Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ _____ Информатика и системы управления _____

КАФЕДРА _____Системы обработки информации и управления_____

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

_«Обработка и тематический анализ текстовых данных из отзывов сотрудников для выявления ключевых тем и тегов с использованием машинного обучения»_____

Студент ____ИУ5-32М_____ _____ ___**Киричков Е.Е.**___
(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель _____ ___**Гапанюк Ю.Е.**___
(Подпись, дата) (И.О.Фамилия)

2024 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ

Заведующий кафедрой ___ИУ-5___
(Индекс)
_____ __В.И. Терехов__
(И.О.Фамилия)
« _____ » _____ 20 ____ г.

З А Д А Н И Е
на выполнение научно-исследовательской работы

по теме __«Обработка и тематичекий анализ текстовых данных из отзывов сотрудников для выявления ключевых тем и тегов с использованием машинного обучения»_____

Студент группы _____ИУ5-32М_____

_____Киричков Евсей Евгеньвич_____
(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)

_исследовательская_____

Источник тематики (кафедра, предприятие, НИР) _кафедра_____

График выполнения НИР: 25% к _5_ нед., 50% к _9_ нед., 75% к _13_ нед., 100% к 17 нед.

Техническое задание *Провести обработку и тематический анализ текстовых данных из отзывов сотрудников, сделать легковесную систему по обработке новых цитат.*

Оформление научно-исследовательской работы:

Расчетно-пояснительная записка на _20_ листах формата А4.
Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

_____

_____

Дата выдачи задания « _9_ » __сентября__ 20_24_ г.

Руководитель НИР _____ ___**Гапанюк Ю.Е.**___
(Подпись, дата) (И.О.Фамилия)

Студент _____ ___**Киричков Е.Е.**___
(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

Оглавление

Введение

В условиях современного бизнеса отзывы сотрудников представляют собой ценный источник информации, отражающий внутренние процессы компании, уровень удовлетворенности персонала и основные проблемные области. Эффективное использование этих данных позволяет компаниям выявлять направления для улучшения, оптимизировать управленческие решения и повышать общую эффективность работы организации. Однако значительные объемы текстовых данных требуют применения автоматизированных методов для их обработки и анализа.

Тематическое моделирование является одним из ключевых подходов к анализу текстовой информации. Оно позволяет выделять скрытые темы в текстах, выявлять основные направления обсуждений и классифицировать данные для последующего использования. При этом качественная предобработка текстов играет важную роль, обеспечивая снижение шума и повышение точности анализа.

Целью данной работы является разработка подхода к обработке и тематическому анализу текстовых данных из отзывов сотрудников для выявления ключевых тем и тегов с использованием машинного обучения.

Для достижения этой цели решаются следующие задачи:

- Предварительная обработка текстовых данных, включая удаление шума, нормализацию и лемматизацию;
- Настройка методов тематического моделирования для выделения ключевых тем;
- Оценка качества полученных результатов на основе соответствующих метрик;
- Анализ распределения тем и визуализация их структуры;
- Реализация возможности анализа новых текстов.

Практическая значимость исследования заключается в создании системы, способной автоматизировать обработку и анализ текстов. Результаты работы могут быть использованы для улучшения управления персоналом, повышения

вовлеченности сотрудников и формирования рекомендаций по развитию корпоративной культуры.

Настоящее исследование представляет собой комплексный подход к анализу текстовых данных, объединяющий этапы обработки, тематического анализа и визуализации результатов.

1. Аналитический обзор

Тематическое моделирование является ключевым инструментом анализа текстовых данных, позволяя выявлять скрытую структуру тем и ключевые слова в больших корпусах текстов. В условиях современного информационного потока, характеризующегося огромными объемами данных, этот метод становится незаменимым в задачах анализа отзывов, автоматической рубрикации и систематизации информации. Популярными подходами к тематическому моделированию являются вероятностные и алгебраические методы. Среди вероятностных моделей особое место занимает латентное размещение Дирихле (LDA), обеспечивающее мягкую кластеризацию терминов. Этот метод позволяет учитывать многозначность слов и связывать их с несколькими темами в зависимости от контекста. Алгебраические методы, такие как неотрицательное матричное разложение (NMF), также находят применение благодаря высокой интерпретируемости тем, достигаемой за счет использования весов терминов. Однако несмотря на свои преимущества, обе методологии имеют ограничения. Например, LDA не учитывает порядок слов и требует точной настройки гиперпараметров, а NMF больше подходит для текстов со значительным объемом данных.

Современные подходы, такие как битерм-моделирование (BTM), эффективно решают проблемы, связанные с разреженностью данных в коротких текстах. BTM работает на уровне корпуса, анализируя совместную встречаемость пар слов, что делает его подходящим для задач анализа коротких сообщений, таких как твиты или отзывы.

Эффективность тематического моделирования во многом определяется качеством предварительной обработки текстов. Основные этапы включают

удаление стоп-слов, лемматизацию, токенизацию, а также учет специфики текстов, например, выделение биграмм и именованных сущностей. Эти процедуры способствуют уменьшению шума в данных и повышению качества получаемых тем.

Для оценки качества тематических моделей широко используются количественные и качественные методы. Среди количественных подходов популярны метрики когерентности тем, такие как Normalized Pointwise Mutual Information (NPMI), которые измеряют степень согласованности ключевых слов в рамках темы. Качественные методы включают задачи на определение лишнего слова, где человек должен выявить термин, не относящийся к теме, а также вручную назначаемые метки тем, что позволяет проверить интерпретируемость результатов. Эти методы предоставляют важную обратную связь о том, насколько эффективно модель выделяет темы и распределяет их по документам.

Практическое применение тематического моделирования охватывает широкий спектр задач, включая автоматическую рубрикацию новостных текстов, анализ отзывов и выявление трендов. Например, применение тематического моделирования для анализа отзывов сотрудников позволяет не только структурировать данные, но и выявлять ключевые проблемы и предпочтения, что способствует улучшению процессов принятия решений. В таких задачах модели могут быть дополнены процедурами автоматического назначения меток тем, что увеличивает их практическую ценность. Эти возможности делают тематическое моделирование важным инструментом для решения задач, связанных с обработкой больших объемов текстовой информации.

2. Методы и подходы

Для достижения поставленных целей анализа текстовых данных из отзывов сотрудников были использованы современные методы обработки текстов и тематического моделирования. Работа была разделена на несколько

этапов: предобработка текстовых данных, построение тематической модели и оценка её качества.

Предобработка текстов играет ключевую роль в обеспечении качества анализа. Основными шагами на этом этапе являются очистка текстов от шумовых элементов, таких как цифры, знаки препинания и стоп-слова. Также важным процессом является лемматизация, которая позволяет привести слова к их базовым формам и тем самым уменьшить количество уникальных терминов в корпусе. Для этой задачи были использованы инструменты, обеспечивающие точный морфологический анализ, такие как pymorphy2. Дополнительно учитывались сокращения и специфические выражения, характерные для текста отзывов сотрудников, что позволило повысить качество обработки. Также применялась токенизация и выделение биграмм для учета устойчивых словосочетаний, что улучшает результаты моделирования.

На этапе построения тематической модели использовались методы вероятностного анализа. Основной моделью было выбрано латентное размещение Дирихле (LDA), поскольку она обеспечивает интерпретируемость тем и широко используется в анализе текстов. Для построения модели были сформированы корпус и словарь, включающие нормализованные слова и их частотные характеристики. Каждое слово представлялось в формате «мешка слов», что позволило использовать модель для выявления статистических закономерностей в данных. Гиперпараметры модели, такие как количество тем, были выбраны на основе предварительных экспериментов и оценки метрик качества.

Для оценки качества моделей применялись как количественные, так и визуальные подходы. Среди количественных методов использовались метрики когерентности тем, которые измеряют согласованность ключевых слов внутри каждой темы. Эти показатели дают возможность количественно оценить, насколько хорошо слова темы связаны между собой. Дополнительно применялись методы визуализации, такие как интерактивные диаграммы,

которые позволяют наглядно представить распределение тем в корпусе и их взаимосвязи. Визуализация способствовала более глубокому пониманию распределения тем и их структуры.

Все перечисленные методы были интегрированы в единую систему анализа, что позволило автоматизировать процесс обработки отзывов сотрудников и тематического моделирования. Такая структура работы обеспечивает возможность дальнейшего использования разработанных подходов для анализа новых данных и адаптации к изменяющимся задачам.

3. Анализ и обработка данных

В процессе анализа отзывов сотрудников был использован корпус текстов, содержащий как формализованные, так и неформализованные высказывания. Исходные данные включали различные по структуре тексты, от коротких предложений до длинных развернутых описаний. Примеры цитат из отзывов демонстрируют наличие в текстах специфических особенностей: повторы букв, опечатки, неформальная лексика, а также элементы разговорной речи. Например:

1. «У нас среди ночи в районе 55 часов упала полка с водой в стеклянной таре, то есть, посреди этой лужа, стекла, всё, это уже под утро, кстати, это не первый раз...»
2. «Программы повышения квалификации через гильдию ЦМФ, хитрые, хитрые. У всех, кто хочет повышаться, есть свои предпочтения и направления, а гильдия толкает свое, не всегда то, что нужно сотрудникам...».

Эти примеры подчеркивают необходимость тщательной предобработки данных для устранения шумов и подготовки текстов к анализу.

На этапе предобработки была проведена очистка данных от шумов, включая удаление пунктуации, лишних пробелов, цифр и специальных символов. Также были исправлены распространенные опечатки и нормализованы слова с учетом морфологических особенностей русского языка. Для этой задачи использовались инструменты, позволяющие проводить

лемматизацию и токенизацию текста. Особое внимание уделялось сохранению значимых слов и конструкций, таких как аббревиатуры и специальные термины (например, «ЦМФ», «Белые Змеи»).

Сравнение облаков слов до и после обработки демонстрирует результаты предобработки. На рисунке 1 представлено облако слов до очистки, где видны повторяющиеся элементы, опечатки и шумовые данные. После предобработки, как видно на рисунке 2, облако слов стало более структурированным и отражает ключевые термины и темы, содержащиеся в текстах.

Рисунок 1 – Облако слов до обработки

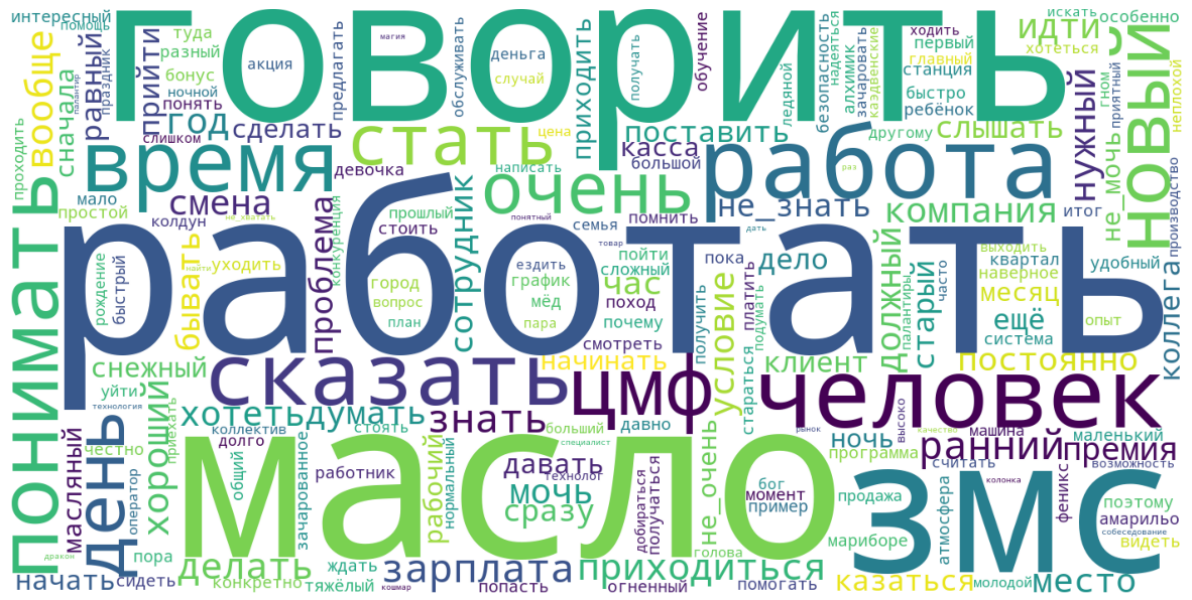

Рисунок 2 – Облако слов после обработки

После завершения всех этапов предобработки был проведен анализ распределения длины текстов в итоговом корпусе. Это позволило убедиться в том, что данные структурированы и готовы для последующего анализа. На рисунке 3 представлена гистограмма распределения длины текстов после обработки. Большинство текстов имеет длину от 10 до 40 слов, что указывает на их компактность и информативность.

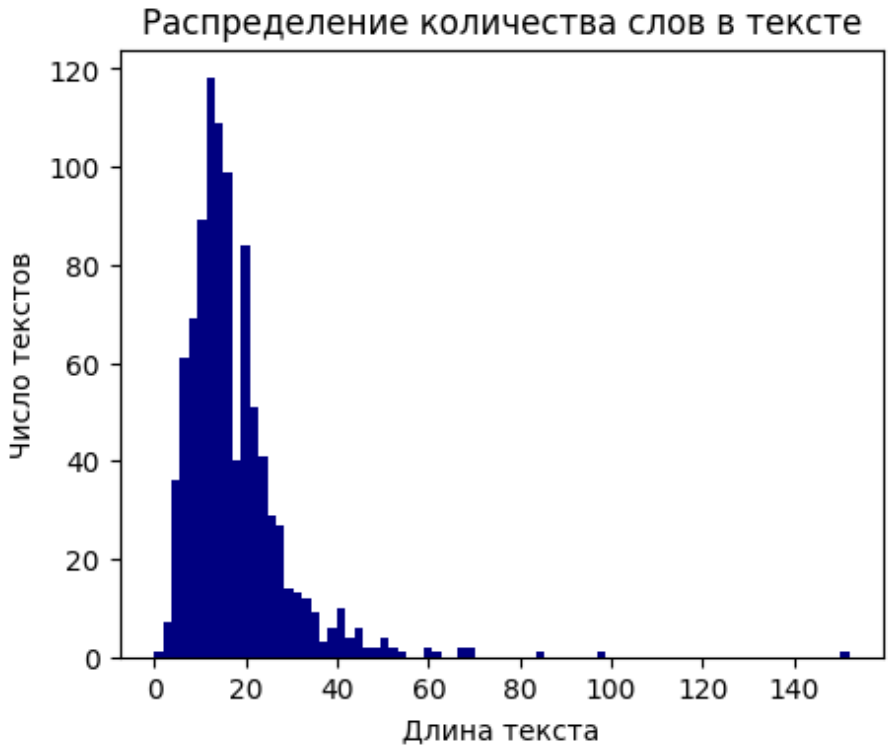

Рисунок 3 - Гистограмма распределения длины текстов после обработки

4. Построение модели

Для тематического моделирования текстовых данных была выбрана модель латентного размещения Дирихле (LDA), реализованная в библиотеке Gensim. Этот метод позволяет выделить скрытые темы в текстах на основе вероятностного распределения слов по темам. Перед обучением модели был сформирован корпус текстов, предварительно обработанный и нормализованный. Для представления данных использовался метод "мешок слов", где каждый текст описывается вектором частот слов.

Обучение модели включало настройку ключевых параметров, таких как количество тем и параметры распределения альфа и бета, влияющих на

разреженность распределений тем в текстах и слов в темах соответственно. Для реализации была использована библиотека Gensim, обеспечивающая удобный интерфейс для построения и тестирования тематических моделей.

Для определения оптимального количества тем использовались метрики когерентности, такие как c_v и u_mass. Эти метрики измеряют согласованность ключевых слов внутри тем и позволяют выбрать такое количество тем, которое обеспечивает наибольшую интерпретируемость результатов.

На рисунке 4 показана зависимость коэффициента когерентности c_v от количества тем, а на рисунке 5 — когерентность u_mass. Для выбора числа тем было использовано пересечение данных двух метрик. Оптимальное значение было выбрано на уровне 8 тем, где достигается баланс между интерпретируемостью тем (по c_v) и стабильностью модели (по u_mass).

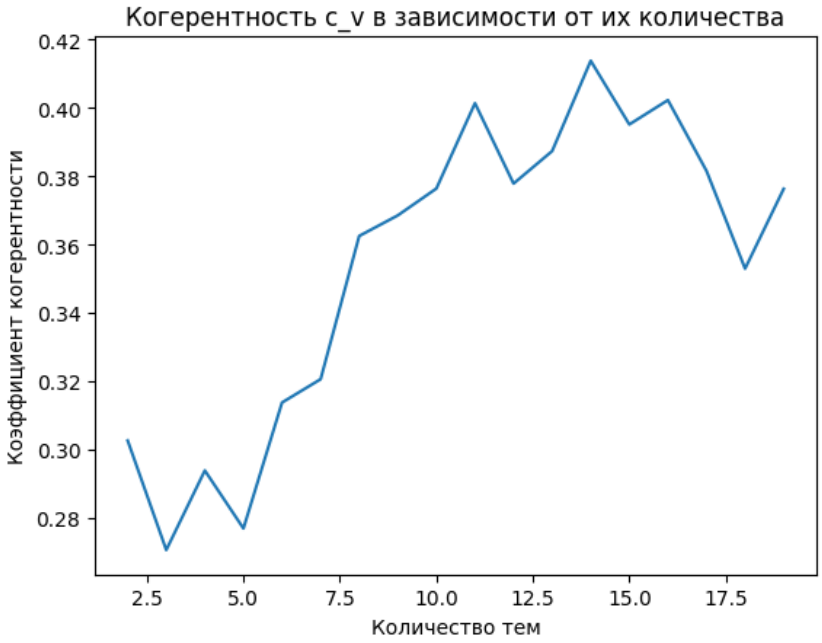

Рисунок 4 - Когерентность c_v

Рисунок 5 - Когерентность u_mass

Итоговая модель LDA была обучена с учетом выбранного количества тем. Визуализация тем, построенная с использованием библиотеки pyLDAvis, представлена на рисунке 6. Для построения карты интертопических расстояний в pyLDAvis использовался алгоритм t-SNE, который сохраняет локальные структуры данных, что позволяет точно отразить взаимосвязи между темами. Карта демонстрирует, насколько темы различаются друг от друга, а также их относительные размеры, основанные на частоте встречаемости тем в текстах.

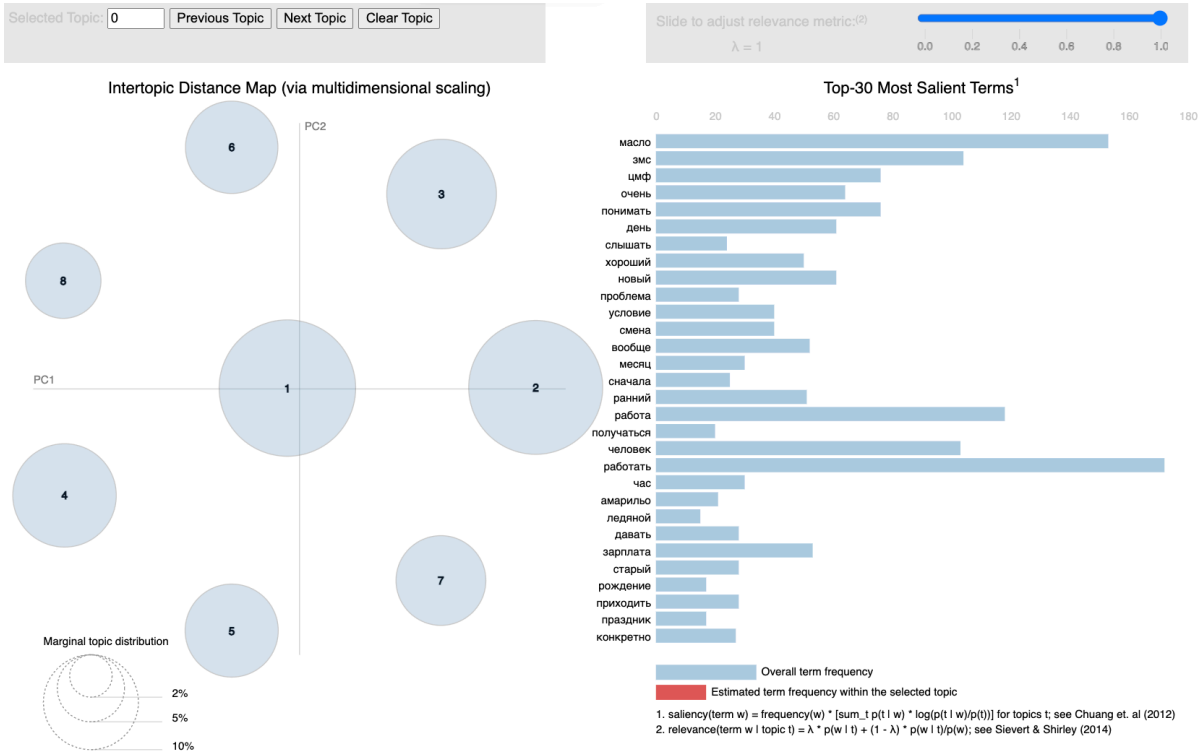

Рисунок 6 – pyLDAvis

5. Обсуждаемые темы и их визуализация

В результате тематического анализа отзывов сотрудников модель выделила 8 ключевых тем, каждая из которых отражает определенные аспекты работы. Для каждой темы были определены характерные ключевые слова, которые помогают интерпретировать её содержание (первые 5 для каждой темы показаны на диаграмме). На основе итоговой модели был выполнен подсчет количества текстов, относящихся к каждой теме, что позволяет понять, какие аспекты работы вызывают наибольший интерес или обсуждение.

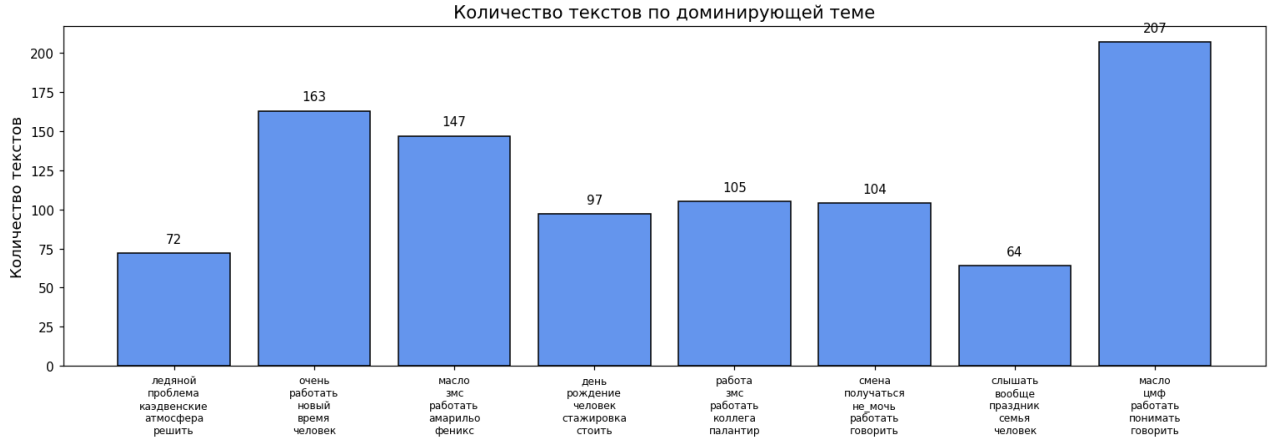

Рисунок 7 – Количество текстов по доминирующим темам

Однако следует отметить, что модель LDA, применённая для анализа, имеет ограниченную эффективность на коротких текстах, так как такие данные могут не содержать достаточного количества информации для точного выделения тем. Это может приводить к менее чётким границам между темами или увеличению шума в результатах. Тем не менее, качественная предобработка текстов помогла минимизировать эти проблемы.

Для визуализации результатов тематического моделирования был применен алгоритм t-SNE, который позволяет сократить размерность данных и отразить кластеризацию текстов в пространстве на основе их тематической принадлежности. Алгоритм t-SNE сохраняет локальные структуры данных, что делает его особенно полезным для анализа взаимосвязей между темами и распределения текстов внутри них.

На рисунке 8 представлена диаграмма кластеризации, где каждая точка соответствует отдельному тексту, а кластеры отражают принадлежность текстов к различным темам. Цветовые маркеры визуализируют распределение текстов по темам, что позволяет наглядно оценить степень их различия и пересечения.

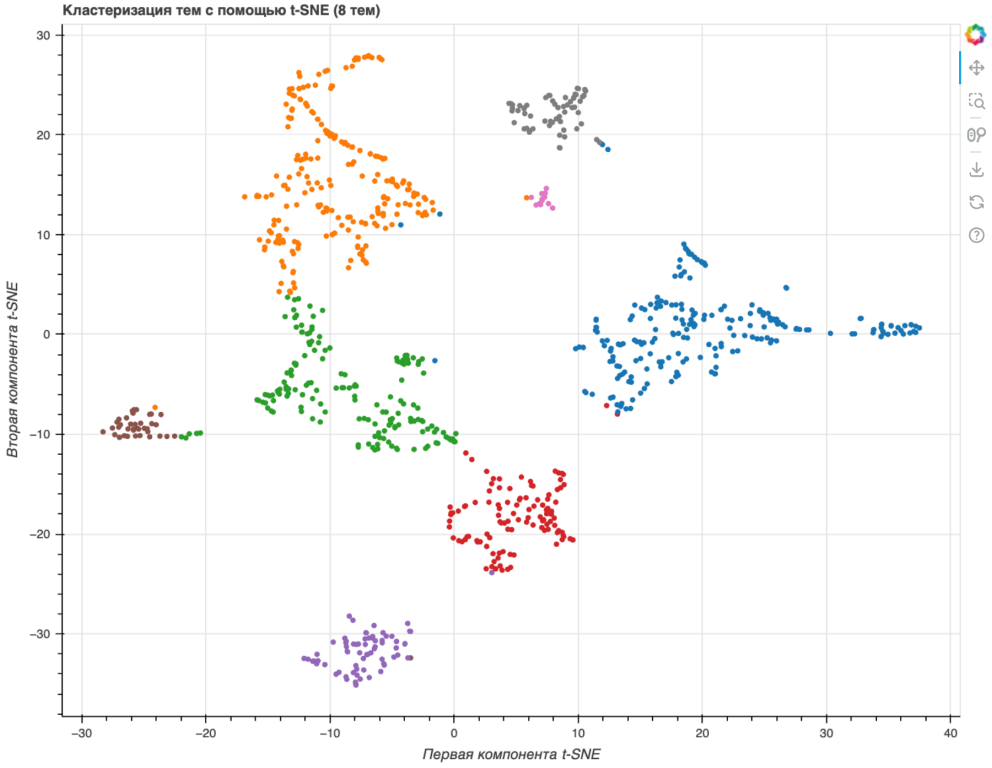

Рисунок 8 – Диаграмма кластеризации с использованием t-SNE

Результаты показывают, что темы имеют чёткие границы, но также присутствуют небольшие области пересечения, что указывает на взаимосвязь отдельных аспектов обсуждаемых вопросов. Например, темы, связанные с карьерным ростом, могут частично пересекаться с темами, посвященными рабочим условиям или логистическим процессам.

6. Легковесная демо-система

Для демонстрации возможностей модели и упрощения её использования была реализована легковесная веб-приложение, построенное с использованием фреймворка Streamlit. Система предоставляет пользователю удобный интерфейс для загрузки данных, анализа цитат и получения результатов тематического моделирования.

Основные возможности:

1. Интерфейс загрузки данных

   Пользователь может загрузить данные в формате JSON вручную через текстовое поле или с использованием файлового ввода. В системе отображается предварительный просмотр загруженных данных, что позволяет убедиться в корректности формата.

2. Анализ цитат

   После загрузки данных система автоматически применяет обученную модель LDA к текстам. Результаты анализа включают:

   - Доминирующую тему для каждой цитаты.
   - Ключевые слова, характеризующие тему.

Результаты анализа отображаются в таблице, где каждая запись включает идентификатор текста, номер доминирующей темы и список ключевых слов, связанных с темой.

3. Экспорт результатов

   Пользователь может скачать результаты анализа в формате CSV для дальнейшего использования или интеграции с другими системами.

Для обеспечения удобства развертывания приложение упаковано в Docker-контейнер. Это позволяет запускать систему на любой платформе,

поддерживающей Docker, без необходимости установки зависимостей или настройки окружения.

Преимущества и перспективы

- Легковесность: Минимальные требования к ресурсам позволяют запускать систему на локальном компьютере или в облаке.
- Гибкость: Поддержка различных форматов данных делает систему универсальной для анализа текстов в разных контекстах.
- Удобство использования: Благодаря простому интерфейсу пользователи, не обладающие техническими навыками, могут быстро проанализировать данные.
- Расширяемость: Система может быть дополнена новыми функциями, такими как визуализация распределения тем или анализ дополнительных текстовых атрибутов.

Легковесная демо-система предоставляет практическую возможность использовать результаты тематического анализа в реальных задачах, таких как анализ обратной связи сотрудников или выявление ключевых проблемных зон в организации. Она демонстрирует модель LDA в удобной и доступной форме.

Анализ интервью Цинтрийские Масла Феникса

Выберите способ загрузки данных:

Способ загрузки

Ввести JSON текст вручную

Введите данные в формате JSON


```
[
  {
    "id": 0,
    "quote": "«У ннас среди ночі в райооне 55 часов упала полха с водой в сттеклянной  таре, тоесть, посреди этой лужа, сстеккла, всёё, это уже поод утро, кстати, ээто не перрвыи раз,,, они сами по себе падают, как-то неправилььно рассчитывают, мы же должны выставлять по определённой картинке, у нас при мне было уже пару раз, чтоо сами по себе грохаются эти полки с буттылками»\n\n\n"
  },
  {
    "id": 1,
    "quote": "Программы повышения квалификации через гильдию ЦМФ, хитрые, хитрые. У всех, кто хочет повышаться, есть свои предпочтения и направления, а гильдия... гильдия
```


Загруженные данные:

	id	quote
0	0	«У ннас среди ночі в райооне 55 часов упала полха с водой в сттеклянной таре,
1	1	Программы повышения квалификации через гильдию ЦМФ, хитрые, хитрые. У в

Анализ цитат...

Результаты анализа:

	id	topic_num	keywords
0	0	5	смена, получаться, не_мочь, работать, говорить, сначала, час, понять, должный
1	1	1	очень, работать, новый, время, человек, сказать, говорить, идти, месяц, старый

Скачать результаты в CSV

Рисунок 9 - Демо-система

Заключение

В ходе работы была выполнена комплексная обработка текстовых данных для тематического моделирования, направленная на выявление основных тем в корпусе текстов. Основная цель состояла в том, чтобы понять структуру данных, выявить скрытые темы и оценить их распространённость и значимость в тексте. В качестве метода была выбрана LDA (Latent Dirichlet Allocation) — популярная техника тематического моделирования, которая представляет документы как комбинации скрытых тем и позволяет интерпретировать содержимое текста более структурировано.

1. Предобработка данных. На первом этапе были проведены различные операции по очистке текста, включая удаление лишних символов, эмодзи, чисел и пунктуации, а также исправление орфографии и замена сокращений. Кроме того, текст был лемматизирован, и стоп-слова были удалены. Были учтены нюансы, такие как обработка отрицаний и специфических терминов, что позволило сохранить смысловые связи в текстах. В результате предобработки тексты стали более пригодными для анализа, а также повысилось качество получаемых тем.
2. Создание и оценка LDA-модели. С помощью библиотеки gensim была обучена модель LDA с оптимальным количеством тем, подобранным на основе метрик качества. Мы использовали метрику: Когерентность (метрики u_mass и c_v) — для оценки интерпретируемости тем. Когерентность позволила нам выбрать такое количество тем, при котором темы остаются интерпретируемыми, а модель не теряет общей связности
3. Визуализация тем. Для лучшего понимания структуры тем была выполнена визуализация с использованием t-SNE и библиотеки PyLDAvis. t-SNE позволила отобразить темы в виде кластеров в двумерном пространстве, что облегчило анализ связей между темами и выявление их группировок. PyLDAvis, в свою очередь, визуализировала распределение слов в темах, что позволило оценить важность ключевых слов и их роль в описании тем.

Эти визуализации дали наглядное представление о структуре данных и помогли выявить наиболее значимые темы.

4. Анализ распределения тем. Была выполнена количественная оценка распределения текстов по темам. В результате анализа было выяснено, какие темы являются доминирующими и наиболее распространёнными в корпусе, а какие встречаются реже.

Дальнейшее развитие работы может включать более глубокий анализ интерпретируемости тем, улучшение методов предобработки для специфических корпусов, а также применение нейронных сетей для более точного тематического моделирования коротких текстов (BERT).

Полученные результаты использованы при написании приложения на Streamlit для выделения ключевых тем и тегов.

Список литературы

1. Билбро Р., Бенгфорт Б. Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка. – 2019.
2. Воронцов К.В. Вероятностное тематическое моделирование. URL: www.machinelearning.ru. – 2013.
3. Тематическое моделирование // Википедия, свободная энциклопедия. URL: http://en.wikipedia.org/wiki/Тематическое_моделирование.
4. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet Allocation // Journal of Machine Learning Research. – 2003. – Т. 3. – Вып. 4–5. – С. 993–1022. DOI: 10.1162/jmlr.2003.3.4-5.993.
5. Nature. Validation of Topic Models and Applications // Scientific Reports. – 2024. DOI: https://www.nature.com/articles/s41598-024-61738-4.