Страничка для ТЗ

Содержание

Todo list

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 данный 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 Введение

Заполнить введение

2.2 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]

2.3 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).

Добавить описание TM, NLP, IR

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.)

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

Переосмыслить систематизацию 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. В данной работе "Назначение ключевых слов"рассматриваться не будет.

Изучив работы [8] методы могут быть разделены на следующие группы:

1) статистический подход;

2) машинное обучение;

Или более детализировано:

1) статистический подход;

2) лингвистический подход;

3) подход через машинное обучение;

4) комбинированный;

5) остальное;

Так же стоит отметить что способы извлечения ключевых слов можно категоризировать на контролируемые (supervized) и неконтролируемые (unsupervized) методы. Самым распространенным методом из первой категории является KEA (keyphrase extraction algorithm). Данный метод использует Наивный Байесовский алгоритм для обучения и извлечения ключевых выражений. Двумя главными проблемами контролируемых методов является обязательное наличие тренировочных данных с в ручную размеченными ключевыми словами и привязанности к доменной области, на которой они обучались 1.

Рис. 1. Классификация методов извлечения ключевых слов

2.4.1 Статистические методы

Статистические методы извлечения ключевых слов работают на основе численных данных, говорящих о встречаемости слова в тексте. К статистическим методам относятся:

1) TF-IDF;

2) YAKE;

3) Rake

TF-IDF - это аббревиатура, скрывающая за собой "Term Frequency - Inverse Document Frequency"(Частота термина и инвертированная частота в документе).

Еще в 1958 году Ганс Петер Лун в своей статье "Автоматическое создание литературных рефератов"предложил, что "частота появления слов в статье обеспечивает полезное измерение значимости слов"что до сих пор, вероятно, является одни из самых важных аспектов в области "Информационного поиска". Данный метод используется во всех известных поисковых системах от Goggle и Yahoo и заканчивая специализированными поисковыми решениями, такими как ElasticSearch и ManticoreSearch [12]. Как можно понять сверху идет речь про TF.

В 1972 Карен Спарк Джонес в работе "A statistical interpretation of term specificity and it's application in retrieval"в журнале "Journal of Documentation"[11] предложил "Полнота описания документа - это количество содержащихся в нем терминов, а специфичность термина - это количество документов к которым он относится". В будущем данная работа станет известна как "inverse document frequency"или IDF.

С помощью tf-idf вместо представления термина в документе его необработанной частотой (количество вхождений) или его относительной частотой (количество терминов, деленное на длину документа), каждый термин взвешивается путем деления частоты термина на количество документов, корпусы которых содержат данное слово.

Yake - Yet Another Keyword Extractor. Данный метод был разработан компанией LIAAD и первое упоминается в работе "YAKE Keyword extraction from single documents using multiple local features"[14] Это легковесный автоматический подход извлечения ключевых слов, основанный на статистических характеристиках текста, извлеченных из отдельных документов, для выбора наиболее важных ключевых слов текста. Авторы данного способа выделяют следующие особенности [14]:

1) не требует обучения

2) не требует корпуса тематически заготовленных текстов;

3) не зависит от области использования;

4) не зависит от языка;

5) масштабируемость;

6) результат не зависит от частоты термина;

RAKE - Rapid Automatic Keyphrase Extraction. К особенностям относят [15]:

1) не требует обучение;

2) не зависит от области применения;

3) не привязан к определенному языку;

4) работает на одиночном документе

Во время работы над RAKE у авторов стояла задача разработать механизм, метод извлечения ключевых слов, эффективно работающий с отдельными документами, позволяя примение к динамическим коллекциям, легко применимый к новым доменам.

Преимуществами чисто статистического подхода являются универсальность алгоритмов извлечения ключевых слов, простота реализации и отсутствие необходимости в трудоемких и время-затратных процедурах построения лингвистических баз знаний. Несмотря на указанные преимущества статистических методов извлечения ключевых слов, чисто статистические методы часто не обеспечивают удовлетворительного качества результатов. При этом область их применения ограничена языками с бедной морфологией, такими как английский, где частотность словоформ одной лексемы велика. Чисто статистические модели извлечения ключевых слов, удовлетворительно работающие, например, на материале английского языка, не пригодны для естественных языков с богатой морфологией, в частности, для русского языка, где каждая лексема характеризуется большим количеством словоформ с низкой частотностью в каждом конкретном тексте [9].

К статистическим методаи использующие графовые методам относятся:

1) Textrank;

2) SingleRank

3) ExpandRank

4) TopicRank;

5) TopicalPageRank

6) PositionalRank

7) MultipartiteRank

8) Rake;

2.4.2 Лингвистические

2.4.3 Гибридные

2.5 Заключение

3 Конструкторский

3.1 Введение

3.2 RAKE

3.3 YAKE

3.4 Заключение

4 Технологический

4.1 Введение

4.2 Системные требования

4.3 Язык программирования

4.4 Среда разработки

4.5 Заключение

5 Эксперементальный раздел

6 Заключение

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*

10. (Для понимания TF-IDF) Understanding Inverse Document Frequency: On Theoretical Arguments for IDF URL: *https* : *//www.researchgate.net/publication/*238123710$_U$*nderstanding*$_I$*nverse*$_D$*ocume*

11. A statistical interpretation of term specificity and its application in retrieval // URL: $http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.115.8343\&rep=rep1\&type=pdf$

12. TF-IDF in a nutshell // URL: $https://towardsdatascience.com/tf-idf-in-a-nutshell-b0ff082fbbc$

13. // URL: $https://courses.ischool.berkeley.edu/i256/f06/papers/luhn58.pdf$

14. YAKE // URL: $https://www.sciencedirect.com/science/article/pii/S0020025$

15. Automatic Keyword Extraction from individual Document // URL: $https://www.researchgate.net/publication/227988510_Automatic_Keyword_Extraction_$