МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«ТЮМЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

ШКОЛА КОМПЬЮТЕРНЫХ НАУК
Кафедра программного обеспечения

РЕКОМЕНДОВАННО К ЗАЩИТЕ
В ГЭК

Заведующий кафедрой
Кандидат наук,
М.С. Воробьева
_____ 2024 г.

ОТЧЕТ ПО ДИСЦИПЛИНЕ: ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ ПОДДЕРЖКИ И ПРИНЯТИЯ РЕШЕНИЙ

НА ТЕМУ: РАЗРАБОТКА СЕРВИСА ДЛЯ АНАЛИЗА ТЕКСТОВ НАУЧНЫХ ПУБЛИКАЦИЙ СБОРНИКА «МАТЕМАТИЧЕСКОЕ И ИНФОРМАЦИОННОЕ МОДЕЛИРОВАНИЕ»

02.03.03 Математическое обеспечение и администрирование информационных систем

Выполнили работу (групповой проект) студенты 4 курса очной формы обучения	Загайнова Евгения Олеговна Калимова Алтынай Есенбаевна
Руководитель Профессор, д. пед. н.	Захарова Ирина Гелиевна

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

В условиях стремительного развития цифровых технологий научная деятельность всё чаще связана с необходимостью работы с электронными архивами публикаций. Современные конференции и научные журналы, как правило, представляют свои материалы в формате PDF-сборников, в которых статьи включены в состав общего документа с единым оглавлением и структурированными метаданными. Примером такого формата является сборник конференции «Математическое и информационное моделирование», объединяющий сотни научных работ, приведённых к единому стилевому и содержательному стандарту. Несмотря на доступность таких архивов, их использование сопряжено с рядом трудностей: поиск нужной информации зачастую осуществляется вручную, отсутствуют инструменты семантического анализа, а навигация по содержанию ограничена. Это создаёт значительные временные затраты и снижает эффективность работы исследователей и авторов научных статей. Особенно актуальна эта проблема при попытке анализа структуры и содержания сборника: выявления наиболее часто публикуемых тем, характерных формулировок аннотаций, используемых терминов и типов иллюстративного материала.

Переход научных изданий в электронный формат открывает новые возможности для автоматизации таких процессов. Однако большинство PDF-сборников, размещённых в архивах (например, eLibrary Тюменского государственного университета), изначально не предполагают встроенных функций поиска, фильтрации или анализа. Это обуславливает необходимость создания специализированных программных средств, способных не только извлекать ключевые компоненты публикаций, но и предоставлять пользователю интерфейс для работы с полученными данными.

Целью настоящей работы является разработка программного сервиса, обеспечивающего автоматизированную обработку PDF-сборников научных публикаций, структурирование содержимого и реализацию пользовательского веб-интерфейса с поддержкой семантического поиска и интеллектуальной фильтрации.

Для достижения цели были поставлены и реализованы следующие задачи:

1. Изучить методы обработки и анализа научных статей;
2. Спроектировать систему хранения данных;
3. Разработать модули для обработки и анализа научных текстов:
    3.1. Разработать модуль предобработки и извлечения метаданных;
    3.2. Разработать модуль фильтрации;
    3.3. Разработать модуль семантического поиска по аннотациям;
    3.4. Разработать модуль визуализации результатов;
4. Реализовать сервис для анализа научных публикаций.

В процессе реализации проекта особое внимание уделялось не только техническим аспектам, но и созданию условий, способствующих сохранению работоспособности и общего физического состояния. В период разработки были предусмотрены регулярные физические упражнения и активные перерывы, направленные на профилактику утомления и поддержание высокой концентрации внимания. Такой подход позволил избежать переутомления и обеспечить стабильную продуктивность при выполнении трудоёмких задач анализа и программирования.

Также в ходе работы соблюдались требования безопасной организации рабочего процесса. Все этапы выполнения проекта сопровождались контролем за уровнем нагрузок и мерами по поддержанию комфортной и безопасной среды, что особенно важно в условиях длительной работы за компьютером и высокой интеллектуальной активности.

МЕТОДЫ И ТЕХНОЛОГИИ АНАЛИЗА НАУЧНЫХ ТЕКСТОВ

1.1 Структура и особенности научных публикаций

Научные статьи являются основным способом передачи результатов исследований и широко применяются в академическом сообществе. Независимо от предметной области, подавляющее большинство публикаций имеют стандартизированную структуру, включающую следующие ключевые элементы:

- **Заголовок** — краткое отражение содержания исследования.
- **Аннотация** — сжатое описание целей, методов и результатов.
- **Ключевые слова** — термины, отражающие тематику и предметную область.
- **Основной текст**, как правило, включает: введение, материалы и методы, результаты, обсуждение, заключение.
- **Список литературы** — перечень использованных источников и научных заимствований.
- **Сведения об авторах** и, при наличии, их организациях.

В контексте PDF-сборников, таких как издание «Математическое и информационное моделирование», статьи представлены в едином документе, где структура каждой работы визуально повторяется, но не всегда строго формализована с точки зрения машинной обработки. Это создаёт дополнительные вызовы при автоматическом извлечении информации. Однако такая однородность позволяет применять устойчивые шаблоны для выделения элементов — через регулярные выражения.

Метаданные (заголовки, авторы, аннотации, ключевые слова) имеют особую ценность в автоматическом анализе, поскольку занимают малый объём и несут высокую информативную нагрузку. Их извлечение и анализ позволяют реализовать фильтрацию, индексирование, тематическое моделирование и поиск по смыслу без необходимости обращения ко всему полному тексту.

1.2 Методы обработки естественного языка

Для автоматизации анализа научных текстов используются методы обработки естественного языка (Natural Language Processing, NLP), объединяющие широкий спектр подходов, направленных на извлечение, структурирование и интерпретацию текстовой информации. Эти методы позволяют не просто считывать текст, а «понимать» его содержательную структуру, выявлять ключевые элементы, сопоставлять по смыслу и строить на этой основе интеллектуальные системы анализа.

Лексический анализ — это первый и один из самых важных этапов обработки текста. Его задача — разложить текст на отдельные элементы (лексемы), пригодные для дальнейшего анализа. Он позволяет превратить сплошной поток символов в набор осмысленных единиц, с которыми можно работать количественно и логически.

Методы предобработки текста:

Токенизация — разбиение текста на отдельные элементы (токены), такие как слова, числа, знаки препинания. Это позволяет преобразовать сплошной поток символов в логически осмысленные единицы.

Удаление стоп-слов — исключение из текста слов, не несущих смысловой нагрузки, но часто встречающихся

Нормализация текста — приведение текста к единому виду. Сюда входит:

- перевод всех символов в нижний регистр;
- удаление лишних пробелов, пунктуации, спецсимволов;
- устранение шумов, таких как HTML-теги, даты, числа.

Лемматизация — процесс приведения слова к его словарной (начальной) форме — лемме. Это важно для снижения морфологического разнообразия языка, особенно русского.

Стемминг — метод грубого усечения слова до основы (стемы), не всегда совпадающей с леммой. Используется для ускоренной фильтрации и индексации текста.

Синтаксический анализ представляет собой этап обработки текста, направленный на выявление грамматических связей между словами и

построение формального представления структуры предложений. В рамках синтаксического анализа определяются роли слов в предложении, их иерархия, типы подчинения и взаимосвязи. Это позволяет понять, как слова взаимодействуют друг с другом, какие являются главными, а какие зависимыми, что критически важно для точной интерпретации научного текста.

Среди ключевых методов синтаксического анализа выделяются:

- Контекстно-свободные грамматики (CFG), основанные на формальных правилах, описывающих допустимые конструкции языка. Эти грамматики используются для построения синтаксических деревьев, описывающих структуру предложений.
- Зависимостный парсинг (Dependency Parsing), который фокусируется на установлении пар зависимостей между словами. Каждое слово рассматривается в отношении зависимости от другого, и формируется дерево, в котором отображаются направленные связи между лексемами.
- Конституентный анализ (Constituency Parsing), направленный на определение фразовых структур и вложенности в предложениях. С его помощью можно выделить синтаксические единицы, такие как именные или глагольные группы.

Синтаксический анализ является фундаментом для последующего семантического анализа, так как он обеспечивает корректную разметку структуры предложения, определяя, какие части текста связаны между собой логически и грамматически.

Семантический анализ представляет собой этап обработки текста, целью которого является интерпретация смысла слов, фраз и предложений. Он выходит за рамки грамматической структуры и фокусируется на содержательной стороне текста. Семантический анализ применяется для оценки смысловой близости между различными текстами, идентификации ключевых понятий и установления смысловых связей.

Ключевыми направлениями семантического анализа являются:

- Векторизация текста, заключающаяся в преобразовании слов, предложений или целых документов в числовые векторы. Эти векторы отражают семантические характеристики текста и позволяют

использовать математические методы для сравнения и кластеризации.

- Семантическое сопоставление, обеспечивающее определение степени смысловой близости между текстами. Оно используется, например, для реализации интеллектуального поиска, при котором запрос пользователя сравнивается с текстами научных статей по смыслу, а не по формальным совпадениям слов.
- Тематическая классификация, направленная на определение принадлежности текста к определённой тематической категории на основе анализа его содержательной части.

1.3. Семантический анализ текстов

Семантический анализ является одним из ключевых направлений обработки текстовой информации, целью которого является выявление, интерпретация и использование смысловых связей между словами, фразами и фрагментами текста. В отличие от лексического и синтаксического анализа, ориентированных на форму и структуру, семантический анализ сосредоточен на содержании — то есть на том, "что именно" говорится в тексте, а не только "как" это сказано.

Современные подходы к семантическому анализу сочетают методы лингвистического моделирования, статистического анализа и машинного обучения. Основной задачей является создание представлений текста, позволяющих сравнивать, классифицировать и искать документы на основе их смысла. В рамках данной работы реализованы несколько таких подходов, каждый из которых детально описан ниже.

1.3.1. Векторизация текстов

Процесс векторизации текста заключается в его преобразовании из символьной формы в числовое представление, пригодное для машинной обработки. Векторизация позволяет перейти от текстовой информации к структурированному пространству признаков, отражающих семантику текста.

Методы векторизации:

- **TF-IDF (Term Frequency – Inverse Document Frequency)** — один из наиболее простых и интерпретируемых подходов. TF отражает частоту термина в документе, а IDF — его уникальность в корпусе. Метод не учитывает порядок слов и контекст, но хорошо подходит для задач базового тематического моделирования и фильтрации. Реализация возможна через библиотеки scikit-learn, где объект TfidfVectorizer позволяет построить матрицу признаков.
- **Word2Vec** — метод обучения эмбеддингов слов на основе их контекста. Существует две архитектуры: CBOW (предсказывает текущее слово по окружению) и Skip-Gram (предсказывает контекст по слову). Модель обучается на корпусе текстов, создавая плотные вектора для каждого слова. Реализация доступна в библиотеке gensim.
- **GloVe (Global Vectors)** — в отличие от Word2Vec, строится на основе глобальной статистики частот совместной встречаемости слов. Итоговые вектора отражают как локальные, так и глобальные зависимости между словами.
- **FastText** — расширение Word2Vec, в котором слова представляются как совокупность n-грамм (например, «машина» → «ма», «аш», «ши», «ин», «на»). Это позволяет обрабатывать редкие и незнакомые слова. Поддерживается в gensim и оригинальной реализации Facebook.
- **BERT** — контекстная модель трансформерного типа, обученная на задаче маскированного языка. Она обрабатывает предложение целиком, понимая смысл слов в зависимости от окружающего контекста. Используется для получения эмбеддингов слов и предложений. Реализуется с помощью transformers от HuggingFace.
- **Sentence-BERT / SentenceTransformer** — адаптация BERT для получения эмбеддингов на уровне предложений и абзацев. Включает предобученные модели, которые можно использовать без дообучения, или дообучить на собственных парах предложений с помощью триплетных потерь. Библиотека sentence-transformers предоставляет удобный интерфейс для применения векторизации и оценки сходства.

1.3.2. Семантическое сопоставление текстов

После того как текст представлен в векторной форме, можно вычислять его семантическую близость к другим текстам. Это позволяет находить статьи, схожие по смыслу, даже если формулировки различаются.

Наиболее распространённые метрики:

- **Косинусное сходство** — рассчитывается как косинус угла между двумя векторами. Используется для оценки семантической близости и реализуется через sklearn.metrics.pairwise.cosine_similarity или scipy.spatial.distance.cosine.
- **Евклидово расстояние** — измеряет абсолютное расстояние между точками в векторном пространстве. Подходит при равномерном распределении данных, но чувствителен к масштабам.
- **Манхэттенское расстояние** — сумма абсолютных разностей по всем координатам, применяется реже, но может быть устойчивым при наличии выбросов.

Реализация семантического сопоставления в проекте осуществляется путём сравнения векторов аннотаций и пользовательского ввода. После получения векторов из модели SentenceTransformer, для каждой статьи рассчитывается косинусное сходство с запросом, и результаты сортируются по убыванию значения. Наиболее близкие статьи подаются на выход в интерфейс.

1.3.3. Тематическое моделирование

Этот метод направлен на автоматическое выявление тем в большом массиве текстов. В основе лежит предположение, что каждый документ содержит несколько латентных тем, каждая из которых может быть описана распределением слов.

Реализация тематического моделирования:

- **LDA (Latent Dirichlet Allocation)** — одна из самых популярных моделей, реализованная в gensim.models.LdaModel. Требует входной матрицы "слово-документ" и формирует вероятностные распределения тем. При этом каждая статья может относиться сразу к нескольким темам с разной степенью вероятности.

- **NMF (Non-negative Matrix Factorization)** — работает через разложение матрицы частот слов на матрицы с тематическими и документными коэффициентами. Простой в интерпретации и применении через sklearn.decomposition.NMF.
- **BERTopic** — современный подход, объединяющий векторизацию на основе BERT, кластеризацию (например, с использованием HDBSCAN) и выделение ключевых слов на уровне темы. Позволяет получить интерпретируемые темы даже в малых выборках. Библиотека bertopic предлагает готовый инструментарий для анализа и визуализации.

В проекте тематическое моделирование используется как дополнительный модуль, позволяющий анализировать общее распределение тем в PDF-сборниках и рекомендовать пользователю релевантные области знаний.

1.3.4. Классификация и категоризация текстов

Автоматическая классификация текста заключается в отнесении его к одной или нескольким предопределённым категориям. В научной среде это может быть классификация по предметным областям, типу исследования, уровню оригинальности и т.д.

Виды классификации:

- **Бинарная** — определение принадлежности к категории (например, генеративный / негенеративный текст);
- **Многоклассовая** — присвоение одного класса из множества (например, «математика», «медицина», «информатика»);
- **Многоуровневая (иерархическая)** — классификация по дереву тем или УДК.

Методы реализации:

- **Логистическая регрессия, SVM** — хорошо работают на TF-IDF векторах, требуют обучения на размеченной выборке.
- **Деревья решений, случайный лес** — подходят при наличии табличных признаков.

- **Нейросетевые модели** — CNN, LSTM и особенно трансформеры, такие как RoBERTa и DistilBERT, дают высокую точность при наличии большого корпуса.

В проекте возможна реализация классификатора по тематикам на основе эмбеддингов аннотаций, с последующей дообучаемой логистической моделью или KNN-классификатором по косинусной мере.

Таким образом, семантический анализ охватывает весь спектр задач — от построения смысловых представлений текста до поиска и категоризации. Его реализация в проекте построена на использовании контекстных моделей (SentenceTransformer), обеспечивающих гибкость, адаптируемость и высокую точность обработки научных публикаций.