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

Кафедра информатики и программирования

ИЗВЛЕЧЕНИЕ КЛЮЧЕВЫХ СЛОВ ИЗ РУССКИХ ТЕКСТОВ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Зиборовой Елизаветы 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 2023

ВВЕДЕНИЕ

Актуальность темы. Большое количество данных, генерируемых каждый день, представляет собой как преимущество, так и недостаток. С одной стороны, данные помогают компаниям получить реальную информацию о мнениях людей о продукте или услуге. Такие сведения можно добыть из анализа электронных писем, обзоров продуктов, сообщений в социальных сетях, отзывов клиентов, заявок в службу поддержки и т. д. С другой стороны, есть проблема, как обрабатывать все эти данные, ведь вручную такая работа будет занимать довольно длительное время. И здесь извлечение ключевых слов из текста играет важную роль.

Ключевые слова помогают и обычным пользователям в таких вопросах, как выбор товара с помощью краткой выдержки из всех отзывов; запрос фактической информации из новостных статей и интернет-ресурсов; выбор фильма с использованием обширного словаря ключевых жанров и др.

Извлечение ключевых слов (Keyword Extraction) — это метод анализа текста, который автоматически извлекает из текста часто используемые и наиболее значимые слова и выражения. Этот метод помогает обобщить содержание текстов и выделить основные темы обсуждения.

Для обработки естественного языка, в частности, извлечения ключевых слов используются методы машинного обучения.

Цель бакалаврской работы – изучение методов извлечения ключевых слов, их реализация и анализ полученных результатов.

Поставленная цель определила **следующие задачи**:

1. Изучить основные понятия и определения процесса извлечения ключевых слов.
2. Рассмотреть методы извлечения ключевых слов.
3. Выполнить реализацию методов извлечения ключевых слов.
4. Проверить работу реализованных методов на размеченных данных.
5. Ознакомиться с методами оценки качества.
6. Провести сравнительный анализ реализованных методов и оценить их.

Методологические основы извлечения ключевых слов представлены в работах: Дубинина Е.Ю. [1], Тихонова Е.В., Косычева М.А. [2], Ноздрина Т.Г. [3], Ванюшкин А.С., Гращенко Л.А. [4], Jones K.S. [5], Mihalcea R., Tarau P. [6], Rose S., Engel D., Cramer N., Cowley W [7].

Теоретическая значимость бакалаврской работы.

Полученные результаты исследования подтверждают проблематику стандартных методов извлечения ключевых слов.

Структура и объём работы. Бакалаврская работа состоит из введения, 4 разделов, заключения, списка использованных источников и 3 приложений. Общий объем работы – 50 страницы, из них 46 страниц – основное содержание, включая 16 рисунков и 16 таблиц, цифровой носитель в качестве приложения, список использованных источников информации – 25 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Извлечение ключевых слов» посвящен рассмотрению основных понятий извлечения ключевых слов, этапов извлечения и описанию методов TF-IDF, TextRank, RAKE.

Выделение ключевых слов из неструктурированных данных является одним из методов обработки такой информации. Этот метод позволяет выделить наиболее важные и значимые слова, которые могут быть использованы для анализа и сравнения данных. Извлечение ключевых слов помогает предприятиям принимать более обоснованные решения, оптимизировать свою деятельность и повышать эффективность бизнес-процессов. Ключевые слова помогают и обычным пользователям в таких вопросах, как выбор товара с помощью краткой выдержки из всех отзывов; запрос фактической информации из новостных статей и интернет-ресурсов; выбор фильма с использованием обширного словаря ключевых жанров и др.

Ключевые слова характеризуются тем, что:

- являются наиболее употребительными (частотными) наименованиями, обозначают признак предмета, состояние или действие;
- представлены значимой лексикой, достаточно обобщены по своей семантике (средней степени абстракции), стилистически нейтральны, не оценочны;
- связаны друг с другом сетью семантических связей, пересечения значений;
- более половины слов ядра тематического компонента состоит из ключевых слов, а минимальный набор ключевых слов приближается к инварианту содержания при их логическом упорядочении;
- набор ключевых слов состоит из 5-15 или 8-10 слов, что соответствует объему оперативной памяти человека, в тексте содержится 25-30% ключевых слов;

- набор ключевых слов определяет контексты слов, обладающих максимальной предсказуемостью.

Несмотря на обширный объем специализированных и междисциплинарных исследований, посвященных ключевым словам, до настоящего времени не существует единой методики обнаружения ключевых слов человеком. Экспериментально подтверждено, что данная операция осуществляется интуитивно людьми и зависит от личностных и гендерных особенностей. Это приводит к сложностям при разработке методов и алгоритмов извлечения ключевых слов для вычислительной техники. Отсутствие четких формализованных моделей и размытые определения в области компьютерной лингвистики и других инженерных дисциплин затрудняют создание и верификацию соответствующего инструментария.

Извлечение ключевых слов из текста представляет собой сложный процесс, так как их характеристики проявляются на нескольких уровнях текстового анализа, включая морфологический, лексический, синтаксический и прагматический. Поэтому для их распознавания требуются методы, которые могут быть многоэтапными и сложными. Исследование литературы показывает, что существуют три последовательных этапа в современных алгоритмах извлечения ключевых слов: предобработка, распознавание (классификация) и постобработка.

Методы извлечения ключевых слов делятся по следующим признакам:

- наличие элементов обучения и подходы к его реализации;
- тип математического аппарата системы распознавания, обусловленного формой информации представления признаков ключевых слов;
- тип используемых для реализации метода лингвистических ресурсов.

Рассмотрим подробнее три метода различных по математическому аппарату распознавания, а именно статистический TF-IDF, структурный TextRank и гибридный RAKE.

TF-IDF - статистическая мера, которая применяется для оценки значимости определенного слова в контексте документа, который является частью коллекции документов или корпуса. Вес слова пропорционален частоте употребления этого слова в документе, и обратно пропорционален частоте употребления слова в других документах коллекции.

TextRank представляет собой графическую модель ранжирования, используемую для обработки текста. Она основана на концепции представления текста в виде графа, где слова представлены вершинами, а связи между ними - ребрами графа. Для оценки важности каждого слова используется классическая метрика PageRank.

RAKE (Rapid Automatic Keyword Extraction) — еще один алгоритм извлечения ключевых слов на основе графа. Алгоритм основан на наблюдении, что ключевые слова часто состоят из нескольких слов и обычно не включают стоп-слова или знаки препинания

Второй раздел «Методы оценки» посвящен описанию оценок качества, с помощью которых далее на практике будут оцениваться реализованные методы.

К показателям качества работы системы относят полноту (Recall), точность (Precision) и F-меру (F).

Полнота - это доля истинно положительных классификаций. Полнота показывает, какую долю объектов, реально относящихся к положительному классу, мы предсказали верно. Полнота демонстрирует способность алгоритма обнаруживать данный класс вообще.

Точность - доля правильных ответов модели в пределах класса — это доля объектов действительно принадлежащих данному классу относительно всех объектов которые система отнесла к этому классу. Именно введение precision не позволяет нам записывать все объекты в один класс.

F-мера - “гармоническое среднее” точности и полноты.

Третий раздел «Программные инструменты Python для реализации» посвящен обоснованию выбора языка Python для реализации и описанию необходимых библиотек.

Обработка естественного языка (NLP - Natural Language Processing) — это мощная технология, которая помогает извлечь огромную пользу из текстовых данных.

В связи с тем, что методы анализа текста широко применяются в области машинного обучения, необходимо использовать язык программирования, который обладает обширным набором научных и вычислительных библиотек. В качестве наиболее подходящего для этой цели языка может выступать Python, который включает в себя значительный набор мощных библиотек, таких как Scikit-Learn, NLTK, Gensim, spaCy, NetworkX и Yellowbrick.

Четвертый раздел «Практическая часть» посвящен реализации рассмотренных методов на размеченном наборе данных.

В качестве датасета были выбраны habrahabr_0.jsonlines, habrahabr_1.jsonlines, habrahabr_2.jsonlines, habrahabr_3.jsonlines. Каждый файл представляет собой набор из 1000 различных статей с сайта habr.com.

По каждой статье в датасете выделен текст статьи, заголовок, содержание, url-адрес и ключевые слова вручную автором.

Для извлечения ключевых слов были выполнены следующие этапы: предварительная обработка данных (удаление стоп-слов, токенизация, лемматизация), распознавание ключевых слов с помощью рассмотренных методов TF-IDF, TextRank, RAKE и постобработка в виде вывода списка ключевых слов.

В итоге получилось, что метод RAKE сработал лучше других. Это может быть объяснено тем, что алгоритм является гибридным, а следовательно имеет в себе преимущества статистических и структурных методов. Также количество ключевых слов, выделенных автором наборов данных и извлеченных с помощью методов, в отдельных статьях сильно

различалось, что снижало оценку качества. Здесь стоит помнить о проблематике задачи, а именно о сложности при разработке методов извлечения ключевых слов, основываясь на интуитивном алгоритме людей.

ЗАКЛЮЧЕНИЕ

В ходе данной работы были изучены основные понятия и определения процесса извлечения ключевых слов, рассмотрены методы извлечения ключевых слов - TF-IDF. TextRank, RAKE, произведена их реализация и проверка на размеченных данных на языке программирования Python, а также проведен сравнительный анализ методов. Лучший результат из рассмотренных алгоритмов показал RAKE с оценкой качества: точность - 34%, полнота - 42%, f-мера - 37%.

Данная задача является довольно сложной и качество сильно зависит от исходных данных.

Основные источники информации:

1 Дубинина Е.Ю. Выделение ключевых слов текста научной статьи в процессе создания автоматического реферата / Е.Ю. Дубинина // Вестник ВГУ. Серия: Филология. Журналистика. – 2020. – № 1. – С. 26–28.

2 Тихонова Е.В., Косычева М.А. Эффективные ключевые слова: стратегии формулирования / Е.В. Тихонова, М.А. Косычева // Health, Food & Biotechnology. – 2021. – Vol. 3 (4). – P. 7–15.

3 Ноздрина Т.Г. Особенности восстановления текстов – оригиналов на основе ключевых слов / Т.Г. Ноздрина // Современные проблемы науки и образования. – 2015. – № 1-2. – С. 167

4 Ванюшкин А.С., Гращенко Л.А. Методы и алгоритмы извлечения ключевых слов / А.С. Ванюшкин, Л.А. Гращенко // Новые информационные технологии в автоматизированных системах. – 2016. – №19.

5 Jones K.S. A statistical interpretation of term specificity and its application in retrieval. // J. Doc. – 1972. – Vol. 28. – P. 11–21.

6 Mihalcea R., Tarau P. TextRank: Bringing Order into Text // Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing – 2004. – P. 404–411.

7 Rose S., Engel D., Cramer N., Cowley W. Automatic Keyword Extraction from Individual Documents. // Text Min. Appl. Theory. – 2010. – P. 1–20.