Страничка для ТЗ

Содержание

1 Введение

В 21 веке текстовая ткань современного общества претерпело радикальные изменения в связи с продолжающейся информационной революцией. Количество документов, доступных в Интернете и в других место, ошеломляет. Люди, предприятия, группы, организации, учреждения и правительство не только оставляет "цифровые следы"при использовании Интернета. Миллионы пользователей интернета, профессионалов или любителей создают миллиарды веб-страниц и документов. Каждый день создается огромное количество онлайн-текстов для различных целей, по разным вопросам, в разных странах, на всевозможных языках и в многообразных онлайн средах: пользовательский контент в блогах и на сайтах социальных сетей, электронная почта, блоги, новости, научные работы и т.д. Более того, по всему миру государства, институты, библиотеки, музеи цифвровизируют свою материалы и выкладывают его всемирную паутину, что бы информацию для бизнеса, науки, исследований, развлечений можно было получить через любое доступное нам устройство: телефон, планшет, компьютер и т.д. [2]

Традиционные медиа такие как газеты и телевидение быстро мигрируют в интернет. Новостные газеты или другие СМИ обновляют новостые ленты почти в реальном времени, что позволяет интересующимся получать свежую информацию. Поисковые системы только усугубили ситуацию, делая все больше и больше документов доступными всего в несколько нажатий клавиш на вашей клавиатуре. Таким образом, интернет и веб контент стали наиболее эффективными ресурсами для исследования современной экономики, культуры, политики, человеческого общения и взаимодействия людей. [2]

На сегодняшний день, количество опубликованных документов достигает 1 биллиона веб-страниц [1]. Такое очень огромное колличество информации делает задачу индексирования и поиска достаточно затруднительной, тем более преобладающие большинство документов не имеет ключевых слов (выражений) отсутствие которых заставляет пользователя полностью прочитать документ что бы получить общее представление о информации. Проставлять в ручную ключевую информацию для текста быстро превращается в раздражающую задачу. При таком огромном количестве документов ручное проставление является невозможным. Для того что бы автоматизировать дан-

ный процесс часто используются программы для извлечения ключевых слов, которые используется для поиска ключевой идее текста и извлечения/создания ключевых слов текста Обычно результат данной работы представляет из себя от 5 — 15 ключевых значений, которые предстваляют информацию пользователю или специальным машинам общую информацию о документе.

Целью данной работы является разработка метода извлечения ключевых словосочетаний или слов из текста электронных документов. Для достижения поставленной выше цели необходима решить следующие задачи:

1) Анализ темы и предметной области

2) Изучить существующие методы решения поставленной цели

3) Реализовать алгоритмы для извлечения ключевых слов.

4) Тестирование и замер результатов реализаций

5) Анализ полученных результатов и сопоставление их друг с другом

6) Вывод по итогам проекта

2 Аналитический раздел

2.1 Понятие ключевого слова

Первые попытки теоретического решения проблемы выделения ключевых ("опорных "обобщающих") слов была предпринята в работе А.Н. Соколова Внутренняя речь и мышление [6]. Основы современного понимания ключевых слов, можно сформулировать следующим образом [7]:

1) ключевые слова отображают тему текста;

2) их упорядоченность в наборе ключевых слов может трактоваться как эксплицитно невыраженная тема текста;

3) набор ключевых слов рассматривается как один из минимальных вариантов "текста";

4) такого типа "текст"характеризуется "ядерной"цельностью и минимальной связностью

Ключевые слова - это одно или многокомпонентные лексические группы, отражающие содержание документа [3]

2.2 Извлечение ключевых слов

Извлечение ключевых слов (Keyword extraction) - это задача по автоматическому определению набора терминов которые наилучшем образом описывают объект документа. При изучении терминов, представляющих наиболее релевантную информацию, содержащуюся в документе, используется различная терминология: ключевые фразы, ключевые сегменты, ключевые термины, или просто ключевые слова. Все выше перечисленные синонимы имеют одну и туже функцию - охарактеризовать обсуждаемую тему в документе [4]. Извлечение маленького множество элементов представляющих из себя от одного и более терминов из одного документа является важной проблемой в "Информационном поиске"(Information Retrieval, IR), "Интеллектуальном анализе текста"(Text mining, TM) и в "Обработке естественного языка"(Natural Language Processing, NLP).

Ключевые слова нашли широкое применение в запросах к системам информационного поиска, по сколько их легко определить, пересмотреть, запомнить и поделиться. По сравнению с математическими сигнатурами, они независимы от любого корпуса и могут применяться в нескольких корпусах и системах ИП [5] Так же ключевые слова используются для улучшения функциональности Информационно поисковых систем. Другими словами они могут быть использованы для создания автоматического индекса для коллекции документов или, в качестве альтернативы, могут использовать для представления документов в задачах категоризации или классификации [1].

Извлечение краткого изложения - это основная задача многих IR и NLP приложений включая в себя автоматическое индексирование, обобщение, управление документами, высокоуровневое семантическое описание, категоризацию или кластеризацию текста, документов или веб-сайтов, поиск по категориям, создание словарей для конкретной области, распознавание имен, определение тем, отслеживание и т.д. Благодаря тому что назначение ключевых слов документам в ручную является очень дорогостоящей, трудоемкой и утомительной задачей и дополнительно к этому количество доступных цифровых документов растет, автоматическое извлечение ключевых слов привлекло интерес исследователей в последние несколько лет. Хотя приложения для извлечения ключевых слов обычно работаю с отдельными документами, извлечение так же используется для более сложных задач (Извлечение из коллекции текстов, всего веб-сайта и т.п.)

2.3 Общая схема извлечения ключевых слов из текста

Общая схема извлечения ключевых слов из текста практически одинакова для всех используемых методов и состоит из следующих шагов:

1) предварительная обработка текста:

2) а) исключение элементов маркировки;

   б) приведение слова к словарной форме;

   в) удаление стоп слов, не несущих смысловой нагрузки (предлоги, союзы, частицы, местоимения, междометия и т.д.)

3) отбор кандидатов в ключевые слова;

4) фильтрация кандидатов в ключевые слова (анализ значимых признаков для каждого кандидата)

2.4 Систематизация методов

Методы назначения ключевых слов можно условно разделить 2 категории:

1) назначение ключевых слов;

2) извлечение ключевых слов;

Оба они вращаются вокруг одной и той же проблемы - выбора лучшего ключевого слова. При назначение ключевых слов, оные выбираются из контролируемого словаря терминов или предопределенной таксономии, а документы подразделяются на классы в соответствии с их содержанием. Извлечение ключевых слов обогащает документ ключевыми словами, которые явно упоминаются в тексте. Слова, встречающиеся в документе, анализируются с целью выявления наиболее репрезентативных из них, обычно исследую 2 свойства источника (частота и длина). Обычно извлечение ключевых слов не используется предустановленный словарь для определения ключевых слов.

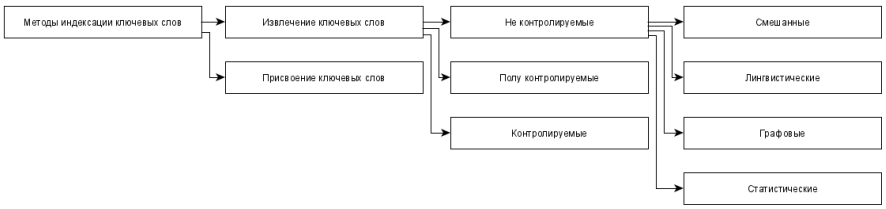

Рис. 1. Классификация методов извлечения ключевых слов

НУЖНО ДОБАВИТЬ ПРО SUPERVIZED И Т.Д

Из ходя из работы [?] методы могут быть разделены на следующие группы:

1) статистический подход;

2) машинное обучение;

Или более детализировано:

1) статистический подход;

2) гравфовый подход;

3) лингвистический подход;

4) подход через машинное обучение;

5) остальное;

2.4.1 Статистические методы

Статистические методы извлечения ключевых слов работают на основе численных данных, говорящих о встречаемости слова в тексте.

2.4.2 Графовые

2.4.3 Лингвистические

2.4.4 Гибридные

2.5 Вывод

Список использованных источников

1. YAKE! Keyword extraction from single documents using multiple local features // URL: *https* : *//www.sciencedirect.com/science/article/abs/pii/S*0020025519308588 (Дата обращения 12.01.2022)

2. Textual Analysis: A Beginner's Guid // URL: *http* : *//www*1.*cs.columbia.edu/ sbenus/Teaching/APTD/McKee*$_C$*h*1.*pdf* (Дата обращения 26.01.2022)

3. Automatic keyphrases extraction based on NLP and statistical methods // URL: *https* : *//www.researchgate.net/publication/*220827238$_A$*utomatic*$_K$*eyphrase*$_E$*xtractio* (Дата обращения 08.02.2022)

4. Keyword extraction from a single document using centrality measures // URL: *https* : *//www.researchgate.net/publication/*221205058$_K$*eyword*$_E$*xtraction*$_f$*rom*$_a$*Sing* (Дата обращения 08.02.2022)

5. Michael W. Berry Text Mining Application and Theory

6. А.Н. Соколов Внутренняя речь и мышление // URL: *https* : *//search.rsl.ru/ru/record/*01008431174 (Дата обращения 08.02.2022)

7. Современные методы автоматизированного извлечения ключевых слов из текста

8. Automatic keyword prediction using Google similarity distance // URL: *https* : *//www.sciencedirect.com/science/article/pii/S*0957417409006745

9. Методы и модели автоматического извлечения ключевых слов // URL: *https* : *//cyberleninka.ru/article/n/metody − i − modeli − avtomaticheskogo − izvlecheniya − klyuchevyh − slov*