Тема

Метод автоматического извлечения **ключевых *слов/словосочетаний*** из текста документов.

Общая информация

С появлением интернета, у каждого человека появился доступ к неограниченному, постоянно пополняющемуся источнику данных: будь это новости, информационные посты или научная литература. На сегодняшний день, количество опубликованых документов достигает 1 билиона веб-страниц. Такое **хтоническое** количество 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 используются для поиска ключевой идее текста и извлеченинния\создания ключевых слов текста. Обычно результат данной работы представляет из себя от 5 — 15 ключевых значений, которые предсталяю информацию пользователю или специальным машинам общую информацию о документе.

Извлечение ключевых слов может быть полезно огромному количеству людей, которые хотят познакомится с той или иной темой. На пример: людям которые хотят познакомится с историей о короновирусе или студенту который хочет изучить определенный предмет.

Проблема:

Выделение общей информации из текста, с целью структуризации и ускорения поиска информации

Методы:

1. Статистические методы
    1. MF
    2. TF-ISF
    3. TF-IDF
    4. CSI

5. EB
6. TR
7. YAKE
8. KEA

2. Гибридные
    1. Использующие статистические методы и Машинное обучение

Алгоритм работы:

1. На вход подается документ, новость, статья (Думаю тут можно выбрать, может стоить даже поработать над новостными агрегаторами)
2. Документ обрабатывается методом

Возможные вариации тем:

1. Работа не только с бумажными источниками а на пример с Хабром или другом новостным агрегатором.
2. Получение доработонного\нового метода
3. Рекомендация по использованию уже существующих методов

Ограничения:

1. Тексты документов написаны с использованием Русского языка.
2. Для алгоритма не требуется большая документов или достаточно наличие нескольких.

Пример литературы:

1. https://www.sciencedirect.com/science/article/pii/S0957417416301464
2. https://reader.elsevier.com/reader/sd/pii/S0020025519308588?token=7AE28F9AF12ECA6A6A91C8E85932EAB17EE892D969E16EE6C1D0A4F9ACAFDCBAC7F1F9399657A734670F8E3BB94286BE&originRegion=eu-west-1&originCreation=20211219113500