1.4051) |

[6] Для биграмм и триграмм – с двумя и тремя пробелами в конце.
[7] Для биграмм и триграмм – с двумя и тремя пробелами в начале.

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.**

Таблица 5.2. – Значения $\pi$ и $\gamma$ для произведений современной поэзии

| Элементы текста | Число элементов алфавита | Порядок элементов алфавита | $\pi$-эффективность распознавания автора | Оптимальный $\gamma$-полуинтервал |
|---|---|---|---|---|
| униграммы | 36 | A B C | 0.98 | [0.0268; 0.0423) |
| | | C B A | 0.98 | [0.0268; 0.0423) |
| | | по ↘ | 0.98 | [0.0384; 0.0415) |
| | | по ↗ | 0.98 | [0.0384; 0.0415) |
| биграммы | 1296 | A B C | 0.98 | [0.2318; 0.2816) |
| | | C B A | 0.98 | [0.2318; 0.2816) |
| | | по ↘ | 0.98 | [0.2484; 0.2745) |
| | | по ↗ | 0.98 | [0.2484; 0.2745) |
| триграммы | 46656 | A B C | 0.98 | [1.3885; 1.7054) |
| | | C B A | 0.98 | [1.3885; 1.7054) |
| | | по ↘ | 0.98 | [1.5556; 1.6453) |
| | | по ↗ | 0.98 | [1.5556; 1.6453) |

Таблица 5.3. – Значения $\pi$ и $\gamma$ для произведений современной прозы

| Элементы текста | Число элементов алфавита | Порядок элементов алфавита | $\pi$-эффективность распознавания автора | Оптимальный $\gamma$-полуинтервал |
|---|---|---|---|---|
| униграммы | 36 | A B C | 0.96 | [0.0285; 0.0336) |
| | | C B A | 0.96 | [0.0285; 0.0336) |
| | | по ↘ | 0.91 | [0.0165; 0.0236) |
| | | по ↗ | 0.91 | [0.0165; 0.0236) |
| биграммы | 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) |
| триграммы | 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) |

**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 $\pi$ эффективности на основе порядков ABC и CBA расположения $N$-грамм ($N = 1, 2, 3$) равны;

3) значения коэффициентов $\pi$ эффективности на основе порядков

расположения $N$-грамм ($N = 1, 2, 3$) по убыванию (↘) или возрастанию (↗) также равны;

4) значение коэффициента $\pi$ эффективности на основе порядка ABC и CBA расположения $N$-грамм ($N = 1, 2, 3$) не ниже значения, основанного на порядке расположения $N$-грамм ($N = 1, 2, 3$) по убыванию (↘) или возрастанию (↗);

5) коэффициент $\pi$ 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 $\pi$ для произведений современной прозы несколько ниже аналогичных значений для произведений классической и современной поэзии;

7) полуинтервалы оптимальных значений $\gamma$ 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 $\pi$ эффективности $\gamma$-классификатора, а также и полуинтервалы оптимальных значений $\gamma$. В **§§ 5.2.-5.4.** исследуются другие допустимые варианты.

В **§§ 6.1 – 6.7 главы 6** дано подробное описание программного комплекса **«THR»** (text homogeneity recognition), предназначенного для распознавания однородности текста.

В **заключении** диссертационной работы приведены основные выводы и результаты.

## 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 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.

**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.

## СПИСОК ПУБЛИКАЦИЙ АВТОРА ПО ТЕМЕ ДИССЕРТАЦИИ

### В журналах, рекомендованных ВАК:

**[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 шифра 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.

**АКАДЕМИЯИ МИЛЛИИ ИЛМҲОИ ТОҶИКИСТОН**
Институти математикаи ба номи А. Ҷӯраев

**ВАЗОРАТИ МАОРИФ ВА ИЛМИ ҶУМҲУРИИ ТОҶИКИСТОН**
Донишгоҳи техникии Тоҷикистон ба номи академик М.С. Осимӣ

---

ТДУ 811::81’33::519.25

*Ба ҳуқуқи дастнавис*

**ҚОСИМОВ Абдунаби Абдурауфович**

# ҚОНУНИЯТҲОИ ОМОРИИ ШИНОХТИ ЯКҶИНСАГИИ МАТН БО ИСТИФОДА АЗ γ-ТАСНИФГАР

**А В Т О Р Е Ф Е Р А Т И**

диссертатсия барои дарёфти дараҷаи илмии доктори илмҳои техникӣ аз рӯйи ихтисоси **05.13.11** – “Таъминоти математикӣ ва барномавии мошинҳои ҳисоббарор, муҷтамаъҳо ва шабакаҳои компютерӣ”

Душанбе – 2024

Диссертатсия дар шуъбаи тарҳрезии математикии Институти математикаи ба номи А. Ҷӯраеви Академияи миллии илмҳои Тоҷикистон ва дар кафедраи системаҳои автоматикунонидашудаи идоракунии Донишгоҳи техникии Тоҷикистон ба номи академик М.С. Осимӣ омода гардидааст.

**Мушовири илмӣ:** **Усмонов Зафар Ҷураевич**, доктори илмҳои физика ва математика, академики АМИТ, профессор,

**Муқарризони расмӣ:** **Пруцков Александр Викторович,** доктори илмҳои техникӣ, Муассисаи давлатии буҷетии федералии таҳсилоти олии касбии “Донишгоҳи давлатии радиотехникии Рязан”, профессори кафедраи «Математикаи амалӣ ва ҳисобӣ»

**Одинаев Раим Назарович**, доктори илмҳои физикаю математика, профессор, мудири кафедраи “Моделсозии математикӣ ва компютерӣ”-и Донишгоҳи миллии Тоҷикистон

**Раҳимов Нодир Одилович,** доктори илмҳои техникӣ, профессор, мудири кафедраи “Таъминоти барномавии технологияҳои иттилоотӣ”-и Донишгоҳи технологияҳои иттилоотии Тошкент ба номи Муҳаммад ал-Хоразмӣ, Ҷумҳурии Узбекистон

**Муассисаи пешбар:** Муассисаи таълимии таҳсилоти касбии олии байнидавлатии Донишгоҳи славянии Русияву Тоҷикистон

Ҳимояи диссертатсия санаи 20 сентябри соли 2024, соати 14:00 дар ҷаласаи шурои диссертатсионии якдафъаинаи 6D.KOA-049 дар назди Донишгоҳи техникии Тоҷикистон ба номи академик М.С. Осимӣ, дар суроғаи шаҳри Душанбе, хиёбони академикҳо Раҷабовҳо, 10 баргузор мегардад.

Бо диссертасия ва автореферати он дар китобхона ва сомонаи расмии Донишгоҳи техникии Тоҷикистон (https://web.ttu.tj/tj/elonho/78) шинос шудан мумкин аст.

Автореферат “____” _____ соли 2024 тавзеъ шудааст.

Хоҳишмандем тақризҳоро нисбати автореферат дар ду нусха бо муҳри муассиса ба суроғаи зерин ирсол намоед: 734042, ш. Душанбе, хиёбони академикҳо Раҷабовҳо, 10 тел: (+992 37) 227-37-81, e-mail: sultonzoda.sh@mail.ru

Котиби илмии
шурои диссертатсионӣ
номзади илмҳои техникӣ, дотсент **Султонзода Ш.М.**

## ТАВСИФНОМАИ УМУМИИ ТАҲҚИҚОТ[1]

**Мубрамии мавзӯи таҳқиқот.** Рисолаи мазкур қисми ҷудонашавандаи масъалаҳои илмии ҷаҳонӣ – коркарди автоматии иттилоот бо забони табиӣ мебошад, ки яке аз масъалаҳои мубрами башарият, эътироф гардидааст. Бо умеди ҳалли бомуваффақияти охирин, масъалаи қобилияти тамаддуни муосир дар идора кардан, ба тартиб даровардан, дарк кардан ва истифода бурдани ҷараёни бузургӣ донишҳои бо фаъолияти худ тавлидшуда алоқаманд аст.

Яке аз паҳлуҳои ин мушкилот тарҳрезии низомҳои худкори шинохти навоварӣ ва ҳадафмандии иттилоот мебошад, ки масъалаҳои марбут ба талфифа, асардуздӣ, монандӣ, муайян кардани муаллиф, шабоҳати асар бо тарҷумаи он ва ғайраро дар бар мегирад. Дар робита ба рушди технологияҳои иттилоотӣ, тадқиқот дар ин соҳаи илм дар саросари ҷаҳон ба таври назаррас афзоиш ёфта истодааст. Нашрияҳои сершумори илмии ҳамаи мамлакатҳои ба дараҷаи олӣ тараққикарда роли махсуси ин масъаларо нишон медиҳад, аз он ҷумла таъсири бевоситаи онро ба инкишофи илму техника, ба пешрафти соҳаи низомҳои зеҳни сунъӣ ва татбиқи миқёсан калон дар иқтисодиёти ҷаҳон мебошад.

Маҳз муҳимияти масъалаи интихоби мавзуи рисолаи диссертатсионӣ ҳамин аст, ки дар Қарори Ҳукумати Ҷумҳурии Тоҷикистон “Дар бораи тасдиқи барномаи истифода ва рушди технологияи иттилоотӣ дар забони тоҷикӣ” аз 06.06.2005 №188 ва дар солҳои 2020-2040 “Бистсолаи омӯзиш ва рушди илмҳои табиатшиносӣ, дақиқ ва риёзӣ дар соҳаи илм ва маориф”, аз 31.01.2020 № 1445 ва Президенти Ҷумҳурии Тоҷикистон, Пешвои миллат, муҳтарам Эмомалӣ Раҳмон дар Паёми солонаи худ ба Маҷлиси Олӣ дар охири соли 2021 ба намояндагони Ҳукумат дастур доданд, ки технологияҳои муосирро дар бахшҳои гуногуни иқтисодиёти кишвар таҳия ва васеъ истифода намуда, Стратегияи миллии рушди Ҷумҳурии Тоҷикистон оиди “Зеҳни сунъӣ”-ро қабул ва татбиқ намоянд, 21-уми декабри 2021 сол.

**Дараҷаи таҳқиқи мавзӯи илмӣ.** Муҳимияти масъалаи илмии зикршударо корҳои назариявӣ ва амалии муҳаққиқони дохиливу хориҷӣ тасдиқ мекунад. Аҳамияти назариявии масъала бо омӯзиши маҷмӯи масъалаҳои марбут ба ташаккул ва тадқиқи мувофиқати СР (симои рақамӣ) дар асоси тақсимоти басомади унсурҳои алифбоиӣ гуногуни матн барои шинохти навоварӣ, талфифа, асардуздӣ, монандӣ, муайянкунии муаллиф ва рамзи асарҳои илмӣ алоқаманд мебошад. Муҳимияти ин гуна корҳо бо муайян кардани хусусиятҳои хоси матн алоқаманд буда, ки гарчанде таҳти назорати эҷодкори худ набошад ҳам, дар бораи услуби муаллиф ва ҳатто хислатҳои фардии муаллиф маълумоти бавосита доранд. Арзиши амалии масъала ба фаъолияти маъмурии давлатӣ вобаста аст, ки дар он коркарди автоматии иттилооти матнӣ ба мадди аввал меояд; дар криминалистика, ки ба муайян кардани ҷинояткор дар асоси ҷиноятҳояш ва муаллифони матнҳои анонимӣ лозиманд; ба соҳаи маориф ва илм, ки дар он ҳам донишҷӯён ва ҳам кормандони псевдо-илмӣ ҳангоми анҷом додани корҳои курсӣ ва дипломӣ ва

[1] Рақамгузории бобҳо, зербобҳо, формулаҳо, ҷадвалҳо, расмҳо ва ғайра дар автореферат мувофиқи рақамгузории онҳо дар диссертатсия истифода мешавад.

барои ба ҳимоя пешниҳод кардани рисолаҳои номзадӣ ва докторӣ аз талфифа, монандкунӣ ва асардуздӣ истифода мекунанд.

Дар ҳамин ҳол, дар хориҷи дур бошад, кор дар ин соҳа, дар робита бо рушди технологияҳои иттилоотӣ ба таври ҷиддӣ тақвият ёфт. Барои исботи ин далел ба корҳои 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 ва S. Waugh, рӯй овардан кофӣ аст.

Дар Русия ба чунин монанд кор тадқиқотҳои А.А. Шелупанов, Р.В. Мешеряков, А.С. Романов, А.В. Куртукова, А.В. Прутсков, Л.С. Ломакина, А.В. Мордвинов, А.С. Суркова, Д.В. Ломакин, А.З. Панкратова, В.Б. Родионов, С.С. Буденков, М.С. Сементсов, М.Д. Ломакина, А.А. Сарев, С.С. Скоринин, И.Д. Чернобаев, А.А. Домнин, В.В. Поддубного, В.П. Фоменко, Т.Г. Фоменко, Н.А. Морозов, А.А. Марков, Д.В. Хмелев, Е.И. Болшакова, А.А. Носков, О.В. Пескова, Е.В. Ягунова, В.В. Александров, Л.Л. Иомдин, М.В. Арапов, В.К. Финн, А.А. Барсегян, М.С. Куприянов, И.И. Холод, А.И. Башмаков, В.С. Белов, Г.Г. Белоногов, А.А. Хорошилов, Ю.Г. Зеленков, А.П. Новоселов, Б.А. Кузнетсов, М.Б. Болдин, Г.И. Симонова, Ю.Н. Тюрин, А.А. Болшаков, Р.Н. Каримов, А.А. Боровков, И.И. Бистров, Б.В. Тарасов, С.И. Радоманов, В.Н. Вапник, А.Я. Червоненкис, Н.К. Верешагин, В.Н. Волкова, А.А. Денисов, Т.А. Гаврилова, А.С. Дмитриев, А.П. Еремеев, Н.Г. Загоруйко, Л.А. Заде, М. Кендалл, А. Стюарт, А.Н. Кирдин, А.Ю. Новоходко, В.Г. Сарегородтсев, А.Н. Колмогоров, А.С. Костишин, В.Н. Кучуганов, И.В. Безсуднов, Д.В. Ландэ, Э. Леман, А.В. Леоненков, Н.Н. Леонтева, Н.В. Лукашевич, Г.Я. Мартиненко, А.С. Мелничук, Л.Н. Мурзин, А.С. Штерн, Г.В. Напреенко, В.А. Негуляев, А.А. Орлов, А.И. Орлов, А.А. Поликарпов, И.Н. Пономаренко, Д.М. Сибулко, А.П. Рижов, Ю.Б. Сафронова, И.П. Севбо, Э.Ф. Скороходко, Ю.Г. Сметанин, М.В. Улянов, А.С. Пестова, Г.Я. Солганик, В.М. Солнтсев, А.А. Харкевич, Г. Хетсо, Я.З. Сипкин, И.Г. Чекунов, А.А. Рогов, Ю.В. Сидоров, А.Ю. Комиссаров, Е.В. Шарапова, Р.В. Шарапов, О.Г. Шевелев, М.А. Марусенко, Ю.Н. Павлов, А.В. Седов, Е.А. Тихомирова, В.В. Дягилев, А.А. Схая, А.О. Шумская, С.В. Бутаков ва З.И. Резанова бахшида шудаанд.

Доир ба масъалаи шинохтӣ якҷинсагии матн дар Тоҷикистон З.Ҷ. Усмонов, Х.А. Тошхуҷаев, Х.Т. Мақсудов, М.А. Умаров, М.А. Исмоилов, Х.А. Худойбердиев, О.М. Солиев, Ш.Н. Ашӯрова, Г.М. Довудов, А.А. Каримов, М.М. Қаюмов, П.Э. Зулфиқорова, Ҷ.Х. Баҳовудинов, С.М. Пиров, Н.М. Қурбонов, М.Ё. Муҳсинзода, Н.О. Қосимова, О.А. Қосимов, Б.Б. Иномов, Д.Э. Қосимов, М.М. Фозилова, Ш.С. Саидов, Д.Н. Комилов ва К.С. Бахтеев, кор кардаанд ва кор карда истодаанд.

Ҳамаи ин гуфтаҳо муҳимияти мавзӯи интихобшудаи диссертатсияро ифода

мекунад, махсусан, бо он мақсад, ки тадқиқот дар чунин самтҳои муҳим дар Тоҷикистон аввалин бор оғоз меёбад ва мустақиман ба коркарди рушди системаи амнияти иттилоотии давлатӣ дар ояндаи наздик алоқаманд аст.

Диссертатсияи мазкур ба таҳқиқи масъалаи муайянкунии ҳамгунии порчаи матн бахшида шудааст.

**Мақсади кор** – Алгоритмизатсияи раванди муайянкунии якҷинсагии матн ва амалӣ намудани он дар намуди маҷмааи барномаи компютерӣ.

**Вазифаҳои таҳқиқот.** Барои ноил шудан ба мақсад чунин масъалаҳо матраҳ шуданд:

1) ду коллексияи матнҳои электрониро ташкил кардан, ки якум барои пешаки тестиронӣ ва дуюмин барои баҳодиҳии оянда истифодаи γ-таснифгар[2] пешбинӣ шудааст;

2) таҳқиқи симои рақамии матн (СРМ) барои шинохтани муаллифи матн;

3) муқаррар кардани самаранокии омории истифодаи γ-таснифгар барои муайян кардани муаллифони асарҳо;

4) муайян кардани андозаи ҳадди ақали матни ношинос, ки барои муайянкунии муаллифи он мувофиқ аст;

5) тафтиши самаранокии истифодабарии элементҳои баланд басомади СРМ барои муайян кардани муаллифи матн;

6) муқаррар намудани самаранокии омории истифодаи γ-таснифгар ва таҳқиқи лоиқии СР дар асоси тақсими басомади гуногуни алифбои элементи матн барои шинохти аломатҳои дигари якҷинсагии матн ба монандӣ: мавзӯи матн, забон, гурӯҳи забонҳо, асл ва тарҷумаи он, услуби асарҳо, рамзи асарҳои илмӣ ва ғайра;

7) таҳқиқи қонуниятҳои омории шинохтӣ якҷинсагии матнҳо дар пайкараҳои осори адабиёти бадеӣ;

8) муайян кардани самаранокии истифодаи γ-таснифгар барои шинохтӣ муаллифи асарҳои ба таври сунъӣ тавлидшудаи муаллифон;

9) тадқиқи таъсири тартиби СР матн ба шинохтӣ якҷинсагии матн бо истифода аз γ-таснифгар;

10) тарҳрезӣ ва татбиқи комплекси барномаҳои компютерӣ барои шинохт (муайян кардан)-и якҷинсагии матн дар асоси СР матнии гуногун ва γ-таснифгар.

**Объекти таҳқиқот** – пайкараи матнҳои чопӣ ва хусусиятҳои он дар забонҳои гуногун.

**Предмети таҳқиқот** – шинохтӣ якҷинсагии асар дар асоси γ-таснифгар (сегонаи математикӣ) ва басомади гуногуни хусусиятҳои матн.

**Навгониҳои илмии** диссертатсия чунин ифода мегарданд:

1) тадқиқи ахборотии аломатҳои ғайрианъанавӣ барои тавсифи миқдории матнҳои тоҷикӣ гузаронида шуд;

2) самаранокии омории амсилаи математикии $\pi$ дар муайянсозии муаллифони асарҳои назмии шоирони классикии тоҷик дар асоси триграммаҳо

[2] Усманов, З.Д. Классификатор дискретных случайных величин // ДАН РТ, 2017, Т.60, №7-8, С. 291-300 ва Алгоритм настройки кластеризатора дискретных случайных величин // ДАН РТ, 2017, Т.60, №9, С. 392-397.

(π=1.00), назми муосир бо истифода аз унираммаҳо (π=0.98) ва насри муосир аз рӯи тақсимоти дарозии ҷумла (дар калимаҳо) (π=0.96) муқаррар карда шуд;

3) самаранокии омории 100% ҳангоми тадбиқи ченаки γ-таснифгар ва усули ҳамсояи наздик (аз рӯи масофа) барои муайян кардани муаллифҳои асарҳо – пай дар пай камшавии порчаи матни ҳаҷмаш аз 7000 калима (40000 рамз) то 20 калима (100 рамз), муқаррар карда шуд;

4) бо мақсади кам кардани ҳаҷми протседураҳои ҳисобкунӣ имконияти самаранок истифода бурдани на ҳамаи элементҳои СР матн, балки фақат элементҳои баландбасомад пайдо карда шуд;

5) самаранокии омории тадбиқи γ-таснифгар ва мувофиқати СР дар асоси тақсимоти басомади унсурҳои алифбоиӣ гуногуни матн барои шинохти аломатҳои дигари якҷинсагии матн ба монандӣ: мавзӯи матн, забон, гурӯҳи забонҳо, асл ва тарҷумаи он, услуби асарҳо ва рамзи асарҳои илмӣ муайян карда шуд;

6) қонуниятҳои омории муайян намудани муаллифон ва забонҳои асарҳо бо ёрии γ-таснифгар дар пайкараи осори адабии бадеӣ таҳқиқ карда шуданд;

7) γ-таснифгар ва усули ҳамсояи наздиктарин бо матнҳои тасодуф гирифташуда барои муайян кардани аломатҳои якҷинсагии матн дар матнҳои кам ва пайкараи матнҳо санҷида шуданд, ки дақиқати кофӣ баландро доданд;

8) самаранокии тадбиқи γ-таснифгар барои муайян кардани муаллифи асари сунъии бо воситаи шабакаҳои нейронии LSTM (Long short-term memory) тартиб шудаи “Шоҳнома”-и А. Фирдавсӣ муқаррар карда шуд;

9) таъсири тартиби СР матн ба муайян кардани якҷинсагии (ҳамгунии) матн бо истифода аз γ-таснифгар тадқиқ карда шуд;

10) бори нахуст дар Тоҷикистон комплекси барномаҳои компютерии ба объект нигаронидашуда сохта шуд, ки қобили шинохти (муайянкунии) якҷинсагии матнҳои номаълум дар асоси СР-и гуногун ва γ-таснифгар дар байни шумораи зиёди матнҳо мебошад.

**Аҳамияти назариявии диссертатсия** дар он зуҳур меёбад, ки дар он усули нави таснифоти бузургиҳои тасодуфии фосиладор (γ-таснифгар) мавриди санҷиши таҷрибавӣ қарор ёфт ва самаранокии татбиқи он дар масъалаҳои муайянкунии ҳамгунии матни номаълуми чопӣ барои ҳар гуна забонҳои табиӣ, ки алифбои ҳуруфӣ доранд, муқаррар карда шуд.

**Арзиши амалии кор** дар он аст, ки тадбиқи комплекси нармафзори компютерии дар он сохташуда дар *фаъолияти маъмурии давлатӣ* барои автоматикунонии раванди коркарди иттилооти матнӣ, *дар соҳаи криминология* барои муқаррар намудани муаллифи матнҳои номаълум, *дар соҳаи маориф ва илм* оид ба ошкор намудани асардуздӣ дар корҳо (лоиҳаҳо)-и курсӣ ва дипломӣ, аз он ҷумла рисолаҳои номзадӣ ва доктории барои ҳимоя пешниҳод шаванда, равона шудааст.

Комплекси барномаҳо таҳти унвони «**THR**» (text homogeneity recognition) дар ташкилотҳои зерин тадбиқ карда шуд:

1. Академияи Вазорати корҳои дохилии Ҷумҳурии Тоҷикистон.

2. Кумитаи давлатии Амнияти миллии Ҷумҳурии Тоҷикистон.

3. Институти забон ва адабиёти ба номи Рӯдакии АМИТ.

4. Институти математикаи ба номи А. Ҷӯраеви АМИТ.

5. Донишгоҳи техникии Тоҷикистон ба номи академик М.С. Осимӣ.

Ин комплекс ба истифодаи васеъи моделҳои математикӣ ва сатҳи баланди барномасозӣ ба даст оварда шуда, барои рушди забони тоҷикӣ бо истифодаи имкониятҳои технологияҳои иттилоотӣ равона шудааст.

Комплекси барномаи мазкур ҳам аз нигоҳи забоншиносии компютерӣ ҳам аз нигоҳи адабиётшиносӣ хеле муҳим буда, барои расонидани ёрии амалӣ ба муҳаққиқони соҳаҳои забон, адабиёт, математика ва технологияҳои иттилоотӣ нигаронида шудааст. Аз он ҷумла 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.

**Нуқтаҳои ба ҳимоя пешниҳодшаванда:** исботи таҷрибавии тадбиқи самараноки γ-таснифгар бо ёрии СР-и гуногуни матн барои шинохти якҷинсагии иттилооти матнӣ.

**Дараҷаи эътимоднокии натиҷаҳои** ба даст омада, ки дар онҳо бо якчанд экспериментҳои ҳисобии силсилавӣ бо дараҷаи саҳеҳияти кофӣ баланд γ-таснифгар ва методи ҳамсояи наздик ҳангоми муайян кардани аломатҳои гуногуни якҷинсагӣ дар матнҳои кам ва пайкара тасдиқ шудаанд.

**Мутобиқати диссертатсия ба шиносномаи ихтисоси илмӣ.** Мундариҷаи таҳқиқоти диссертатсияи зерин ба бандҳои 1, 3, 4, 5 ва 7-и шиносномаи ихтисоси илмии 05.13.11 – “Таъминоти математикӣ ва барномавии мошинҳои ҳисоббарор, муҷтамаъҳо ва шабакаҳои компютерӣ”, мувофиқат мекунад:

– амсилаҳо, усулҳо ва алгоритмҳои банақшагирӣ ва таҳлили барнома ва низоми барномавӣ, инчунин тағйирдиҳии эквивалентӣ, верификатсия ва тестиронии онҳо;

– амсилаҳо, усулҳо, алгоритмҳо, забонҳо ва воситаҳои барномавӣ барои ташкили таъсири барнома ва низоми барномавӣ;

– низоми идоракунии манбаи маълумот ва дониш;

– низоми барномаҳои ҳисоби рамзҳо;

– интерфейсҳои инсон-мошин; амсилаҳо, усулҳо, алгоритмҳо ва воситаҳои барномавии мошинии графикӣ, визуализатсия, коркарди тасвир, низоми ҳақиқати виртуалӣ, муоширати мултимедӣ.

**Саҳми шахсии довталаби дарёфти дараҷаи илмӣ дар таҳқиқот.** Рисолаи докторӣ натиҷаи тадқиқоти беш аз 10 солаи муаллиф буда, дар шуъбаи “Тарҳрезии математикӣ”-и Институти математикаи ба номи А. Ҷӯраеви Академияи миллии илмҳои Тоҷикистон ва Донишгоҳи техникии Тоҷикистон ба номи академик М.С. Осимӣ омода карда шудааст. Гузориши масъалаҳо бо ҳамҷоягии мушовири илмӣ амалӣ шудааст. Натиҷаҳои асосии кори диссертатсионӣ аз тарафи муаллиф мустақилона ба даст оварда шудаанд.

**Тасвиб ва амалисозии натиҷаҳои диссертатсия.** Натиҷаҳои ҷудогонаи кор дар конференсияҳо ва семинарҳои зерин пешниҳод ва муҳокима гардидаанд:

– семинарҳои илмӣ-тадқиқотӣ дар Институти математикаи АМИТ, Донишкадаи политехникии Донишгоҳи техникии Тоҷикистон ба номи академик М.С. Осимӣ дар шаҳри Хуҷанд ва Донишгоҳи Славянии Тоҷикистону Русия дар солҳои 2011-2024;

– конференсияи байналмилалии илмию амалии "Тайёр намудани мутахассисони бозори меҳнат дар шароити ҳамгироии муассисаҳои таҳсилоти олии касбии давлатҳои хориҷӣ ва ҶТ", соли 2013, шаҳри Душанбе;

– конференсияи байналмиллалии "Помир: проблемаҳои мубрами рушди илму техника", соли 2013, шаҳри Хоруғ;

– мизи мудаввари якуми байналмилалии "Мушкилоти арзишҳои маънавию иҷтимоии ҷавонони муосир дар Русия ва Осиёи Марказӣ ва роҳҳои ҳалли онҳо", соли 2013, шаҳри Абакан;

– семинарҳои илмӣ-тадқиқотӣ дар "Технологияҳои нави иттилоотӣ дар автоматикунонии системаҳо" дар солҳои 2014, 2016, 2018, 2019, шаҳри Москва;

– конференсияи байналмилалии илмӣ-амалии "Дурнамои рушди илм ва маориф", соли 2016, шаҳри Душанбе;

– конференсияи байналмилалии "Камоли Хуҷандӣ: ташаккули адабиётшиносӣ, равобити адабӣ ва худшиносии миллӣ", 28-29 октябри соли 2016, шаҳри Хуҷанд;

– конференсияи илмӣ-амалии байналмилалии "Нақши технологияҳои иттилоотию коммуникатсионӣ дар рушди инноватсионии иқтисодии Ҷумҳурии Тоҷикистон", соли 2017, шаҳри Душанбе;

– конференсияи байналмилалии илмӣ "Проблемаҳои математикии муосир ва ҳалли онҳо", 14-15 июни соли 2017, шаҳри Душанбе, шаҳри Кӯлоб;

– конференсияи умумирусиягии илмӣ-амалии "Ҳолат ва дурнамои рушди IT-маълумот", 2019, Ҷумҳурии Чуваш;

– конференсияи илмию техникии байнидонишгоҳҳои ҳарсолаи донишҷӯён, аспирантҳо ва мутахассисони ҷавон ба номи Е.В. Арменский, МИЭМ НИУ ВШЭ, Бахши №1 "Математика ва моделсозии компютерӣ", соли 2020, шаҳри Москва;

– конференсияи 8-уми байналмилалии илмӣ-амалии "Илм ва амалия: татбиқи ҷомеаи муосир", 26-28 декабри соли 2020, Манчестер, Британияи Кабир;

– конференсияи XVI байналмилалӣ оид ба забоншиносии компютерӣ ва когнитивӣ TEL-2020, 12-13 ноябри соли 2020, Қазон, Русия;

– конференсияи ҷумҳуриявии илмию назариявӣ дар мавзӯи "Иқтисоди рақамӣ ва зарурати ҷорӣ намудани низоми нави ҳисобҳои миллӣ", 17 феврали соли 2021, шаҳри Душанбе;

– конференсияи XI байналмилалии илмӣ-техникии "Технологияҳои кушодаи семантикӣ барои тарҳрезии системаҳои зеҳнӣ", Технологияҳои кушодаи семантикӣ барои системаҳои интеллектуалӣ (OSTIS-2021), 16-18 сентябри соли 2021, шаҳри Минск, Ҷумҳурии Беларус;

– конференсияи байналмилалии илмӣ-амалии "Илмҳои техникӣ ва таълими муҳандисӣ барои рушди устувор", 12-13 ноябри соли 2021, Донишгоҳи техникии

Тоҷикистон ба номи академик М.С. Осимӣ, шаҳри Душанбе;

– конференсияи байналмилалӣ бахшида ба хотираи профессор А. Тарасов ва О.В. Казарин дар мавзӯи “Ҳамкории байни донишгоҳҳо, ташкилотҳои илмӣ ва муассисаҳои фарҳангӣ дар соҳаи амнияти иттилоотӣ ва бехатарии технологияҳо”, 19 ва 20 апрели соли 2022, шаҳри Москва;

– конференсияи байналмилалии илмию амалии “XII хонишҳои Ломоносов”, бахшида ба 30-солагии пойдории муносибатҳои дипломатӣ байни Ҷумҳурии Тоҷикистон ва Федератсияи Русия, Филиали Донишгоҳи давлатии Москва ба номи М.В. Ломоносов дар шаҳри Душанбе, 29-30 апрели соли 2022;

– конференсияи VI байналмилалии илмӣ-амалии “Ҷанбаҳои глобалӣ ва минтақавии рушди устувор”, 26-28 феврали соли 2022, шаҳри Копенгаген, Дания;

– конференсияи XXII байналмилалии “Информатика: проблемаҳо, усулҳо, технологияҳо” (IPMT-2022), Донишгоҳи давлатии Воронеж, 10-12 феврали соли 2022, шаҳри Воронеж;

– конференсияи байналмилалии илмӣ-амалӣ дар мавзӯи “Рақамикунонӣ ва зеҳни сунъӣ” бахшида ба “Бистсолаи омӯзиш ва рушди фанҳои табиатшиносӣ, дақиқ ва риёзӣ дар соҳаи илму маориф (солҳои 2020-2040)”, Донишгоҳи техникии Тоҷикистон ба номи академик М.С. Осимӣ, 2023, Душанбе;

– конференсияи байналмилалии “Проблемаҳои муосири математика” бахшида ба 50-солагии Институти математика ба номи А.Ҷӯраеви Академияи миллии илмҳои Тоҷикистон, 26-27 майи соли 2023, Душанбе;

– мураббии беҳтарин – 2023: маҷмуаи IV байналхалқии китобҳои коркунони илмӣ ва педагогӣ, 2023, Остана;

– конференсияи байналмилалии илмию амалии “Комёбиҳои нав дар соҳаи илмҳои табиатшиносӣ ва технологияи информатсионӣ” бахшида ба “Бистсолаи омӯзиш ва рушди фанҳои табиатшиносӣ, дақиқ ва риёзӣ дар соҳаи илму маориф (солҳои 2020-2040)”, 2023, Душанбе, ДСРТ.

**Интишорот аз рӯйи мавзӯи диссертатсия.** Доир ба мавзӯи рисола 73 асари илмӣ чоп шудааст, ки аз он 34 (14 бе ҳаммуаллифӣ) мақолаҳо дар нашрияҳои аз тарафи КОА-и назди Президенти Ҷумҳурии Тоҷикистон тавсияшуда, 30 гузориш дар маҷаллаҳои илмӣ ва конференсияҳои байналмилалӣ, ду монография ва ду дастури таълимӣ, инчунин панҷ пойгоҳи додаҳо ва барномаҳои компютерӣ, ки ҳамчун объекти моликияти зеҳнӣ ба қайд гирифта шудаанд, [1-М-73-М].

**Сохтор ва ҳаҷми диссертатсия.** Рисолаи илмӣ аз мундариҷа, шаш боб, хулоса ва аз 397 рӯйхати адабиёти истифодашуда иборат аст. Муҳтавои диссертатсия дар 271 саҳифа дарҷ гардида, 107 ҷадвал ва 9 расмро дар бар мегирад.

Муаллиф ба мушовирони илмӣ – доктори илмҳои ф.-м., профессор, академики АМИТ Усмонов З.Ҷ. ва доктори илмҳои ф.-м., профессор, академики АМИТ Раҳмонов З.Ҳ., инчунин ба кормандони Донишкадаи политехникии Донишгоҳи техникии Тоҷикистон ба номи академик М.С. Осимӣ дар шаҳри Хуҷанд, Институти математикаи ба номи А. Ҷӯраеви АМИТ ва Донишгоҳи техникии Тоҷикистон ба номи академик М.С. Осимӣ миннатдории махсуси худро баён мекунад.

## ҚИСМИ АСОСИИ ТАҲҚИҚОТ

**Мавод ва методҳои таҳқиқот.** Маводи таҳқиқот ба омӯзиши масъалаи шинохтӣ якҷинсагии порчаи матнҳо бахшида шудааст. Аломатҳои янҷинсагӣ – азбаски масъала ба матн марбут аст, аломатҳои зерини якҷинсагӣ барои таҳқиқот гирифта мешаванд: шинохти муаллифӣ, мавзуъҳои матн, забон, гурӯҳбандии забонҳо, асл ва тарҷумаи он, услуби асарҳо ва рамзҳои осори илмӣ.

Дар кор барои шинохтӣ якҷинсагии матнҳо моделҳои риёзии қабули қарор истифода мешаванд, ки дар байни онҳо махсусан муваффақанд: шабакаҳои нейронӣ, мошини ёрирасони векторӣ, усули ҳамсояи наздик (аз рӯйи масофа) ва γ-таснифкунанда, ки ба наздикӣ дар Институти математикаи ба номи А. Ҷӯраев АМИТ таҳия шудааст, истифода бурда шудаанд.

**Натиҷаҳои таҳқиқот.** Мазмуни мухтасари натиҷаҳои бобҳои кори диссертатсиониро меорем.

Дар **муқаддима** аҳамияти мавзӯи кори диссертатсионӣ, асоснок карда шуда, мақсад ва вазифаҳои рисола, проблемаи илмӣ шаклдиҳӣ, объект, мавзуъ, усулҳои тадқиқот муайяни муқаррароти асосии барои дифоъ пешниҳод карда шуда, навгонии илмӣ, аҳамияти назариявӣ ва амалии натиҷаҳои ба даст омада ифода гардида ва дар охир дар бораи апробатсияи кор маълумот дода мешавад.

Дар **боби 1** *“Мафҳумҳои асосӣ ва таърифҳо”* баррасии адабиёт (мақолаҳо ва нашрияҳо) пешниҳод карда мешавад, масъалаи мушкилотро барои шинохтӣ худкори якҷинсагии матн тасвият мешавад, мафҳумҳоеро, ки дар оянда васеъ истифода мешаванд, оварда шуда, тавсифи муфассали γ-таснифгар нишон дода мешавад, алгоритм ва тавсифи мухтасари он вазифаҳоеро, ки дар бобҳои дигар омӯхта мешаванд, оварда шудаанд. Дар аввал қайд кардан лозим аст, ки кори γ-таснифгар дар матнҳои шумораашон кам (коллексияҳои моделӣ) таҳқиқ карда мешавад. Андозаҳои хурд дар аввал барои тадқиқоти пешакӣ истифода мешаванд ва танҳо пас аз ба даст овардани натиҷаҳои назаррас дар коллексияҳои моделӣ, усулҳои коркарди истифодашуда дар пайкараҳои матнҳо тадқиқ карда мешаванд, ба боби 4 нигаред.

Дар **§ 1.1** шарҳи адабиёт (мақолаҳо ва интишорот) оид ба шинохти худкори якҷинсагии матн оварда шудааст. Татбиқи усулҳои амсиласозии математикӣ дар муайян кардани якҷинсагии матнҳои ношинос ба модели матн, яъне тавсифи миқдории объекти таҳқиқот асос ёфтааст. Ҳоло тибқи ҳисобҳои Ҷ. Рудман [3], тақрибан 1000 гурӯҳи аломатҳо ҳамчун модели матнӣ, аз ҷумла хусусиятҳои морфологӣ, лексикӣ, қимобӣ, синтаксисӣ, сохторӣ, мундариҷаи мушаххас ва ғайра истифода мешаванд. Ғайр аз гуфтаҳои боло қайд кардан бамаврид аст, ки дар монографияи А.А. Шелупанов, А.С. Романов[4] ва Р.В. Мещеряков, баррасии васеи корҳо оид ба муайянкунии монандии матн дар асоси СР-и гуногун ва

[3] Rudman, J. The state of authorship attribution studies: Some problems and solutions // Computers and the Humanities, 1998, Vol. 31, pp. 351-365.

[4] Романов, А.С., Шелупанов, А.А., Мещеряков, Р.В. Разработка и исследование математических моделей, методик и программных средств информационных процессов при идентификации автора текста // -В-Спектр, Томск, 2011, 188 с.

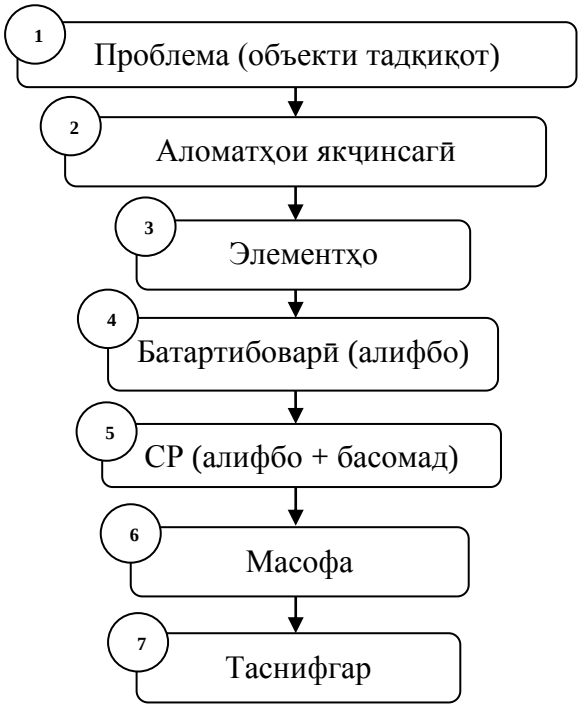
методҳои татбиқшавандаи таснифгар пешниҳод карда шудааст. Дар зербоби навбатӣ он масъалаҳое, ки дар ин рисола бояд ҳал шаванд, ифода меёбад.

Дар **§ 1.2** мураттабсозии масъалаҳо тасвир шудааст, ки ҳалли онҳо тасвири пурраи самаранокии истифодаи γ-таснифгарро барои муайян кардани якҷинсагии осор нишон медиҳад. Ҳафт тавсифи асосии мушкилоте мавҷуданд, ки ҳангоми шинохти якҷинса будани объектҳо ба миён меоянд, ки дар расми 1.1 нишон дода шудаанд.

1. Проблема ё объекти тадқиқот – ҷараён ё падидае мебошад, ки аз ҷониби муҳаққиқ барои омӯзиш ё ҳамчун қисми дониши илмӣ гирифта мешавад, ки худ дарк мекунад. Мисоли объектҳои омӯзиш инҳоянд: матн, тасвирҳо, садо, формула, коди барномавӣ ва ғайра. Дар ин рисола объекти омӯзиш матн аст.

2. Аломатҳои якҷинсагӣ – азбаски масъала ба матн марбут аст, аломатҳои зерини якҷинсагӣ ба назар гирифта мешаванд: муайянкунии муаллиф, мавзуъҳои матн, забон, гурӯҳбандии забонҳо, асл ва тарҷумаи он, услуби асарҳо ва рамзҳои осори илмӣ. Ин проблемаҳои илмӣ дар бобҳои 2-4 таҳқиқ карда мешаванд.

3. Элементҳо – мисоли элементҳои матн метавонанд ҳарфҳои алифбои забони табиӣ, *N*-граммаҳо ва ҳиҷоҳои ҳарфӣ, аломатҳои китобат, морфемаҳо, шаклҳои калима, дарозии калимаҳо, ҷумлаҳо ва параграфҳо (дар символҳо ва калимаҳо), анаграммаҳо ва ғайра бошанд.




**Расми 1.1. - Тавсифи масъала**

4. Батартибоварӣ (алифбо) – агар элементҳо муайян бошанд (яъне интихобшуда), он гоҳ натиҷа аз тартиби ҷойгиршавии элементҳо (яъне интихоби алифбо) вобаста аст. Ин масъала дар боби 5 дида баромада мешавад.

5. CP – тавсифи миқдории объекти омӯзишӣ буда, модели математикии онро тақсимоти басомадҳои элементҳои алифбо меномем. Мисолҳои CP-и матнӣ ин қатори чандомади рамзии, алифбои ва калимавии *N*-граммаҳо, дарозии калима ва

ҷумлаҳо ва ғайра мебошанд.

6. Масофаи байни матнҳо – ба маънои васеъ, дараҷаи (ченаки) дурӣ ё наздик будани матнҳо аз ҳамдигар аст. Барои ҳисоби масофа формулаҳои зиёде мавҷуданд, масалан: γ-таснифгар, масофаи евклидӣ, коэффисиенти коррелятсия, Смирнов-Колмогоров, Фишер-Синдекор ва ғайра.

7. Таснифгар – барои шинохти матнҳои якҷинса, ба ҷуз СР, моделҳои зиёди математикии қабули қарор истифода мешаванд, ки дар байни онҳо махсусан шабакаҳои нейронӣ, мошини ёрирасони векторҳо ва γ-таснифгари ба наздикӣ дар Институти математикаи ба номи А. Ҷӯраев АМИТ таҳияшуда муваффақ мебошанд. Дар **§§ 1.3** ва **1.4**-и навбатӣ тавсифи моҳияти γ-таснифгар оварда мешавад.

Дар **§ 1.3** истилоҳот ва мафҳумҳое оварда мешавад, ки барои тавсифи модели математикии матн истифода мешаванд.

**1.3.1. Масъалаи муайянкунии муаллифи асар.**

Бигзор $\mathbb{A} = \{A_i\} -$ рӯйхати муаллифони $A_i, i = \overline{1, \alpha},$ ва $\mathbb{T} = \{T_j\}$ – баъзе маҷмӯи матнҳои ба онҳо тааллуқ доштаи $T_j, j = \overline{1, \beta}$ бошанд. Бигзор, ки $\mathbb{T}$ ба ду қисм тақсим карда шавад, $\mathbb{T} = \mathbb{T}_1 + \mathbb{T}_2$ , аз он $\mathbb{T}_1$ барои коркади қоидаи мувофиқат (инъикос) “матн → муаллиф” (**масъалаи 1** – омӯзиши модели математикӣ), $\mathbb{T}_2$ – барои тафтиши самаранокии қоидаи коркунӣ (**масъалаи 2** – тестиронии модели математикӣ).

Мавҷудияти робитаи байни матн ва муаллифи он асоси сабкшиносии муосир мебошад. Аз нигоҳи омор услуби муаллиф – ин падидаи эҳтимолист. Аслан, ҳама гуна унсурҳо ё аломатҳои дар матнҳо мавҷудбуда бо баъзе басомадҳое пайдо мешаванд, ки таҳти назорати муаллиф нестанд ва бо вуҷуди ин, маълумоте доранд, ки эҷодкори онҳоро тавсиф мекунанд.

Инак масъалаи ошкор сохтани муаллифи матн, бо ду модели математикӣ сару кор гирифт: яке тавсифи миқдории (намунаи) матн ва дигар моделе, ки барои қабули ҳалли масъала яъне таснифгар ба кор меравад. Тавсифи матн ва моделҳо – ниҳояд зиёданд. Дар айни замон ҷуфти моделҳои гуногун бо мақсади тадқиқот истифодашаванда тавсиф шудаанд. Имкониятҳои зиёди мувофиққати элементҳои ҷуфт сабаби он аст, ки мутахассисон ба масъалаи шаклгирии назарияи умумӣ сарукор надоранд, фақат бо ҷустуҷӯи ҷуфти самаранок барои ҳалли масъалаҳои мушаххас муайян карадани муаллифи асар истифода мебаранд.

Масъалаи мавриди муҳокима қарор гирифта яке аз ҷузъҳои махсуси проблемаи умумии сохтани низомҳои шинохти образҳо, ки аз коркарди оптималии протседура барои таснифоти образҳо ва муайянкуни объектҳо иборат аст. Аз ин рӯ, тамоми дастовардҳо дар таҳияи системаҳои шинохтан дар ҳалли масъалаҳои муайянкунии муаллиф истифода мешаванд.

**1.3.2. СР-и матни чопӣ.**

Як қатор таърифҳоро пешниҳод мекунем, ки дар бобҳо ва зербобҳои минбаъда истифода хоҳанд шуд.

**Таърифи 1.3.1.** ***Алифбо –*** *батартибории маҷмӯи элементҳои матн*, нигаред ба **§ 1.2**.

Мисоли элементҳои матн ҳарфҳои забони табиӣ, рамз ва аломатҳои китобат, *N*-грамма ва ҳиҷоҳои ҳарфӣ, лемма ва морфемаҳо, реша ва асосҳои калимаҳо, шаклҳои калима, калимаҳои махсус ва *N*-граммаҳои калидӣ, дарозии калима ва ҷумлаҳо ва ғайра ҳастанд. Ҷамъи элементҳо, *ки бо ягон роҳ батартиб оварда* шудаанд, алифборо ташкил медиҳанд.

**Таъриф 1.3.2.** *Қатори чандомади элементҳои алифбо* ***СР-и матн*** *номида мешавад.*

Аз ин рӯ, СР-и матн – ин ҷуфти аз як тараф элементҳои батартибёфтаи матн ва аз тарафи дигар, маълумот дар бораи тақсимоти чандомади вохӯрӣ дар матн худи элементҳоро дарбар мегирад. Чунин мисолҳо ба монандӣ тақсимоти чандомади рақамҳо, *N*-граммаи ҳарф ва калимаҳо, дарозии калима ва ҷумла ва ғайра шуда метавонанд.

СР-и матни *T* дар шакли ҷадвали зерин навишта мешавад:

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

ки сатри якум – рақами тартибии (индексҳои) элементҳои алифбо ($m$ – шумораи элементҳо) ва сатри дуюм – басомади нисбии вохӯрии онҳо дар $T$, дар баробари ин $\sum_{k=1}^{m} p_k = 1$ мебошад.

СР боз дар намуди функсияҳои дискретӣ ифода карда мешавад:

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

**1.3.3. Масофа байни СР-и матнҳо.**

Бигзор $T_1$, $T_2$ – ҷуфти матнҳои ихтиёрӣ, ки аз рӯйи як алифбо муайян карда шуда бошад ва

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

СР-и мувофиқи онҳо бо чунин функсияҳои дискретӣ ифода меёбад $\alpha = 1, 2$, ва $s = 1, \cdots, m$.

**Таърифи 1.3.3.** *Рақами мусбати* $\rho\,(T_1, T_2)$, *ин масофаи байни матнҳои* $T_1$ *ва* $T_2$ *мебошад, ки бо формулаи зерин муайян мегардад*

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

яъне масофаи байни ду матн ҳамчун ҳисоби масофаи максималии бо меҳвари ординатаи байни функсияҳои дискретии $F^{(1)}(s)$ ва $F^{(2)}(s)$ ва ба коэффитсиенти вазнии $\sqrt{m/2}$ зарбшаванда аст. Инчунин қайд мекунем, ки ҳангоми $\rho\,(T_1, T_2) = 0$ будан маънои як будани СР-и $T_1$ва $T_2$-ро дорад, на балки худи матнҳоро.

**1.3.4. Фарзияи $\mathbb{H}$ “якҷинсагӣ” хусусиятҳои услуби муаллиф.**

Дар эҷодиёти муаллифон муайян шудани “якҷинсагии” ин ё он хусусиятҳои услубӣ дар осори онҳо, истифодаи калима, синтаксис, таркиб, интонатсия, ритм ва ғайра зоҳир мегардад. Ин мафҳумро шарҳ надода, синонимҳои он: “монандӣ”, “якхелагӣ”, “ҳамгунӣ”, “якнамуда”, “авлодӣ” ва ғайраро барои фаҳмидан оварда мешавад. Ҳамаи онҳо ба мафҳуми услуби муаллиф баста шудаанд, ки эҷодиёти

муаллифро дар заминаи ҳамкорони худ аз ҷомеаи нависандагон муайян мекунад.

Фарзияи $\mathbb{H}$, ки бо маънои пурмазмуни масъалаи таҳқиқшаванда алоқаманд буда барои ҳалли масъалаи 1 тавассути интихоб ва минбаъд танзимкунии модели математикӣ истифода мешавад. Бештар табиатан чунин ифода меёбад:

ФАРЗИЯИ $\mathbb{H}$. *Осори як муаллиф – "якҷинса" ва муаллифони гуногун – "ғайриякҷинса" мебошанд.*

Осор – мафҳуми васеъ дорад. Он бо маҷмӯи хусусиятҳо хос аст. Дар ин сурат хосияти "якҷинсагӣ"-и асарҳоро метавон ҳамчун аломатҳои алоҳида ё маҷмӯи "якҷинсагӣ"-и онҳоро шарҳ дод. Аз ин рӯ, шакли фарзияи тағйирёфтаи зеринро метавон мавриди баҳс ифода кард:

ФАРЗИЯИ $\mathbb{H}^*$. *Аломатҳои муайян дар ҳамаи асарҳои як муаллиф "якҷинса" ва дар эҷодиёти муаллифони дигар "ғайриякҷинса" ҳастанд.*

Аз ин нуқтаи назар маълум мешавад, ки чаро муҳаққиқоне, ки дар таҳқиқи шинохти муаллифии матн иштирок доранд, на ба кулли матнҳо, балки бо аломатҳои хоси он сару кор доранд. Ҳамин тавр, масалан, тақсимоти униграммаҳои ҳарфӣ, биграммаҳо, триграммаҳо (бо фосила ва бе фосила), ҳиҷоҳо, морфемаҳо, *N*-граммаҳои калимаҳо, дарозии ҷумла ва абзатсҳо ва бисёр аломатҳои дигар, ки барои муайян кардани муаллифони порчаҳои матнӣ истифода бурда мешаванд.

Дар адабиёт мисолҳои зиёди риоя нашудани фарзияти зерин вуҷуд дорад, аммо он ҳамчун наздикшавии аввалин ба ҳолати воқеӣ қабул карда мешавад, ки имкон медиҳад фарзият ба модели математикӣ табдил дода шавад.

Дар параграфи **1.4** тасвири методе, ки дар кор усули қабули қарор бо воситаи $\gamma$-таснифгар мебошад, оварда мешавад, нигаред ба зернависи 2 дар саҳифаи 3.

γ-таснифгар – ин сегонаи математикие мебошад, ки аз СР-и матнӣ, формулаи масофа байни матнҳо ва алгоритми омӯзишӣ дар асоси назир иборат аст.

**1.4.1. Модели математикии $\mathbb{H}$-фарзият.**

Бигузор $\gamma$ – баъзе адади мусбат бошад.

**Таърифи 1.4.1.** *Матнҳои $T_1, T_2$, **$\gamma$-якҷинса** номида мешаванд, агар*

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

*ва **$\gamma$-ғайриякҷинса агар***

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

Нобаробарии (1.5) ва (1.6) тафсири математикии (моделии) фарзияти $\mathbb{H}$ мебошанд.

**Таърифи 1.4.2.** *$\gamma$-классификатор – алгоритме, ки аз як параметри ҳақиқии $\gamma$ вобаста аст ва ба матни $\mathbb{T}_1$ муаллифаш аз рӯйхати $\mathbb{A}$ мувофиқат мекунад.*

Маълум аст, ки аз қимати $\gamma$ якҷинсагӣ ва ғайриякҷинсагии ҷуфти матнҳо, инчунин дараҷаи иҷроиши фарзият вобастагӣ дорад. Якҷинса будани ҳамаи матнҳои як муаллиф дар ҳудуди модели математикии нобаробарии (1.5) буда,

ғайриякҷинсагии ду матни муаллифони гуногун – иҷроиши нобаробарии (1.6) мебошад. Фарзияти $\mathbb{H}$ дар он сурат вайрон карда мешавад, ки дар ҳолати ҷуфти матнҳои як муаллиф ба ҷои нобаробарии (1.5), нобаробарии (1.6) иҷро мегардад, инчунин ҳангоми ҷуфти матнҳои муаллифашон гуногун нобаробарии (1.5) қонеъ мегардад, ба ҷои он ки нобаробарии (1.6) иҷро гардад.

Бигзор $\tau = \tau(\gamma)$ – ҳосили ҷамъи шумораи вайрон шудани фарзияти $\mathbb{H}$ дар ду маврид бошад: иҷро нагардидани нобаробарии “якҷинсагӣ” дар ҳолати ду матне, ки ба як муаллиф тааллуқ дошта ва қонеъ нагардидани нобаробарии “ғайриякҷинсагӣ” дар маврид ду матне, ки муаллифашон гуногунанд. Он гоҳ, барои $\gamma$-и муайян гардида нишондиҳандаи иҷроиши фарзият бо қиммати $\pi$ муқаррар карда мешавад, ки бо формулаи зерин ҳисоб карда мешавад:

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

дар он $L$ – рақами масофаҳои байни ҷуфтҳои матнҳо аз зерколлексияи $\mathbb{T}_1$ аст. Аз ин формула бармеояд, ки $\pi$ метавонад қимматро аз порчаи $[0, 1]$ қабул намояд, илова бар ин $\pi = 0$, агар $\tau = L$ ва $\pi = 1$, агар $\tau = 0$ бошад. Дар ҳолати аввал, фарзияти $\mathbb{H}$ корношоям дониста шуда, аммо дар мавриди дуюм – бо маълумоти интихобшавандаи омӯзишӣ пурра мутобиқат мекунад.

Ғайр аз ин самаранокии $\gamma$-таснифгар аз қиммати параметри $\gamma$ вобаста аст, инчунин аҳамияти ёфтани чунин қимати он муҳим, ки дар он ҳангом қиммати $\pi$ максималӣ бошад. *Махсусан дар ҳамин танзимкунии $\gamma$-таснифгар дар маълумотҳои интихобии омӯзишӣ ба мақсад мувофиқ аст.* Агар чунин танзимкунӣ қобили қабул бошад, дар ин ҳолат оиди ҳалли **масъалаи 1** – омӯзиши $\gamma$-таснифгар ҳарф зад.

**Мулоҳиза.** Ба он таваҷҷуҳ кунем, ки фарзиятҳои $\mathbb{H}$ ва $\mathbb{H}^*$, барои муайян кардани муаллиф ва хусусиятҳои услуби муаллиф равона гардида, инчунин ба ҳалли мақсадҳои дигар истифода карда шаванд.

Барои мисол, агар асарҳоро аз рӯйи мавзуъҳои гуногун фарқ кардани бошем, пас $\mathbb{H}^{**}$ – фарзиятро барои танзимкунии $\gamma$-таснифгар дар шакли зерин ифода мекунанд: *ҳама гуна асарҳое, ки дар як мавзӯъ навишта шудаанд “якҷинса” мебошанд, аммо дар мавзуъҳои гуногун онҳо “ғайриякҷинса” ҳастанд*. Ва боз нобаробариҳои (1.5) ва (1.6)-ро метавон ҳамчун тафсири математикии (моделии) $\mathbb{H}^{**}$– фарзият истифода кард.

Мисоли дигар – муайян кардани забонҳои асарҳост. Дар ин маврид $\mathbb{H}^{**}$ – фарзият дар шакли андаке осон тағйир дода тасвият мешавад: *ҳамаи асарҳое, ки ба як забон навишта шудаанд “якҷинса” буда, вале асарҳои дар забонҳои гуногун “ғайриякҷинса” мебошанд*. Ва боз нобаробариҳои (1.5) ва (1.6) метавонанд ҳамчун тафсири математикии (моделии) $\mathbb{H}^{**}$– фарзият амал кунанд.

Қайд кардан муҳим аст, ки самаранокии фарзиятҳо на танҳо аз $\gamma$-таснифгар, балки аз бодиққат интихоб шудани CP-и объекти омӯзишӣ низ вобаста аст.

Дар чор боби навбатии ин рисола мо муайян кардани якҷинсагии матнро дар асоси CP-и гуногуни матн ва $\gamma$-таснифкунанда дар байни миқдори зиёди матнҳо меомӯзем.

Дар **боби 2** *“Тадқиқоти самаранокии шинохти якҷинсагии матн дар мисоли коллексияи моделии осори бадеӣ”* мо ба намунаи матнҳои кам, ки аз се қисм иборат аст: осори классикони адабиёти тоҷику форс, шоирони муосир ва нависандагони муосир, рӯ овардем. Ҳар як қисм аз 10 асарӣ, бо ду асари панҷ муаллиф иборат аст. Аломатҳои миқдории дараҷаи баланд ба имконияти истифодаи шинохти муаллифи матн дар мисоли коллексияи ками осори бадеии забони тоҷикӣ, инчунин забони ӯзбекӣ санҷида шуда, ба сифати методи тадқиқот γ-таснифгар ва ҳамсояи наздиктарин гирифта шуд. Мақсади мо на танҳо муайян кардани фарқиятҳо дар ҳаҷм ва ҷойгиршавии нимфосилаҳои оптималии γ, балки муайян кардани шумораи вайронкунии фарзияти якҷинсагӣ, ҳисоб кардани коэффитсиенти самаранокии муайян кардани муаллифон дар асоси асарҳои онҳо ва дар маҷмуъ эҳтимолан шинохтан дар порчаҳои хурди матнӣ мебошад. Порчаҳо аз *“оғоз”*, *“байн”* ва *“охир”-и* асар ҷудо карда, ки “дар ҳудуди” порчаҳои матнии андозаҳои гуногун ба таври тасодуфӣ гирифта шудаанд.

Ҳангоми истифодаи метрикии таснифгар ва усули ҳамсояи наздиктарин (бо масофа) муаллифони порчаҳои матнии ҳаҷмашон пайдарпай аз 7000 калима (40000 рамз) то 20 калима (100 рамз) камшаванда муайян карда шуд.

Дар **боби 3** “*Муайян кардани аломатҳои якҷинсагӣ”* мо ба таҳқиқи масъалаи навбатии муҳим рӯ меорем: оё аломатҳои дигари якҷинсагиро дар асоси γ-таснифгар муайян кардан мумкин аст, ба монандӣ, мавзӯи матн, забон, нусхаи асл ва тарҷумаи он, услуби асарҳо, рамзрамзҳои корҳои илмӣ ва ғайра. Маълум аст, ки ҳалли ин гуна масъалаҳо аҳамияти бағоят муҳими амалӣ дорад.

Дар **§ 3.1**, мо ба таҳқиқи муайян кардани муаллиф ва мавзӯи асарҳо дар асоси γ-таснифгар шуруъ кардем. Дар мавриди аввал ҳамчун фарзияи корӣ тасдиқ дар бораи янҷинса будани асарҳои як муаллиф ва ғайриякҷинсагии асарҳои муаллифони гуногун қабул карда мешавад; дар ҳолати дуюм – якҷинса будани асарҳо дар як мавзуъ ва ғайриякҷинсагӣ дар мавзуъҳои гуногун. Дар мисоли коллексияи хурди ***C*** асарҳои бадеии давраи шӯравӣ таъсири муштараки СР, фазои метрикӣ ва таснифгари матн барои қабули қарор оиди “якҷинсагӣ” ва “ғайриякҷинсагӣ”-и асарҳо таҳқиқ карда мешавад. Бо ёрии γ-таснифгар ҷуфти асарҳои М.А. Шолохов, Н. Островский, Б. Полевой, К. Симонов, А. Фадеев, Д. Фурманов, А.С. Серафимович ва Ф.Д. Крюков, ки бо воситаи нӯҳ СР-и гуногун намояндагӣ мекунанд, ба сифати воситаи “якҷинсагӣ” тафтиш мешавад.

Дар **§ 3.2** татбиқи γ-таснифгар барои шинохти автоматии забони асар дар мисоли коллексияи матнҳои кам аз рӯи чандомади ҳарфҳои алифбо муқаррар карда мешавад. Дар **§ 3.2.5** татбиқи γ-таснифгар барои шинохти автоматии забони асар, ки дар асоси басомади 26 алифбои маъмули лотинӣ таъсис дода шуда ва дар мисоли коллексияи ками 10 матн бо панҷ забон (англисӣ, олмонӣ, испанӣ, итолиёвӣ ва фаронсавӣ) бо хати лотинӣ навишта шудааст, муайян карда мешавад. Модели математикии γ-таснифгар ба намуди сегона тақдим шудааст. Ҷузъи аввали он СРМ – қатори чандомади униграммаҳои ҳарфҳо дар матн мебошад; қисмати дуюм формулаи ҳисоб кардани масофаҳои байни СРМ-ҳо ва сеюмин алгоритми омӯзиши мошинист, ки фарзияи “якҷинсагӣ”-и асарҳои ба як забон навишташуда ва “ғайриякҷинсагӣ”-и асарҳои бо забонҳои гуногун

навишташударо амалӣ менамояд. Мизробкунии алгоритме, ки ҷадвали масофаҳои ҷуфтии байни ҳамаи асарҳои коллексияи хурдро истифода мебарад, аз муайян кардани қиммати оптималии параметри воқеии γ иборат буда, ки барои он ҳадди ақал кам кардани миқдори хатогии вайрон кардани фарзияти “якҷинсагӣ” мебошад. Барои санҷидани таснифгари матнҳо шаш матни тасодуфии иловагӣ гирифта шуд, ки панҷтои онҳо бо матнҳои коллексияи аввала бо ҳамон забонҳо мебошанд. Ҳамчун маълумоти таҷрибавӣ, ки тадқиқоти мо дар он вусъат дода мешавад, маҷмӯаи хурди ***C*** аз 10 асар (матн) интихоб карда шуд, ки байни онҳо

*бо забони англисӣ* **(En):** В. Шекспир “Romeo and Juliet” (Ромео ва Ҷулетта, **en_1**, 25832 *калима*), М. Твен “A Connecticut Yankee in King Arthur's Court” (Янки аз Коннектикут дар дарбори шоҳ Артур, **en_2**, 117257 *калима*);

*бо забони олмонӣ* **(De):** Г. Пиз “Schiff ohne Mannschaft” (Киштӣ бе экипаж, **de_1**, 59695 *калима*), Г. Диана “Das flammende Kreuz: Roman” (Салиби Дурахшон: Роман, **de_2**, 70104 *калима*);

*дар испанӣ* **(Es):** Д.Ҷ. Генрих “El ocaso de la magia” (Шоми ҷодугарӣ, **es_1**, 73300 *калима*), В.Ф. Алберто “Oceano” (Уқёнус, **es_2**, 103596 *калима*);

*дар итолиёвӣ* **(It):** Г. Эд “Elminster: la nascita di un mago” (Элминстер: таваллуди ҷодугар, **it_1**, 127087 *калима*), С. Роберт “Il paradosso del passato” (Парадокс аз гузашта, **it_2**, 69697 *калима*);

*ва ба забони фаронсавӣ* **(Fr):** С. Жорж “Lavinia” (Лавиния, **fr_1**, 13151 *калима*), Б.Мишел ”Les Nymphéas noirs” (Савсани сиёҳи обӣ, **fr_2**, 108137*калима*).

Ҳисобҳо бо формулаҳои (1.1) – (1.4) чилу панҷ ҷуфти масофаҳои $\rho(T_1, T_2)$ байни асарҳои коллексияи ***C*** дар ҷадвали зерин нишон дода шудаанд:

Ҷадвали 3.26. – Масофаҳо байни матнҳои коллексияи ***C***

| Матнҳо | | 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** | |

Дар асоси маълумоти ҷадвали 3.26 натиҷаҳои зерин ба даст омаданд:

– маҷмӯи ҳамаи ҷуфти масофаҳо дар ҳудуди [0.0287, 0.4038] дохил мешаванд, ки дар ин ҳангом масофаи наздиктарин байни ду матни **de_1** ва **de_2** дар забони олмонӣ ва дуртарин байни **de_1** дар олмонӣ ва **fr_2** дар фаронсавӣ аст;

– қиммати оптималии нимфосилаи $\gamma$ дар дохили ҳудудӣ зерин шуд:

$$\gamma^{onm} \in [0.0833;\ 0.1353)\ ; \qquad (3.47)$$

мувофиқи таърифи **1.3.4** ин маънои онро дорад, ки агар масофаи $\rho(T_1, T_2)$ байни ду матн аз қимати $\gamma^{onm}$ ҳудудӣ фосилаи муқарраршуда зиёд набошад, он гоҳ ҷуфти матнҳо ба як забон тааллуқ доранд (масофаҳои мувофиқ дар ҷадвал бо ранги хокистарӣ қайд карда шудаанд); агар аз он зиёд бошад, пас онҳо ба забонҳои гуногун тааллуқ доранд (масофаҳои мувофиқ бе ранг гузошта мешаванд);

– қайд кардан ба маврид аст, ки барои ҳамаи (бе истисно) асарҳои коллексияи ***C*** фарзияти H пурра иҷро шуд ва тафсири математикии он дар шакли таърифи **1.3.4** чунин ба даст оварда шуда:

$$\tau = \tau_{\min} = 0,$$

яъне ҳеҷ яке аз нобаробариҳои (1.5) ва (1.6) вайрон нашуданд;

– дар натиҷа нишондиҳандаи самаранокии модели математикии шинохти забони асарҳои дар ин кор пешниҳодшуда баробар шуд:

$$\pi = \pi_{\max} = 1.$$

**Тестиронӣ.** Ҳамин тариқ, танзимкунии (омӯзиши) γ-таснифгар аз рӯи маълумотҳои коллексияи хурди матнҳои ***C*** бо муваффақ гузашт. Барои тестиронии таснифгар 6 матн ба таври тасодуфӣ интихоб карда шуд:

*бо забони англисӣ* **(En):** Ҷ. Лондон “The Call of the Wild” (Занги ваҳшӣ) (Text_En, 31763 *калима*);

*бо забони олмонӣ* **(De)**: М. Вилли “Die seltsamen Reisen des Marco Polo” (Сафарҳои ҳайратангези Марко Поло) (Text_De, 126607 *калима*);

*бо забони испанӣ* **(Es):** Д. Арне “Misterioso” (Асрор) (Text_Es, 106835 *калима*);

*бо забони итолиёвӣ* **(It):** Ш.Боб “Sfida al cielo” (Даъват ба фалак) (Text_It, 101154 *калима*);

*бо забони фаронсавӣ* **(Fr):** К.С. Доминикович “Fantôme” (Арвоҳ) (Text_Fr, 46089 *калима*);

*ва бо забони руминӣ* **(Po):** Т.Р. Руэл “Întoarcerea regelui” (Бозгашти Подшоҳ) (Text_Ro, 146266 *калима*).

Барои шаш асаре, ки барои санҷиш пешбинӣ шудаанд, СР-и (1.1) ва сипас бо истифода аз формулаҳои (1.2), (1.3), (1.4) масофаҳо то 10 асари коллексияи ***C*** барои ҳар яки онҳо ҳисоб карда шуданд. Қиматҳои мувофиқ дар катакчаҳои ҷадвали 3.27, ки дар буриши сатрҳо ва сутунҳо ҷойгиранд, оварда мешаванд. Дар ҷадвал бо ранги хокистарӣ масофаи наздиктарин байни ҷуфти асарҳои санҷишӣ ва коллексияи аввала қайд карда шудааст.

Натиҷаҳои бадастомада нишон медиҳанд, ки ҳамсояҳои наздиктарини[5] панҷ асари аввал танҳо бо ҷуфти матнҳои коллексияи аввала аз ҷиҳати забон бо онҳо якҷинсаанд. Дар мавриди матни руминӣ (Text_Ro) бошад, тамоми масофаҳои он то даҳ матни коллексияи матнҳо аз қиммати $\gamma^{onm}$ зиёд буданд, нигаред (3.47). Аз

[5] Воронцов, К.В. Математические методы обучения по прецедентам // [Манбаи электронӣ] – Реҷаи дастрасӣ: http://www.ccas.ru/voron ва Дьяконов, А.Г. Анализ данных, обучение по прецедентам, логические игры, системы WEKA, RapidMiner и MatLab (Практикум на ЭВМ кафедры математических методов прогнозирования) // Учебное пособие, М.: Издательский отдел факультета ВМК МГУ имени М.В. Ломоносова, 2010, 278 с.

ин рӯ, тавре ки интизор мерафт, барои Text_Ro дар коллексия ягон объекти якҷинса вуҷуд надошт. Бо вуҷуди ин, ҷолиби қайд он аст, ки γ-таснифгар ҳамчун ҳамсояи наздиктарини он ду асар es_1 ва es_2 ба забони испанӣ ва ду it_1 ва it_2 ба итолиёвӣ нишон дод.

Ҷадвали 3.27. – Масофаҳо байни матнҳо дар коллексияи ***C*** ва шаш асари санҷишии ба таври тасодуфӣ гирифташуда

| **Матнҳо** | | **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 |

**Хулоса.** Пас, γ-таснифгар дар асосӣ $\gamma = \gamma^{onm}$ дар матнҳои тасодуфии гирифташуда бо CP дар асоси басомади 26 ҳарфҳои асосии лотинӣ қобилияти 100% муайян кардани забони асарҳоро тасдиқ кард.

Ҳамин тариқ, сегонаи математикӣ бо ҳамроҳии CP-и матнҳое, ки бо қатори чандомади 26 ҳарфҳои лотинӣ ифода шудаанд, формулаҳои (1.1) – (1.4) барои ҳисоб кардани масофаи байни матнҳо ва алгоритми муайян кардани матнҳои якҷинса барои ҳалли самарабахши масъалаи гузошташуда хеле мувофиқанд. Муаллиф изҳори боварии онро мекунад, ки зиёд шудани миқдори матнҳо монеа барои татбиқи бомуваффақияти γ-таснифгар на танҳо барои шинохти забонҳо, балки барои муайян кардани навъҳои гуногуни якҷинсагии ҳуҷҷатҳои матнӣ нахоҳад шуд. Натиҷаҳои монанд дар **§§ 3.2.1 – 3.2.5** барои забонҳо бо хатти кириллӣ, инчунин бо хати лотинӣ, вале барои дигар мисоли коллексияи матнӣ ба даст оварда шудааст.

Дар **§ 3.2.6**, дар мисоли коллексияи хурде, ки аз 26 матнҳо ба 13 забонҳо (бо 2 асар аз ҳар як забон) ба таври тасодуфӣ тартиб дода шуда, татбиқи γ-таснифгар дар асоси басомади алифбои лотиние, ки барои ҳамаи забонҳо универсалӣ аст, барои шинохти автоматии мансубияти матнҳо ба як гурӯҳи забонҳои славянӣ муқаррар карда шудааст.

Дар **§ 3.3** дар мисоли коллексияи намунавии матнҳо ба забонҳои русӣ ва тоҷикӣ ва тарҷумаи онҳо ба тоҷикӣ ва русӣ бо истифода аз γ-таснифгар ва CP, ки тақсимоти чандомади ҳарфҳои униграммаро дар матнҳо тавсиф мекунад, тадқиқи омории “якҷинсагӣ”-и асарҳои нусхаи асл ва тарҷумашуда гузаронида шуд.

Дар **§ 3.4**, татбиқи γ-таснифгар барои шинохти автоматии рамзи корҳои илмӣ дар асоси қатори чандомади униграммаҳо муайян карда мешавад. Маърузаҳои илмӣ, авторефератҳои олимони гуногун, ки ба забони русӣ навишта шудаанд, гирифта шуданд. Авторефератҳо дар соҳаҳои зерини илмӣ гирифта шудаанд:

таърих, педагогика, сиёсатшиносӣ, филология ва иқтисодиёт. Барои таҷриба коллексияи 10 автореферат, ки ба 5 рамзи ихтисосҳо тааллуқ доранд, барои ҳар як рамз 2 авторефератӣ гирифта шуд:

рамзи 07.00.02: (таърих): 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.)”.

рамзи 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”.

рамзи 23.00.01: (сиёсатшиносӣ): 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”.

рамзи 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”.

рамзи 08.00.01: (иқтисодиёт): 1. Ермакова Е.М. “Особенности современного рынка труда в рыночной и переходной экономике”, 2. Яськин А.В. “Институциональный фактор экономического выбора на современных рынках”.

Маҷмӯи 10 авторефератҳоро ***А-коллексия (модел) меномем.***

Аз рӯи формулаҳои (1.1) – (1.4) 45 ҷуфти масофаҳои $\rho\,(T_1, T_2)$ байни авторефератҳои коллексияи ***А*** ҳисоб карда шуда дар ҷадвали 3.31 оварда шудаанд:

Ҷадвали 3.31. – Масофаи байни авторефератҳои коллексияи ***А***

| **Рамзҳо (Авторефератҳо)** | | **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** | |

Дар асоси маълумоти ҷадвали 3.31 натиҷаҳои зерин ба даст оварда шуданд:

– маҷмӯи ҳамаи ҷуфти масофаҳо дар ҳудуди [0.0393, 0.2341] дохил мешаванд, ки дар ин ҳангом масофаи минималӣ байни рамзҳои 23.00.01 “Автореферат-1” ва 23.00.01 “Автореферат-2” амалӣ карда мешавад ва масофаи максималӣ бошад, байни рамзҳои 10.01.01 “Автореферат-1” ва 08.00.01 “Автореферат-1” аст;

– қиммати оптималии нимфосилаи $\gamma$ дар дохили ҳудудӣ зерин шуд:

$$\gamma^{\text{опт}} \in [0.0610;\ 0.0614);$$

Дар ҷадвали 3.31 катакчаҳои хокистарранг (ки онҳо 6-то мебошанд) вайрон кардани фарзияти таҳияшударо барои ҷуфти мувофиқи авторефератҳо нишон медиҳанд ва аз ин рӯ чунин шуд

$$\tau = \tau_{min} = 6,$$

– дар натиҷа, нишондиҳандаи самаранокии модели математикии шинохти рамзи авторефератҳои дар ин кори илмӣ пешниҳодшуда баробар шуд

$$\pi = \pi_{max} = 0.87$$

**Санҷиш.** Ҳамин тариқ, натиҷаҳои бахши қаблӣ нишон медиҳанд, ки танзимкунии (омӯзиши) γ-таснифгар дар коллексияи хурди матнҳои ***A*** бо муваффақ гузашт. Барои санҷиши таснифгар авторефератҳои зерин гирифта шуданд (ҳамаи онҳо зери рақами 3 навишта шудаанд, то нишон дода шавад, ки онҳо авторефератҳои сеюм аз рамзҳои ихтисосии мувофиқ мебошанд):

*рамзи 07.00.02:* З. Аракелян М.А. “Политическая полиция Российской империи в борьбе с революционным подпольем в 1881-1905 гг.”;

*рамзи 13.00.01:* З. Дуда И.В. “Формирование ценностных ориентаций больных сколиозом школьников в учебно-воспитательном процессе школы-интерната”;

*рамзи 23.00.01*: З. Андреев М.Г. “Роль средств массовой информации в формировании позитивного образа некоммерческих организаций в современной России”;

*рамзи 10.01.01:* З. Левина Е.Н. “Проблема биографизма в творчестве И.С. Тургенева 1840-1850-х годов”;

*рамзи 08.00.01:* З. Добролежа Е.В. “Управление ресурсным обеспечением экономики региона”.

Пас аз ташаккули СР-и авторефератҳо, ки барои санҷидан ва ҳисоб кардани масофа аз рӯи формулаи (1.4) пешбинӣ шудаанд, ҷадвали зерини масофа аз ҳар як авторефератҳои санҷидашуда то ҳамаи 10 авторефератҳои коллексияи аввала оварда мешавад.

Ҷадвали 3.32. – Масофа байни авторефератҳои коллексия ва авторефератҳои санҷидашуда

| **Рамзҳо (Авторефератҳо)** | | **07.00.02** | **13.00.01** | **23.00.01** | **10.01.01** | **08.00.01** |
|---|---|---|---|---|---|---|
| | | З | З | З | З | З |
| **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** |

Дар ҷадвали 3.32 бо ранги хокистарӣ масофаи минималии ҷуфти ячейкаҳо аз авторефератҳои санҷидашуда то авторефератҳои коллексияи ***A мувофиқ*** буда оварда шудааст**.**

Ҳамин тавр, барои автореферати 07.00.02(3) ҳамсояи наздиктарин автореферати 07.00.02(1); барои автореферати 13.00.01(3) ҳамсояи наздиктарин автореферати 08.00.01(2); барои автореферати 23.00.01(3) ҳамсояи наздиктарин автореферати 23.00.01(2); барои автореферати 10.01.01(3) ҳамсояи наздиктарин автореферати 10.01.01(1); барои автореферати 08.00.01(3) ҳамсояи наздиктарин автореферати 08.00.01(2) шуданд.

Мувофиқи маълумоти ҷадвали 3.32 усули ҳамсояи наздиктарин барои муайян кардани рамзҳо аз 5 автореферати санҷидашуда 4-тоаш дуруст ва 1 автореферат хато мебошад.

**Хулоса.** γ-таснифгар бо қимати муқарраршудаи $\gamma = \gamma^{\text{опт}}$ дар авторефератҳои тасодуфии гирифташуда санҷида шуд, ки қобилияти 87% шинохтани рамзи ихтисоси авторефератҳоро тасдиқ кард.

Дар **§ 3.5** дар асоси истифодаи γ-таснифгар дар коркарди 68 асари 7 мактаби адабӣ барои муайянкунии муаллиф ва шинохти услуб дар доираи адабиёти тоҷику форс баҳогузорӣ карда шудааст. Объекти тадқиқоти ин параграф аз силсилаи шоҳасарҳои назми классикии форсии мактаби адибони Хуросон, Ироқ ва Ҳинд, инчунин осори мактаби насри классикӣ, сабки омехта, назми муосир ва насри муосир иборат буда, ки аз рӯйи вақт пайдарпай меоянд. Дар **хулоса** дар ҷадвали зерин аз 7 услуб танҳо 4-тои онро меорем.

Ҷадвали 3.33. – Масофаҳо байни CP-и асарҳои чаҳор услуб

| Муаллифон | Услуби классикии Хуросонӣ | | | | | | Услуби классикии Ироқӣ | | | | Насри классикӣ | | | Насри муосир | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 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** | |

Дар ҷадвал 2 асари А. Рӯдакӣ ва 4 асари А. Фирдавсӣ (мактаби Хуросон), 2 адад аз У. Хайём ва Ҳ. Шерозӣ (мактаби Ироқӣ), 2 адад аз У. Кайковус ва 1 асари М. Ғазолӣ (мактаби насри классикӣ) ва ниҳоят 3 аз С. Айнӣ (мактаби насри

муосир) нишон дода шудааст. Барои чунин омезиши муаллифон бо осори худ қимати $\gamma \in [2.8324; 2.8385)$ ёфта шуд, ки барои он фарзияти якҷинса будани услуб 100% иҷро мешавад. Аз ин рӯ, ҷадвали ҷамъбастӣ ҳамчун нусхаи намунавии асосгузорони мактабҳои адабии тоҷику форс тафсир шудааст.

Дар **боби 4** *"Таҳқиқи қонуниятҳои омории шинохти якҷинсагии матнҳо дар корпуси эҷодиёти осори адабӣ"* чунин ҳалли масъалаи муҳокимашаванда на дар коллексияи матнҳои кам, балки дар корпуси асарҳои бадеӣ мавриди барраси қарор гирифтааст, ки оё метавон натиҷаи қаноатбахш ба даст овард.

Дар **§ 4.1** татбиқи γ-таснифгарро барои шинохти автоматии забони асар дар асоси басомади 26 алифбои умуми бо графикаи кириллӣ дар мисоли корпуси иборат аз 70 матн дар 20 забон (бо 8 асарӣ дар 5 забонҳо: белорусӣ, булғорӣ, русӣ, тоҷикӣ ва украинӣ ва 2 асарӣ ба 15 забони дигар) дида баромада шуд. Дар мисоли пайкара ҳолатҳои 5, 10, 20 забонҳо, инчунин бо 10, 20, 40 матнҳо гирифта шуда, хусусиятҳои истифодаи γ-таснифгарро ҳангоми шинохти забони матн муқаррар карда мешаванд. Барои санҷиши таснифгар се матни тасодуфӣ ба таври иловагӣ гирифта шуданд, ки бо ҳамон забони матнҳои пайкара мувофиқанд. Бо истифода аз усули ҳамсояи (бо масофа) наздиктарин се матни санҷишӣ барои якҷинсагӣ бо ҷуфти мувофиқи асарҳои якзабонӣ тафтиш карда мешаванд. Дар ҷараёни танзимкунӣ амалиётҳои зерин иҷро карда мешаванд:

– коркарди пешакии маълумоти таҷрибавӣ тавассути хориҷ кардани символҳои нолозима аз тамоми асарҳои пайкара ба ғайр аз алифбои кириллӣ;

– ҳисобкунии CP-и (1.1) (чандомадӣ 26 ҳарфи кириллӣ) барои ҳамаи 70 матнҳои коллексияи ***C***;

– ҳисоб кардани формулаҳои (1.2), (1.3) ва (1.4) масофаҳои гуногуни ҷуфти $\rho(T_1, T_2)$ байни матнҳои коллексияи ***C*** (натиҷаҳои таҷрибавӣ дар ҷадвали 4.1 оварда шудаанд);

Ҷадвали 4.1. – Натиҷаи таҷрибаҳо

| **Шумораи забонҳо** | **Шумораи матнҳо** | **Шумораи масофаҳои мутақобила – *L*** | ***τ*-шумораи умумии хатогиҳо** | **Қимати оптималии γ-нимфосила** | ***π*-самаранокии шинохти забон** |
|---|---|---|---|---|---|
| 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 |

Аз рӯи маълумоти ҷадвали 4.1 натиҷаҳои зерин ба даст оварда шуданд:

– қимати оптималии γ чунин шуд

$$\gamma^{опт} \in [0.1001; 0.1638); \qquad (4.7)$$

мувофиқи таърифи **1.4.1** ин маънои онро дорад, ки агар масофаи $\rho(T_1, T_2)$ байни ду матн аз қимати $\gamma^{опт}$ зиёд набошад, пас ҷуфти матнҳо ба як забон тааллуқ доранд; агар зиёд бошад, ба забонҳои гуногун тааллуқ дорад;

– қиммати баландтарини $\pi$=100% коэффитсиенти самаранокии шинохти

забони матн дар корпуси 5 забон бо 10 матн амалӣ карда мешавад;

– коэффисиенти $\pi$ самаранокии забоншиносии асарҳо аз рӯи интихоби 5 забони 20, 40 матн бо қиматҳои аз 92% то 93% муайян карда мешавад, ки ҳамаи хатогиҳо байни матнҳои русӣ ва украинӣ мебошанд. Ин аз он шаҳодат медиҳад, ки ин забонҳо хеле наздиканд;

– коэффитсиенти самаранокии $\pi$ ҳангоми интихоби пайкараи матнии 10, 20 забон бо 20, 40 матн ба 98% ва 99% баробар шуд.

Ҳамин тариқ, сегонаи математикӣ бо гуруҳи СР-и матнҳое, ки бо қатори чандомади 26 ҳарфи кириллӣ ифода ёфтаанд, формулаҳои (1.1) – (1.4) барои ҳисоб кардани масофа байни матнҳо ва алгоритми муайян кардани ҳалли самарабахши матнҳои якҷинса мувофиқанд. Ин тадқиқот нишон медиҳад, ки алифбои ягонаи забонҳоро эҷод кардан мумкин аст.

Дар навбати худ, усули ҳамсояи наздиктарин имкони тақсимоти бехатои се асари иловагӣ гирифташударо аз рӯи забонҳо нишон дод. Муаллиф изҳори боварии онро мекунад, ки зиёд шудани миқдори матнҳо монеа барои татбиқи бомуваффақияти γ-таснифгар на танҳо барои шинохти забонҳо, балки барои муайян кардани навъҳои гуногуни якҷинсагии ҳуҷҷатҳои матнӣ нахоҳад шуд.

Дар **§ 4.2** татбиқи γ-таснифгарро барои шинохти автоматии забони асар дар асоси басомади 26 ҳарфи умумии алифбои лотинӣ бо истифода аз мисоли пайкараи иборат аз 70 матн дар 20 забон (8 асар дар 5 забонҳо: англисӣ, венгерӣ, лотинӣ, литвагӣ ва голландӣ ва 2 асар ба 15 забони дигар) муқаррар карда мешавад.

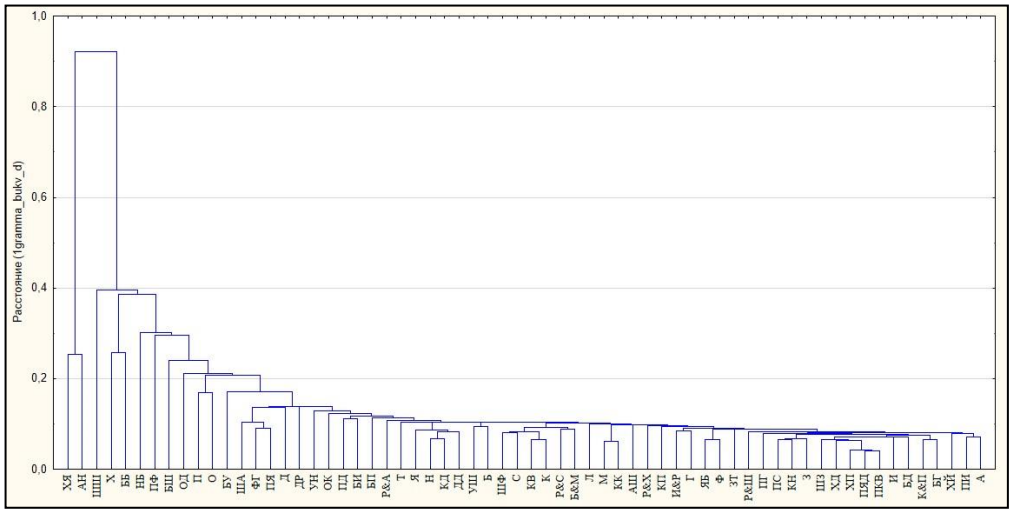
Дар параграфӣ **4.3** татбиқи γ-таснифгар барои шинохти автоматии муаллифи асар аз рӯи басомади 26 ҳарфи алифбои кириллӣ бо истифода аз мисоли пайкараи иборат аз 70 матни шеърии 20 муаллифи тоҷику форс (8 асарҳои 5 муаллиф: А. Суруш, А. Фирдавсӣ, К. Хуҷандӣ, Л. Шералӣ ва Ҷ. Румӣ ва 2 асари 15 муаллифи дигар) муқаррар шудааст. Дар мисоли пайкара интихоби 5, 10, 20 муаллифони асар, инчунин 10, 20, 40 матн баррасӣ гардида, хусусиятҳои истифодабарии γ-таснифгар ҳангоми шинохти муаллифи матн муайян карда мешаванд. Барои санҷидани таснифгар ба таври иловагӣ се матни тасодуфии дигари ҳамон муаллифон гирифта шуд. Бо истифода аз усули ҳамсояи (бо масофа) наздиктарин се матни иловагӣ гирифта шуда барои якҷинса будан бо ҷуфти мувофиқи асарҳои муаллифон тафтиш карда мешавад.

Дар **§ 4.4** муқаддима ва 63 ашъори “Шоҳнома”-и А. Фирдавсӣ бо СР аз рӯи қатори ҳарфҳои алифбои кирилии забони тоҷикӣ дар онҳо гузошта мешавад. Алгоритми таснифгари иерархии агломеративиро истифода мебарем. Ҳамчун масофаи байни объектҳо, мо усули γ-таснифгари ададҳои тасодуфии дискретиро мегирем. Ин параграф танҳо ду қисматро доро аст. Маълумоти гирифташуда дар ҷадвал (матрисаи масофаҳо) ҷойгир карда мешавад. Бо истифода аз усули ҳамсояи наздиктарин аз рӯи матритсаи масофаҳо класстеризатсияи иерархии қисмҳои таркибии асарҳо амалӣ карда мешавад.

Дар расми 4.1 қад-қади меҳвари абсиссаҳо, бо аломати ихтисоршуда, номи шеърҳо мувофиқи принсипи ҳамсояҳои ба ҳам наздиктарин ҷойгир карда шуда, қад-қади меҳвари ординатҳо бошад, шкалаи масофаҳои мутақобилаи байни

шеърҳо оварда шудаанд.

Аз 64 воҳиди таркибии асари “Шоҳнома” достонҳои “Подшоҳи Яздгирд” (ПЯД) ва “Подшоҳи Кайхусрав” (ПКВ) “якҷинса” ба назар мерасанд, ки масофаи байни онҳо $\rho((\text{ПЯД}),(\text{ПКВ})) = 0.0128$ дар қиёс бо дигар асарҳо наздиктарин будааст. Дар баробари ин, достонҳои “Подшоҳии Ардашери Некукор” (АН) ва “Подшоҳии Шопур ибни Шопур” (ШШ), ки дар масофаи аз ҳама дур $\rho((\text{АН}),(\text{ШШ})) = 0.4021$ ҷойгиранд.




Расми 4.1. – Натиҷаи таснифи иерархии достонҳо дар шакли дендрограмма

Сабаби эҳтимолии чунин фосилаи калон байни онҳо дар он аст, ки ҳаҷми ин шеърҳо хеле хурд буда, дар (АН) 181 калима ва дар (ШШ) 352 калима иборат аст. Дар ин замина, ПЯД 9474 калима ва (ПКВ) 35991 калимаро дар бар мегирад. Барои масофаҳо ҳисоби миёна $\rho = 0.0851$ мебошад.

Дар **§ 4.5** тадқиқоти мо оид ба омӯзиши шабакаҳои нейронии ашъори “Шоҳнома”-и А. Фирдавсӣ ва асари сунъӣ тартиб додашудаи он шоир ифода меёбад. Модел вобастагии дарозмуҳлатӣ ва хусусиятҳои синтаксисии пайкарарo омӯхтааст. Самаранокии таснифоти ашъори нав дар барномаи компютерии **“TTA”** (tajik text author) барои муайян кардани муаллифи матн муқаррар карда шуд. Матнҳои ба таври сунъӣ тавлидшудаи $\tilde{T}_1$ ва $\tilde{T}_2$, ки дар барномаи компютерии **“TTA”** санҷида шуд, ба натиҷаҳои зерин муяссар гаштем, ки дар ҷадвали 4.7 оварда шудааст. Натиҷаҳои санҷиш нишон доданд, ки масофаи матнҳои ба таври сунъӣ тавлидшудаи $\tilde{T}_1$ ва $\tilde{T}_2$ то осори А. Фирдавсӣ (Р&С, Б&М) ҳангоми истифодаи униграммаҳо хеле наздиканд (<0.07).

Ҷадвали 4.7. – Таъсирбахшии шеъри ба таври сунъӣ тавлидшуда

| Матни сунъӣ | *N*-граммаҳо | Самарабахшӣ | А. Фирдавсӣ | | Ҷ. Румӣ | | … |
|---|---|---|---|---|---|---|---|
| | | | Р&С | Б&М | ММ1 | ММ2 | |
| $\tilde{T}_1$ (1001 слов) | 1gr. бо пр. | 93 | 0.0656 | 0.0596 | 0.1592 | 0.1524 | |
| | 2gr. бо пр. | 93 | 0.4204 | 0.3627 | 1.0171 | 0.9481 | |
| | 3gr. бо пр. | 96 | 2.5941 | 2.2471 | 6.1823 | 5.7627 | |
| $\tilde{T}_2$ (5002 слов) | 1gr. бо пр. | 100 | 0.0639 | 0.0499 | 0.1537 | 0.1472 | |
| | 2gr. бо пр. | 100 | 0.4048 | 0.3344 | 0.9528 | 0.9196 | |
| | 3gr. бо пр. | 100 | 2.4507 | 2.1481 | 6.0758 | 5.8107 | |

Идомаи ҷадвали 4.7.

| | А. Суруш | | С. Айнӣ | | 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 |

Инчунин, самаранокии таснифоти муайян кардани муаллифи матн 93-100%-ро нишон дод. Ин аз он гувоҳӣ медиҳад, ки сифати матнҳои ба таври сунъӣ тавлидшудаи $\tilde{T}$, ки бо шабакаи нейронии рекурентии LSTM омӯзонидашуда, тавонист баъзе аз хусусиятҳои синтаксисӣ ва услубии достонҳои Шоҳномаро биомӯзад ва комплекси барномаҳои “**ТТА”** ба эҳтимоли зиёд тавонист А. Фирдавсиро ҳамчун муаллиф аз ин матнҳо муайян кунад.

Дар **боби 5** “*Таҳқиқи таъсири тартиби СР-и матн ба муайян кардани якҷинса будани асар”*, бо истифода аз мисоли коллексияи моделӣ, ки тавсифи миқдории асарҳо дар асоси вариантҳои гуногуни батартибории элементҳои алифбо хусусиятҳои истифодаи γ-таснифгар ҳангоми шинохти муаллифи матн ошкор карда шудаанд.

Дар **§ 5.1** дар мисоли коллексияи хурди матнҳо, ки тавсифи миқдории асарҳои онҳо дар асоси вариантҳои гуногуни батартибории $N$-граммаҳои ҳарфи ($N = 1, 2, 3$) бо фосилаҳо, хусусиятҳои истифодабарии γ-таснифгар ҳангоми шинохти муаллифи матн оварда шудаанд.

**5.1.4. Маҷмӯи** $N$-граммаҳо ($N = 1, 2, 3$) вобаста ба батартиборii элементҳояшон дар 4 вариант дида баромада мешаванд:

1) элементҳо бо тартиби алифбо батартиб оварда шуда бо пробел (фосила) ҳамчун элементи охирон ҷойгир карда мешаванд (ҳамчун $ABC$ ифода карда мешавад)[6];

2) элементҳо бо тартиби баръакси алифбо батартиб оварда шуда бо пробел (фосила) ҳамчун элементи аввал ҷойгир карда мешаванд (ҳамчун $CBA$ ифода карда мешавад)[7];

3) элементҳо бо тартиби камшавии басомади вохӯрии онҳо дар матн ҷойгир карда мешаванд (бо рамзи “ ↘” ифода карда мешавад);

4) элементҳо бо тартиби афзоиши басомади вохӯрии онҳо дар матн ҷойгир карда мешаванд (бо рамзи “ ↗” ифода карда мешавад).

Натиҷаҳои коркарди автоматии коллексияи моделии матнҳо дар ҷадвалҳои 5.1-5.3 нишон дода шудаанд.

[6] Барои биграммаҳо ва триграммаҳо – бо ду ва се фосила дар охир.
[7] Барои биграммҳо ва триграммаҳо – бо ду ва се фосила дар аввал.

Ҷадвали 5.1. – Қимматҳои $\pi$ ва $\gamma$ барои асарҳои назми классикӣ

| Элементҳои матн | Шумораи элементҳои алифбо | Тартиби элементҳои алифбо | $\pi$-саҳеҳии шиноҳтӣ муаллиф | Қимати оптималии γ-нимфосила |
|---|---|---|---|---|
| униграммаҳо | 36 | A B C | 0.98 | [0.0354; 0.0447) |
| | | C B A | 0.98 | [0.0354; 0.0447) |
| | | бо ↘ | 0.98 | [0.0337; 0.0342) |
| | | бо ↗ | 0.98 | [0.0337; 0.0342) |
| биграммаҳо | 1296 | A B C | 0.98 | [0.2987; 0.3551) |
| | | C B A | 0.98 | [0.2987; 0.3551) |
| | | бо ↘ | 0.96 | [0.2065; 0.2212) |
| | | бо ↗ | 0.96 | [0.2065; 0.2212) |
| триграммаҳо | 46656 | A B C | 1.00 | [2.1630; 2.1648) |
| | | C B A | 1.00 | [2.1630; 2.1648) |
| | | бо ↘ | 0.96 | [1.2426; 1.4051) |
| | | бо ↗ | 0.96 | [1.2426; 1.4051) |

Дар ин ҷадвал, инчунин дар ду ҷадвалӣ навбатӣ, дар сутуни сеюм барои тавсифи тартиби элементҳои алифбо, аломатҳои дар банди **5.1.4** ифода шуда оварда мешаванд.

Ҷадвали 5.2. – Қимматҳои $\pi$ ва $\gamma$ барои асарҳои назми муосир

| Элементҳои матн | Шумораи элементҳои алифбо | Тартиби элементҳои алифбо | $\pi$-саҳеҳии шиноҳтӣ муаллиф | Қимати оптималии γ-нимфосила |
|---|---|---|---|---|
| униграммаҳо | 36 | A B C | 0.98 | [0.0268; 0.0423) |
| | | C B A | 0.98 | [0.0268; 0.0423) |
| | | бо ↘ | 0.98 | [0.0384; 0.0415) |
| | | бо ↗ | 0.98 | [0.0384; 0.0415) |
| биграммаҳо | 1296 | A B C | 0.98 | [0.2318; 0.2816) |
| | | C B A | 0.98 | [0.2318; 0.2816) |
| | | бо ↘ | 0.98 | [0.2484; 0.2745) |
| | | бо ↗ | 0.98 | [0.2484; 0.2745) |
| триграммаҳо | 46656 | A B C | 0.98 | [1.3885; 1.7054) |
| | | C B A | 0.98 | [1.3885; 1.7054) |
| | | бо ↘ | 0.98 | [1.5556; 1.6453) |
| | | бо ↗ | 0.98 | [1.5556; 1.6453) |

**Хулоса.** Аз натиҷаҳои ҳисобу китоб, ки дар сутунҳои 4 ва 5 оварда шудаанд, хулосаҳои зерин мебароянд:

1) қиммати баландтарини $\pi = 1$ коэффитсиенти самарабахши шиноҳтӣ муаллифи матн барои асарҳои назми классикӣ дар триграммаҳо, ки ҳам аз ABC ва ҳам CBA батартиб оварда шудаанд, амалӣ карда мешавад;

2) қимматҳои коэффитсиенти самаранокии $\pi$ дар батартибории ABC ва CBA-и $N$-граммаҳо ( $N = 1, 2, 3$) баробаранд;

3) қимматҳои коэффитсиенти самаранокии $\pi$ аз рӯи батартибории $N$-граммаҳо ($N = 1, 2, 3$) дар камшавӣ ( ↘) ё афзоиш ( ↗) низ баробаранд;

Ҷадвали 5.3. – Қимматҳои $\pi$ ва $\gamma$ барои асарҳои насри муосир

| Элементҳои матн | Шумораи элементҳои алифбо | Тартиби элементҳои алифбо | $\pi$-саҳеҳии шиноҳтӣ муаллиф | Қимати оптималии γ-нимфосила |
|---|---|---|---|---|
| униграммаҳо | 36 | A B C | 0.96 | [0.0285; 0.0336) |
| | | C B A | 0.96 | [0.0285; 0.0336) |
| | | бо ↘ | 0.91 | [0.0165; 0.0236) |
| | | бо ↗ | 0.91 | [0.0165; 0.0236) |
| биграммаҳо | 1296 | A B C | 0.93 | [0.2216; 0.2272) |
| | | C B A | 0.93 | [0.2216; 0.2272) |
| | | бо ↘ | 0.91 | [0.2386; 0.2568) |
| | | бо ↗ | 0.91 | [0.2386; 0.2568) |
| триграммаҳо | 46656 | A B C | 0.96 | [1.3379; 1.3412) |
| | | C B A | 0.96 | [1.3379; 1.3412) |
| | | бо ↘ | 0.91 | [0.7450; 1.3704) |
| | | бо ↗ | 0.91 | [0.7450; 1.3704) |

4) қимматҳои коэффитсиенти самаранокии $\pi$ дар асоси тартиби ABC ва CBA ҷойгиршавии $N$-граммаҳо ( $N = 1, 2, 3$ ) аз қимматҳое, ки аз рӯи тартиби ҷойгиршавии $N$-граммаҳо ($N = 1, 2, 3$) дар камшавӣ ( ↘) ё афзоиш ( ↗) ба даст омадаанд, кам нестанд;

5) коэффитсиенти саҳеҳии $\pi$ муайян кардани муаллифи асарҳои назми муосир чӣ барои ҳар як $N$-граммаҳо ($N = 1, 2, 3$) ва ҳам барои ҳамаи вариантҳои батартибории онҳо бо қимати 0.98 муайян карда мешавад;

6) коэффитсиентҳои $\pi$ осори насри муосир назар ба осори назми классикӣ ва муосир як андоза пасттар аст;

7) қиммати оптималии нимфосилаи γ барои ду батартибории муқобили ҷойгиршавии $N$-граммаҳо ($N = 1, 2, 3$) якхелаанд.

Аз шумораи зиёди вариантҳои имконпазири батартибории элементҳои матн, танҳо чортои онҳо таҳқиқ карда шуданд: дутои онҳо ба тартиби алифбо алоқаманданд ва дуи дигар ба басомади элементҳо асос ёфтаанд. Маҳз дар ин ду ҳолати батартиборӣ ва баръакси онҳо, масофаи байни ҳама гуна ҷуфтҳои асарҳо баробар шуд, ки дар натиҷа коэффитсиентҳои $\pi$ самаранокии γ-таснифгар баробар шуданд, инчунин қиммати оптималии γ. Дар **§§ 5.2.-5.4.** дигар вариантҳои имконпазир тадқиқ карда мешаванд.

Дар **§§ 6.1-6.7 боби 6** тавсифи муфассали комплекси барномаҳои **“THR”** (text homogeneity recognition), ки барои муайян кардани якҷинсагии матни номаълум пешбини шудааст, дода мешавад.

Дар **охири** диссертатсия хулосаҳо ва натиҷаҳои асосӣ оварда мешаванд.

# ХУЛОСА

**Натиҷаҳои асосии диссертатсия:**

1. Маълумот аз адабиёти гуногун оид ба аломатҳои миқдории матн ва алгоритмҳое, ки барои муайян кардани якҷинсагии асарҳо татбиқ мешаванд, таҳлил карда шуданд. Самтҳои асосии тадқиқот муайян шуданд.

2. Саҳеҳии тадбиқи γ-таснифгар дар маҷмӯи васеи асарҳо барои муайян кардани муаллифони онҳо исбот карда шуд.

3. Самаранокии γ-таснифгар бо дақиқии то 100% барои шинохти муаллифи порчаи матн дар ҳаҷми аз 7000 калима (40000 рамз) то 20 калима (100 рамз) муайян карда шуд.

4. Имконияти комилан кам кардани ҳаҷми протседураҳои ҳисобкунӣ аз ҳисоби истифода бурдани на ҳама, балки фақат элементҳои баландбасомади СР-и матн муқаррар карда шуд.

5. Самаранокии омории тадбиқи қатори чандомади гуногуни алифбои элементҳои матн ва γ-таснифгар (сегонаи математикӣ) барои шинохти аломатҳои дигари якҷинсагии матн ба монандӣ: мавзӯи матн, забон, гурӯҳи забонҳо, асл ва тарҷумаи он, услуби асарҳо ва рамзи асарҳои илмӣ муайян карда шуд.

6. Қонуниятҳои омории муайянкунии муаллиф ва забонҳои асарҳо дар пайкараи осори адабии бадеӣ таҳқиқ карда шуд.

7. Ҳангоми истифода аз таснифгари метрикӣ ва усули ҳамсояи (бо масофа) наздиктарин дар матнҳои тасодуфии барои санҷиш гирифташуда бо дақиқати кофӣ баланд якҷинсагии аломатҳои матн дар коллексияи матнҳои ками гуногун ва пайкараи матнҳо пайдо карда шуд.

8. Самаранокии тадбиқи γ-таснифгар барои муайян кардани муаллиф дар асари сунъӣ тартиб додашудаи “Шоҳнома”-и А. Фирдавсӣ муқаррар гардид.

9. Хусусиятҳои истифодаи γ-таснифкунанда ҳангоми шинохтӣ муаллифи матн дар мисоли коллексияи матнҳои кам, ки дар вариантҳои тавсифи миқдории асарҳои гуногуни батартибовардашудаи $N$-грамм ($N$ = 1, 2, 3)-ҳои ҳарфӣ (бе фосила ва бо фосила) асос ёфтаанд, ошкор карда шуд.

10. Якумин маротиба дар Тоҷикистон дар асоси СР-и гуногуни матн ва γ-таснифкунанда комплекси барномаҳои компютерии ба объект нигаронидашуда барои шинохти (муайянкунии) якҷинсагии шумораи зиёди матнҳои номаълум сохта шуд.

**Тавсияҳо оид ба истифодаи амалии натиҷаҳои таҳқиқот.**

Комплекси тарҳрезишуда барои ба кор бурдан дар худкорсозии раванди коркарди маълумоти матнӣ дар фаъолияти идораи давлатӣ барои муқаррар кардани муаллифи матни номаълум дар соҳаи криминалистика, барои муайян кардани асардуздӣ дар корҳои курсӣ ва дипломӣ ва рисолаҳои номзадӣ ва докторӣ дар соҳаи маориф ва илм, инчунин барои истифода дар омӯзиши проблемаҳои илмии гуногун, ки бо масъалаҳои муайян кардани “якҷинсагӣ”-и матни чопӣ алоқамандӣ доранд, тавсия дода мешавад.

## ФЕҲРИСТИ ИНТИШОРОТИ ДОВТАЛАБИ ДАРЁФТИ ДАРАҶАИ ИЛМӢ

### Дар журналҳои ба КОА дахлдор:

**[1-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. Косимов, А.А.** Определение шифра 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-М]. 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-М]. 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.

**Монография ва дастурҳои таълимӣ:**

**[35-М]. 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-М]. Косимов, А.А.** Амалияи барномасозӣ дар забони Python [Матн] / **А.А. Косимов** // ДТТ ба номи ак. М.С. Осимӣ, Душанбе. – 2023. – 163 с.

**[37-М]. 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-М]. 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.

**Дигар нашрияҳо, таълифот ва маҷаллаҳои конфронс:**

**[39-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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.

**Шаҳодатномаҳо дар бораи бақайдгирии давлатии барномаи компютерӣ барои МЭҲ:**

**[69-М]. 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-М]. 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-М]. 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-М]. 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-М]. 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.

# ШАРҲИ МУХТАСАР

**ба рисолаи диссертатсионии Қосимов Абдунаби Абдурауфович дар мавзӯи “Қонуниятҳои омории шинохти якҷинсагии матн бо истифода аз γ-таснифгар” барои дарёфти дараҷаи илмии доктори илмҳои техникӣ аз рӯйи ихтисоси 05.13.11 - Таъминоти математикӣ ва барномавии мошинҳои ҳисоббарор, муҷтамаъҳо ва шабакаҳои компютерӣ**

**Калидвожаҳо:** матн, забон, пайкара, алифбо, симои рақамии матн, чандомад, фарзияти якҷинсагӣ, γ-таснифгар, омӯзиш, шинохт, санҷиши таснифгар, баҳодиҳии саҳеҳӣ.

**Объекти таҳқиқот:** корпуси матнҳои чопӣ ва хусусиятҳои он дар забонҳои гуногун.

**Усулҳои тадқиқот:** омӯзиши мошинӣ, кодиронии маълумот, усулҳои омори математикӣ, таҷрибаҳои ҳисобӣ, назарияи маҷмӯъ, барномасозии ба объект нигаронидашуда.

Дар рисола раванди муайянсозии якҷинсагии матн алгоритсозӣ шуда, ки барои он:

– информативнокии аломатҳои ғайрианъанавии забоншиносӣ барои тавсифи миқдории матн таҳқиқ карда мешавад;

– дар маҷмӯи васеи асарҳо самаранокии истифодаи γ-таснифгар барои шинохти муаллифони асарҳои пурра муқаррар карда мешавад;

– бо мақсади комилан кам кардани ҳаҷми протседураҳои ҳисобҳо имкони самаранокии истифодаи на ҳама, балки танҳо элементҳои баландбасомади алифбо тадқиқ карда мешавад;

– самаранокии γ-таснифгарро барои шинохти муаллифи порчаи матн дар ҳаҷми аз 7000 калима (40000 рамз) то 20 калима (100 рамз) муайян карда мешавад;

– саҳеҳии омории тадбиқи γ-таснифгар (сегонаи математикӣ) барои шинохти аломатҳои дигари якҷинсагии матн ба монандӣ: мавзӯи матн, забон, гурӯҳи забонҳо, асл ва тарҷумаи он, услуби асарҳо ва рамзи асарҳои илмӣ таҳқиқ карда мешавад;

– қонуниятҳои омории шинохти муаллиф ва забони асарҳо дар корпуси эҷодиёти осори адабӣ омӯхта мешаванд.

**Мақсади кор:** сохтани комплекси барномаҳои компютерии ба объект нигаронида шуда барои муайян кардани якҷинсагии матн.

**Соҳаи татбиқшаванда:** комплекси тарҳрезишуда барои ба кор бурдан дар худкорсозии раванди коркарди маълумоти матнӣ дар фаъолияти идораи давлатӣ барои муқаррар кардани муаллифи матни номаълум дар соҳаи криминалистика, барои муайян кардани асардуздӣ дар корҳои курсӣ ва дипломӣ ва рисолаҳои номзадӣ ва докторӣ дар соҳаи маориф ва илм, инчунин барои истифода дар омӯзиши проблемаҳои илмии гуногун, ки бо масъалаҳои муайян кардани якҷинсагии матни чопи алоқамандӣ доранд, тавсия дода мешавад.

**АННОТАЦИЯ**

**диссертации Косимова Абдунаби Абдурауфовича на тему «Статистические закономерности распознавания однородности текстов с помощью γ-классификатора» на соискание ученой степени доктора технических наук по специальности 05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей**

**Ключевые слова:** текст, язык, корпус, алфавит, цифровой портрет текста, частотность, гипотеза однородности, γ-классификатор, обучение, распознавание, тестирование классификатора, оценка эффективности.

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

**Методы исследования:** машинное обучение, кодирование информации, методы математической статистики, вычислительный эксперимент, теории множеств, объектно-ориентированное программирование.

В работе алгоритмизируется процесс распознавания однородности произведений, для чего:

– исследуется информативность нетрадиционных лингвистических признаков на предмет количественного описания текстов;

– на расширенной коллекции произведений устанавливается эффективность применения γ-классификатора для распознавания авторов полноценных произведений;

– для сокращения объёма вычислительных процедур устанавливается возможность эффективного использования только высокочастотных элементов алфавита;

– устанавливается эффективность γ-классификатора, способного распознавать автора текстового фрагмента размером от величины в 7000 слов (40000 символов) вплоть до 20 слов (100 символов);

– устанавливается статистическая эффективность применения 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;

– исследуются статистические закономерности распознавания авторов и языков произведений на корпусах художественных литературных произведений;

**Целью работы** является создание объектно-ориентированного компьютерного программного комплекса для распознавания однородности текста.

**Область применения:** 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.

# ANNOTATION

**on the dissertation of Kosimov Abdunabi Abduraufovich on the theme "Statistical patterns of recognition of text homogeneity using a γ-classifier" for doctor a degree of technical sciences on a specialty 05.13.11 – Mathematical and software of computers, complexes and computer networks**

**Keywords:** text, language, corpus, alphabet, digital portrait of the text, frequency, homogeneity hypothesis, γ-classifier, learning, recognition, classifier testing, performance evaluation.

**Object of the research:** corpus of printed texts and its characteristics in different languages.

**Research methods:** machine learning, information coding, methods of mathematical statistics, computational experiment, set theory, object-oriented programming.

In the work, the process of recognition of the homogeneity of texts is algorithmized, for which:

– the informativity of non-traditional linguistic features is studied for the quantitative description of texts;

– on an extended collection of works, the effectiveness of the use of the γ-classifier is established for recognizing authors as full-fledged works;

- to reduce the volume of computational procedures, the possibility of effective use of only high-frequency elements of the alphabet is established;

– the efficiency of the γ-classifier is established, capable of recognizing the author of a text fragment in size from 7000 words (40000 symbols) up to 20 words (100 symbols);

- the statistical efficiency of using the γ-classifier (mathematical triad) is established for recognizing other signs of homogeneity, such as text topics, language, groups of languages, the original and its translation, the style of works and ciphers of scientific works;

- the statistical patterns of recognition of authors and languages of works on the corpus of artistic literary works are studied;

**Purpose of the research:** is to create an object-oriented computer software package for recognizing text homogeneity.

**Scope:** the designed complex is recommended for use in automating the process of processing text information in public administration, for establishing the authorship of anonymous texts in the field of forensic science, for detecting plagiarism in course and diploma projects, for submitting candidate and doctoral dissertations in the field of education and science, as well as for use in the study of a wide variety of scientific problems related to the recognition of "homogeneous" printed texts.