Целью курсовой работы является реализовать и сравнить алгоритмы выделения ключевых слов из текстовых данных с целью извлечения важной информации для анализа текстов. В данном отчете рассматриваются современные методы и алгоритмы для выделения ключевых слов из текста.

В ходе работы были проанализированы ключевые литературные источники по обработке естественного языка, включая книги и научные статьи из журналов и с конференций, были использованы статьи из авторитетных журналов и интернет-ресурсов.

Рассмотрены ключевые этапы предобработки текста: токенизация, удаление стоп-слов, определение частей речи (POS-теггинг), лемматизация. Для реализации этих этапов были изучены инструменты и библиотеки Python, такие как NLTK и SpaCy.

Метод TF-IDF (term frequency-inverse document frequency) основан на вычислении значимости слова путем оценки его частоты в документе относительно всего корпуса текстов. Также был рассмотрен графовый алгоритм TextRank. Он может использоваться как для суммаризации текста, выделяя наиболее значимые предложения, так и для извлечения ключевых слов на основе связей между словами, представленных в виде графа.

Особое внимание уделено модели BERT (Bidirectional Encoder Representations from Transformers), которая использует трансформеры для контекстуальной обработки слов. BERT анализирует текст в обоих направлениях и позволяет точно выделять ключевые слова за счет понимания контекста. Дополнительно рассмотрен метод «мешка слов» (Bag of Words), который представляет текст в виде набора слов, учитывая только их частоту и игнорируя порядок. Также изучено использование N-грамм — последовательностей из N слов, которые применяются для анализа устойчивых фраз и извлечения ключевых слов. Преимущества нейронных сетей перед традиционными алгоритмами заключаются в способности учитывать сложные зависимости между словами и контекстуальную информацию.

Проведено сравнение методов на основе критериев точности, полноты, времени обработки, сложности и учёт контекста. TF-IDF и BoW — быстрые и простые алгоритмы, но не учитывают контекст и уступают по качеству. TextRank — эффективен для ключевых слов и суммаризации, но не всегда точен. BERT — лучший метод по качеству (точность, полнота, учёт контекста), но требует больших ресурсов. Нейронные сети — вариативные по сложности и качеству, применимы для широкого круга задач.

В результате работы был проведен обзор современных методов выделения ключевых слов, включая TF-IDF, TextRank, BERT и нейронные сети. Рассмотрены этапы предобработки текста, такие как токенизация, лемматизация и удаление стоп-слов, а также инструменты для их реализации. Сравнение алгоритмов позволило выявить их особенности, сильные и слабые стороны, а также области применения.

Цель курсовой работы: Реализовать и сравнить алгоритмы выделения ключевых слов из текстовых данных с целью извлечения важной информации для анализа текстов.

Проведение обзора и анализа источников по теме выделения ключевых слов, изучение современных методов и алгоритмов, таких как TF-IDF, TextRank, BERT, нейронные сети и N-граммы, а также подходов на основе графов и трансформеров. Также определение критериев для оценки качества алгоритмов с учетом таких параметров, как точность, полнота, время обработки, сложность и учет контекста. Следующим этапом станет реализация выбранных алгоритмов на практике с использованием инструментов Python, таких как NLTK и SpaCy, и их тестирование на текстовых данных. Планируется проведение экспериментов, сравнение результатов, а также анализ полученных данных с целью выявления преимуществ и недостатков каждого метода. Завершением станет подведение итогов работы.

На данный момент выполнены первые две задачи: проведен обзор и анализ источников по теме выделения ключевых слов, а также определены критерии для оценки качества алгоритмов, и выявлены более эффективные.

Тема: Выделение ключевых слов из текстовых данных