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. Жители и 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 создается огромное количество онлайн-текстов для разных целей, по разным вопросам, в разных странах, на разных языках и в 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 в семирную паутину, что бы информацию можно было получить через компьютеры, телевизоры, телефоны, планшеты для бизнеса, науки, исследований, развлечений 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 болье и больше документов доступными всего в несклько 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

В данном разделе описана предметная область. Указана классификация методов извлечения ключевых слов и

2.1 Предметная область

Перед тем как углубимся в предметную область стоит разобраться разницей между Анализом текста (Text Analysis), Интеллектуальным анализом текста (Text mining) и Аналитикой текста (Text Analytics). Анализ текста и Интелектуальный анализ текста это одно и тоже, они синонимичны друг другу.

TODO Дописать сравнение

2.1.1 Что такое Интеллектуальный анализ текста?

Интеллектуальный анализ текста, так же известный как Интеллектуальный анализ текстовых данных - это процесс преобразования неструктурированной, сырой информации в структурированный формат для выявления значимых закономерностей и новых идей.[6] Применяя передовые аналитические методы, такие как метод наивного Байеса, метод опорных векторов (SVM) и другие алгоритмы глубокого обучения, компании могут исследовать и обнаруживать скрытые взаимосвязи в своих неструктурированных данных.

Тест - это один из самых распространенных типов данных в базах данных. В зависимости от базы данных, данные могут быть организованы как:

1) структурированные данные: это данные предствалены в табличном формате с многочисленными строками и столбацами, что упрощает их хранение и обработку для анализа и алгоритмов машинного обучения;

2) не структурированные данные: эти данные не имеют предопределенного формата данных. Он может включать текст из источников, таких как социальные сети или обзоры продуктов, или мультимедийные форматы, такие как видео и аудиофайлы;

3) полуструктурированные данные: как следует из названия эти данные предстваляют собой смесь форматов структурированных и неструктурированных данных. Хотя у него есть некторая организация, у него недостаточно структуры для удовлетворения требований реляционной базы данных. Примера частично структурированных данных включают файл XML, JSON, HTML.

Так как 80% иформации в мере относятся к неструктурированному формуту, интелектуальный анализ текста является черезвычайно выжным.

2.2 Классификация методов извлечения ключевых слов

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. Conceptual Framework Internet statistics // URL: *https* : *//opendata.ellak.gr/wp* − *content/uploads/*2015/04/*ConceptualFrameworkInternetStatistics.pdf*

4. Data Mining // URL: *https* : *//www.ibm.com/cloud/learn/data* − *mining* (Дата обращения 26.01.2022)

5. Text mining // URL: *https* : *//monkeylearn.com/text* − *analysis/* (Дата обращения)

6. IBM Text mining // URL: *https* : *//www.ibm.com/cloud/learn/text* − *mining* (Дата обращения)