МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

# ИССЛЕДОВАНИЕ И ПОСТРОЕНИЕ СИСТЕМЫ РЕКОМЕНДАЦИЙ ДЛЯ ТЕКСТОВЫХ ДАННЫХ С УЧЕТОМ СПЕЦИФИКИ ОБЛАСТИ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студентки 2 курса 273 группы
направления 02.04.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Киреевой Маргариты Николаевны

Научный руководитель:
к.ф.-м.н., доцент _____ М.В. Огнева

подпись, дата

Зав. кафедрой:
к.ф.-м.н., доцент _____ М.В. Огнева

подпись, дата

Саратов 2025

# ВВЕДЕНИЕ

**Актуальность темы.** В современном мире, где объемы информации постоянно растут, рекомендательные системы стали неотъемлемой частью нашей повседневной жизни. Они широко используются в бизнесе, так как позволяют предложить потенциальному клиенту наиболее интересующие его товары, показывать таргетированную рекламу, ориентируясь на его предпочтения и вкусы, определенные из предыдущего опыта общения. Такие системы удобны пользователям, поскольку позволяют выделить из огромного объема информации интересующий контент. Яркими представителями сервисов с рекомендательными системами являются Netflix и Кинопоиск, предлагающие пользователям фильмы на основе оценок уже просмотренного контента, YouTube и Яндекс.Музыка, рекомендующие видео и музыку.

Одним из самых популярных генерируемых пользователями контентом также является текст, например, книги, статьи или описания фильмов. Книжные рекомендации важны, ведь существует множество потенциальных читателей, которым из-за обилия информации в Интернете очень трудно найти книги, которые могли бы им понравиться и привить привычку к чтению. Однако, существующие рекомендательные системы, которые работают с текстовыми данными, часто не учитывают тематику и основываются только на оценках пользователей. При этом такие важные качественные характеристики текста, как его читабельность, грамотность, объем словарного запаса автора и другие, вообще не рассматриваются в рекомендациях. Это приводит к тому, что пользователи получают менее качественные рекомендации и не могут найти интересные и нужные для себя тексты.

В связи с этим возникает потребность в создании системы рекомендаций, которая решала бы эти проблемы. Такая система может значительно повысить качество предоставляемых рекомендаций и улучшить пользовательский опыт. Её разработка актуальна для исследований и практического применения в сервисах для чтения или покупки книг, статей и т.п.

**Цель магистерской работы**: исследование и построение системы рекомендаций для текстовых данных с учетом специфики области.

Поставленная цель определила **следующие задачи**:

1. Выполнить обзор существующих рекомендательных систем для книг.
2. Выполнить обзор источников по данной теме.
3. Выполнить обзор методов тематического моделирования и векторного представления текстов.
4. Рассмотреть методы подсчета коэффициентов читабельности и других качественных характеристик текста.
5. Провести исследование применимости рассмотренных методов тематического моделирования и векторизации текстов без коэффициентов читабельности для построения рекомендаций.
6. Провести исследование применимости в задаче классификации рассмотренных методов подсчета коэффициентов читабельности и их комбинирования с методами тематического моделирования и векторизации текстов.
7. Провести исследование применимости в системе рекомендаций рассмотренных методов подсчета коэффициентов читабельности и их комбинирования с методами тематического моделирования и векторизации текстов.
8. Провести сравнительный анализ исследуемых методов, сделать выводы.

**Методологические основы** исследования и построения системы рекомендаций для текстовых данных с учетом специфики области представлены в работах Прокофьевой И.А., Гордеевой О.А., Кудриной М.А. [1], Пуговкиной Е.Д., Белоусова А.А. [2], Федоренко В.И., Киреева В.С. [3], Коршунова А. В., Гомзина А. Г. [4], Ильвовского Д.А. [5], Falk K. [6].

**Практическая значимость магистерской работы** заключается в исследовании и построении системы рекомендаций для текстовых данных,

которая учитывает не только содержание текста, но и его читабельность, а также обеспечивает релевантные рекомендации.

**Структура и объём работы**. Магистерская работа состоит из введения, 6 разделов, заключения, списка использованных источников и 16 приложений. Общий объем работы – 164 страниц, из них 101 страниц – основное содержание, включая 11 рисунков и 47 таблиц, список использованных источников информации – 38 наименований.

## КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

**Первый раздел «Обзор книжных рекомендательных сервисов»** посвящен описанию и сравнению между собой различных рекомендательных сервисов для книг: Букмейт (Яндекс.Книги), Книгогид, Литрес, Фантлаб, Wattpad, LiveLib, Фикбук.

В результате анализа было выявлено, что в большей части рассмотренных рекомендательных сервисов, кроме Фантлаба и Фикбука, рекомендуются книги на основе оценок пользователей или производится фильтрация по жанрам.

Тематика произведений учитывается только в рекомендациях сервисов Литрес, Wattpad и LiveLib с помощью тегов. Также ни на одном сайте не учитывается качество (читабельность, грамотность и т.д.) текста книги, а это тоже зачастую важно для читателя.

**Во втором разделе «Обзор методов, используемых для построения рекомендаций»** рассматривается понятие рекомендательной системы, различные типы рекомендательных систем, приведен обзор методов и технологий, используемых в научных исследованиях, проведенных другими авторами.

Рекомендательные системы подбирают и предлагают пользователю релевантный контент, основываясь на своих знаниях о пользователе, контенте и взаимодействии пользователя и контента.

В **подразделе 2.1** более подробно обозреваются методы тематического моделирования, векторизации текстов и извлечения ключевых слов.

Так, в **подразделе 2.1.1** рассматривается «мешок слов» — метод, в котором текст представляется в виде вектора, где каждая компонента соответствует количеству вхождений определенного слова в текст.

В **подразделе 2.1.2** рассматривается метод TF-IDF (Term Frequency – Inverse Document Frequency). TF вычисляется, как количество раз, которое токен встречается в документе, а IDF как инверсия частоты, с которой этот токен встречается во всех документах коллекции.

В **подразделе 2.1.3** рассматриваются методы тематического моделирования. Тематическое моделирование является способом построения модели коллекции текстовых документов, которая определяет, к каким темам относится каждый из документов.

Латентно-семантическое индексирование (LSI) — усеченное сингулярное разложение матрицы «мешка слов» до матрицы ранга k.

В вероятностном латентно-семантическом индексировании (PLSI) каждый документ может относиться к нескольким темам с некоторой вероятностью. Недостатки PLSI устранены в модели LDA.

Скрытое (латентное) размещение Дирихле (LDA) — генеративная вероятностная модель, которая позволяет выявить основные темы, присутствующие в коллекции, и распределить документы по этим темам.

В **подразделе 2.1.4** рассматриваются модели дистрибутивной семантики. Они используют для получения векторного представления текста гипотезу, что контекст слова определяется его окружением.

Существуют две разновидности модели Word2Vec — Continuous Bag of Words и SkipGram. Первая модель решает задачу предсказания слова $w_i$ на основании контекста (ближайших слов). Вторая модель является противоположной — зная слово $w_i$, найти его контекст.

Позже идея Word2Vec была развита в модель Doc2Vec, в которую помимо векторов слов входят векторы документа.

В **подразделе 2.1.5** рассматриваются методы выделения ключевых слов:

- RAKE (Rapid Automatic Keyword Extraction);
- YAKE (Yet Another Keyword Extractor);
- TextRank;
- PageRank.

В **подразделе 2.1.6** рассматривается косинусное сходство — метрика, которая используется для измерения сходства между двумя векторами. Она основана на косинусе угла между двумя векторами и изменяется от 0 до 1.

В **подразделе 2.3** рассматриваются методы определения качества текста.

Читабельность текста — оценка легкости его восприятия читателями.

$$\text{Индекс Флэша} = 206{,}835 - 1{,}3 \cdot \frac{\text{кол} - \text{во слов}}{\text{кол} - \text{во предложений}} - 60{,}1 \cdot \frac{\text{кол} - \text{во слогов}}{\text{кол} - \text{во слов}}$$

$$\text{Индекс Флэша} - \text{Кинкейда} = 0{,}45 \cdot \frac{\text{кол} - \text{во слов}}{\text{кол} - \text{во предложений}} - 8{,}38 \cdot \frac{\text{кол} - \text{во слогов}}{\text{кол} - \text{во слов}} - 15{,}59$$

$$\text{Индекс Ганнинга} = 0{,}4 \cdot \left( 0{,}78 \cdot \frac{\text{кол} - \text{во слов}}{\text{кол} - \text{во предложений}} + 100 \cdot \frac{\text{кол} - \text{во сложных слов}}{\text{кол} - \text{во слов}} \right)$$

$$SMOG = \sqrt{\frac{64{,}6}{\text{кол} - \text{во предложений}} \cdot \text{кол} - \text{во многосложных слов} + 0{,}05}$$

$$ARI = 6{,}26 \cdot \frac{\text{кол} - \text{во букв}}{\text{кол} - \text{во слов}} + 0{,}2805 \cdot \frac{\text{кол} - \text{во слов}}{\text{кол} - \text{во предложений}} - 31{,}05$$

**Третий раздел «Исследование применимости методов тематического моделирования и векторизации текстов в системе рекомендаций»** посвящен предобработке найденных наборов данных, применению методов «мешок слов», TF-IDF, LDA, Doc2Vec, TextRank для получения рекомендаций и сравнительному анализу методов.

Использовался корпус произведений нарративной прозы XIX века и датасет фанфиков, с применением четырех способов предобработки текста, включавших лемматизацию, фильтрацию частей речи и расширение стоп-слов. Сравнивалось количество рекомендуемых методами фанфиков, так как для текстов XIX века рекомендовать современные фанфики нелогично. Чем меньше фанфиков рекомендуется для прозы XIX века с помощью метода, тем лучше он подходит для построения рекомендаций.

В **подразделах 3.1, 3.2, 3.3, 3.4, 3.5, 3.6** подробно описывается предобработка текстов и применение методов LDA, «мешок слов», TF-IDF, TextRank, Doc2Vec соответственно.

В **подразделе 3.7** проводится сравнительный анализ методов по количеству рекомендуемых фанфиков и времени выполнения.

Лучшие результаты показал метод «Мешок слов» с первым вариантом предобработки (сохранение существительных, прилагательных, глаголов и наречий с минимальным списком стоп-слов). Для всего корпуса классики он рекомендовал 11 фанфиков, а для «Преступления и наказания» Достоевского предлагал его другие произведения. Doc2Vec при 150 эпохах обучения выдавал лишь 18 фанфиков, также сохраняя релевантность рекомендаций. TF-IDF, LDA и TextRank показали более низкие результаты (391, 323, 645).

«Мешок слов» оказался самым быстрым методом (24-122 с), тогда как Doc2Vec при 150 эпохах требовал до 46 минут, а TextRank — до 3 часов. При оценке времени формирования рекомендаций для готовых моделей Doc2Vec и LDA имели преимущество, но для больших наборов данных «мешок слов» может быть предпочтительнее из-за скорости обучения.

**Четвертый раздел «Исследование применимости методов подсчета коэффициентов читабельности и их комбинирования с методами тематического моделирования и векторизации текстов в задаче классификации»** посвящен вычислению коэффициентов читабельности, расширению набора данных и проведению классификации текстов по жанрам, где в качестве признаков используются вектора текстов или темы в тематическом моделировании и коэффициенты читабельности.

В **подразделе 4.1** описывается новый датасет и его предобработка.

В библиотеке Russian Texts Statistics (ruTS) был найден еще один датасет советских хрестоматий с детской литературой. Этот датасет был объединен с двумя предыдущими для дальнейшего анализа. Был добавлен столбец «genre», обозначающий жанр произведения.

В **подразделе 4.2** описывается программная реализация коэффициентов читабельности и их применение. Функции вычисления коэффициентов читабельности Флэша, Флэша-Кинкейда, SMOG, ARI были найдены в библиотеке ruTS, индекс Ганнинга был реализован самостоятельно.

В **подразделах 4.3, 4.4, 4.5, 4.6** подробно описывается классификация методом k ближайших соседей (k = 5), где в качестве признаков

использовались результаты методов LDA, «мешок слов», TF-IDF, Doc2Vec и коэффициенты читабельности соответственно.

В **подразделе 4.7** проводится сравнительный анализ моделей классификации по F-мере.

В таблице 1 представлены лучшие результаты по каждому случаю.

Таблица 1 – Оценка с помощью макро-усреднения F-меры моделей классификации с использованием лучших комбинаций признаков для каждого случая

| Метод | Комбинация признаков | F-мера |
|---|---|---|
| Тематическое моделирование (LDA) | Тема, вероятность темы | 0.81 |
| Векторизация текстов (Doc2Vec) | Векторизованные тексты | 0.82 |
| Тематическое моделирование (LDA) + коэффициенты читабельности | Тема, вероятность темы, индекс SMOG или (SMOG, ARI) | 0.87 |
| Векторизация текстов (Doc2Vec) + коэффициенты читабельности | Векторизованные тексты, индексы SMOG, ARI | 0.84 |

Таким образом, для решения задачи классификации (а в перспективе и для построения системы рекомендаций) наиболее целесообразным подходом является использование в качестве признаков тем и их вероятностей, выделенных с помощью метода LDA, или текстов, векторизованных Doc2Vec, в сочетании с индексами читабельности SMOG, Ганнинга и ARI. Эти комбинации обеспечивают высокую точность классификации и позволяют учитывать как содержание текстов, так и их читабельность. TF-IDF и «мешок слов», а также коэффициенты Флэша и Флэша-Кинкейда продемонстрировали свою неэффективность в контексте рассматриваемой задачи.

**Пятый раздел «Исследование применимости методов подсчета коэффициентов читабельности и их комбинирования с методами тематического моделирования и векторизации текстов в системе рекомендаций»** посвящен расширению набора данных, исследованию применимости использованных ранее методов для получения рекомендаций, сравнительному анализу методов по проценту соответствия жанров и сравнению рекомендаций с оценками рекомендуемых произведений с LiveLib.

В **подразделе 5.1** описывается расширенный набор данных. Для добавления к ранее найденным датасетам были собраны два корпуса по 50 текстов фэнтези и фантастики соответственно.

В **подразделах 5.2, 5.3, 5.4, 5.5** подробно описывается получение рекомендаций с помощью методов LDA, «мешок слов», TF-IDF, Doc2Vec и коэффициентов читабельности соответственно.

Рекомендации с помощью методов векторизации выводятся таким образом: тексты произведений векторизуются, после к векторам добавляются коэффициенты читабельности, и с помощью косинусного сходства определяются похожие произведения. При использовании метода LDA все произведения распределяются по темам с определенной вероятностью, к которой добавляются коэффициенты читабельности, и в рекомендации выводятся произведения одной темы и комбинированного значения вероятности и коэффициентов читабельности с погрешностью 0.1.

Анализируются рекомендации для произведения «Гарри Поттер и узник Азкабана» с учетом и без учета фанфиков (так как их в датасете очень много), а также по всему датасету вычисляется средний процент соответствия жанра рекомендуемых книг исходной книге.

В **подразделе 5.6** проводится сравнительный анализ методов по проценту соответствия жанров. Анализ проводился с двумя вариантами датасета из-за преобладания фанфиков: четыре жанра (без фанфиков) и пять жанров (с фанфиками).

Лучшие результаты были достигнуты с помощью метода Doc2Vec и коэффициентов SMOG, ARI и Ганнинга без фанфиков — 94.51%, с фанфиками Doc2Vec с добавлением индексов Флэша и Флэша-Кинкейда — 93.99%.

К Doc2Vec близок метод LDA: 92.70% без коэффициентов и 92.13% с ними, но без фанфиков его соответствие жанров значительно ниже (78.77%).

«Мешок слов» и TF-IDF показали худшие результаты: 76.25% и 79.14% соответственно, коэффициенты читабельности их не улучшили.

В **подразделе 5.7** проводится сравнение рекомендаций с оценками на LiveLib. Чтобы сравнить качество рекомендаций с помощью рассмотренных методов, также для каждого из них были выведены по 5 рекомендаций (с исключением книг того же автора) для двух произведений разных жанров:

«Герой нашего времени» и сказки «Кот, петух и лиса». Оценки этих и рекомендуемых книг были взяты с помощью парсинга с платформы LiveLib.

Были выбраны пользователи, которые прочитали «Герой нашего времени» и хотя бы 5 других произведений из полного списка рекомендаций. Таких пользователей оказалось 46. Для «Кот, петух и лиса» таких пользователей получилось 54.

Для оценки качества рекомендаций были реализованы две метрики. Hit Rate показывает, купил ли пользователь рекомендованный товар (1 — да, 0 — нет). Precision (Точность) — доля релевантных рекомендаций среди всех предложенных.

Doc2Vec с коэффициентами читабельности показал лучшие значения метрик. Hit Rate равен единице, то есть хотя бы одна подходящая пользователю книга в рекомендациях присутствует. Значение точности 0.58 можно считать хорошим для литературных произведений, где предпочтения пользователей очень индивидуальны и разнообразны. Коэффициенты читабельности улучшили рекомендации для метода Doc2Vec с 0.38 до 0.58. «Мешок слов» показал точность ниже (0.43).

Здесь точность для Doc2Vec с коэффициентами читабельности равна 0.6, Hit Rate равен 0.96, то есть присутствуют пользователи, которым ни одна книг из рекомендуемых не нравится. Но, в целом, результат можно назвать удовлетворительным.

**Шестой раздел «Сравнение результатов всех этапов исследования»** посвящен обобщению результатов разделов 3, 4 и 5 и подведению итогов.

Опираясь на выводы в этом разделе, наиболее подходящим методом для рекомендаций для текстовых данных с учетом специфики области можно назвать Doc2Vec с добавлением коэффициентов читабельности SMOG, ARI и Ганнинга.

# ЗАКЛЮЧЕНИЕ

В ходе выполнения магистерской работы были решены все поставленные задачи. Были изучены существующие книжные рекомендательные сервисы, был сделан обзор источников по данной теме и методов, необходимых для реализации рекомендательной системы. Так, были рассмотрены методы векторного представления текста: «мешок слов», TF-IDF, такие методы тематического моделирования, как латентно-семантическое индексирование, вероятностное латентно-семантическое индексирование, скрытое размещение Дирихле, модели дистрибутивной семантики Word2Vec и Doc2Vec, методы извлечения ключевых слов из текста и методы определения качества текста. Были изучены методы определения качества текста и коэффициенты читабельности: индексы Флэша, Флэша-Кинкейда, SMOG, Ганнинга, Дейла-Чалл, ARI, и исследована применимость этих коэффициентов в системе рекомендаций. Была исследована и применимость методов «мешок слов», TF-IDF, скрытое размещение Дирихле (LDA), Doc2Vec и TextRank: для рекомендаций лучше всего использовать Doc2Vec в сочетании с коэффициентами читабельности SMOG, ARI и Ганнинга. Таким образом, цель магистерской работы достигнута.

В дальнейшем, тему магистерской работы можно развить до создания рекомендательной системы на основе лучших из исследованных методов.

**Отдельные части магистерской работы были опубликованы:**

1. Киреева М.Н. Исследование методов для построения гибридной рекомендательной системы для текстовых данных с учетом специфики области // В кн.: Научное сообщество студентов: МЕЖДИСЦИПЛИНАРНЫЕ ИССЛЕДОВАНИЯ: сборник статей по материалам CCIII международной студенческой научно-практической конференции. Новосибирск: Издательство ООО «СибАК», 2024. С. 36-41.

**Основные источники информации:**


1. Прокофьева И.А., Гордеева О.А., Кудрина М.А. Реализация рекомендательных алгоритмов для книг и литературных произведений // Труды международного симпозиума "Надежность и качество", Т. 1, 2022. С. 211–216.
2. Пуговкина Е.Д., Белоусов А.А. Использование методов кластеризации текстов на естественном языке в рекомендательных системах. Т. 4. // В кн.: Информационные технологии и нанотехнологии (ИТНТ-2022). Самара: Самарский национальный исследовательский университет имени академика С.П. Королева, 2022. С. 041022.
3. Федоренко В.И., Киреев В.С. Машинное обучение в рекомендательных системах // Современные наукоемкие технологии, 2018. С. 102–106.
4. Коршунов А. В., Гомзин А. Г. Тематическое моделирование текстов на естественном языке // Труды Института системного программирования РАН, Т. 23, 2012. С. 215–244.
5. Ильвовский Д.А. Обзор методов оценки сложности текстов в сфере регулирования банковской деятельности. Т. 2. // В кн.: ИТИС 2022. Москва: Институт проблем передачи информации им. А.А. Харкевича РАН, 2022. С. 163–172.
6. Falk K. Practical Recommender Systems. New York: Manning Publications Co. LLC, 2019.
7. Волосова О.В. Технологии искусственного интеллекта в ULS-системах. 2-е изд. Санкт-Петербург: Лань, 2024.
8. Черемушкин В.К., Бакаева О.А. Математические модели графовых методов в рекомендательных системах. Т. 2. // В кн.: Юность и знания - гарантия успеха - 2023. Курск: Закрытое акционерное общество "Университетская книга", 2023. С. 212–215.
9. Lv S., Wang J., Deng F. A hybrid recommendation algorithm based on user nearest neighbor model // Sci Rep, Vol. 14, 2024. P. 17119.