**Практическая работа №4**

Язык программирования: Python

Библиотеки: PyPDF2, PDFMiner, PyMyStem3, PyMorphy2, Natasha, sklearn

Задание:

Выполнить загрузку данных (сборник статей в формате PDF).

Очистить тексты, токенизировать, лемматизировать.

Определить ключевые слова по метрике TF-IDF.

Сравнить результаты по ключевым словам (в текстах статей указан набор слов).

Доп. баллы:

Использовать еще один алгоритм для выделения ключевых слов, например, Rake, YAKE! или TextRank.

Выполнение:

1. Загрузка данных

Загрузить файл с помощью библиотеки для чтения pdf файлов (PyPDF2, PDFMiner и пр.).

1. Предобработка текста, токенизация и лемматизация

Необходимо разделить весь файл на отдельные статьи.

Далее очистить тексты от стоп-слов, лишних знаков, цифр, ссылок на литературу и прочего.

Для лемматизации можно также использовать NLTK, либо другие библиотеки:

PyMyStem3: <https://pypi.org/project/pymystem3/>

PyMorphy2: <https://pymorphy2.readthedocs.io/en/stable/>

Natasha: <https://www.helenkapatsa.ru/liemmatizatsiia/>

1. Метрика TF-IDF

Метрика TF-IDF статистический показатель, применяемый для оценки важности слова в контексте категории, документа или коллекции документов.

Как правило, TF-IDF определяется для каждого слова. Чем выше значение данного показателя, тем значимее слово в контексте категории, документа, коллекции. При этом данный показатель также позволяет учесть и широкоупотребляемые слова, понизив их значимость в контексте объекта для анализа.

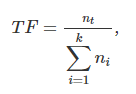
Формула для определения показателя имеет следующий вид:

**TF−IDF=TF\*IDF**

где **TF** — частота слова в конкретном документе,

**IDF** — обратная частота документа (популярность слова).

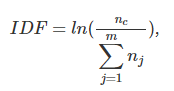
Частота слова в документе определяется по формуле:



где nt количество отдельных слов в документе,

 общее количество всех слов в документе.

Обратная частота документа (также часто называют инверсией частоты) определяется по формуле:



где

nc — количество документов всего,

 — количество документов, в которых содержится интересующее слово.

Векторизация по этой метрике можно произвести с помощью библиотеки sklearn: <https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html>

1. Сравнение результатов

У каждой статьи указаны ключевые слова. Сравнить для статей наиболее важные слова по метрике с тем, что уже указано. Сделать вывод.

В чем плюсы и минусы метрики TF-IDF?

Для каких задач лучше всего использовать данную метрику?

1. Доп. баллы

Используя другой алгоритм выделить ключевые слова. Сравнить результаты с метрикой TF-IDF. Объяснить, как работает алгоритм.