# Тема

Метод автоматическое извлечения ключевых слов из электронного документа на русском языке

# Предметная область

Решаемая задача лежит в областях Информационного поиска (Information Retrieval), обработке текстов на естественном языке (Natural Language Processing) и принадлежит набирающему популярность и актуальному направлению “Data Science”

# Актуальность

В последнее время с учетом роста не только популярности интернета, но и его реального использования во всех сферах жизни, все большее и большее количество текстовой информации не сопровождается ключевыми терминами, что превращает задачу изучения и сбора информации в затруднительную задача.

Необходимость автоматизации такого рода задач требует разработки систем извлечения ключевых слов с возможностью автоматического определения ключевых слов в тексте.

Возможное применение, решаемая задача, может найти в вопросах кластеризации, классификации, информационных системах, создании авторефератов,

# Проблема:

После детального изучения существующих решений, обнаружилось, что совсем небольшое их количество имеют возможность извлечения ключевых слов/словосочетаний из текстов на русском языке.

Так же стоит выделить основные проблемы методов:

1. Для обучаемых методов требуется большого размера корпус текстов
2. Корпусы текстов должны относится к одной области, что говорит о снижения работоспособности в других областях
3. В большинстве своем статистические метод работаю только с частотой вхождения слово в документ или в корпус документов

Однако стандартный статистический метод извлечения весьма удачно проявляет себя на тех естественных языка, в которых есть понятие как прямой порядок и имеют бедную морфологию. Проводить автоматические извлечение ключевых слов их текстов на русском языке представляет из себя не самую простую задачу.

# Описание решения:

Для решения поставленной проблемы предполагается создать ПО, которое будет извлекать ключевые слово\словосочетания из текстов на русском языке и разделять их по ключевым словам

Вход: Текст или набор файлов формата pdf

Выход: Кластеризация текстов по ключевым словам

Цель:

Реализация комбинации методов извлечения ключевых слова/словосочетаний из единичного документа на естественном языке.

Задачи:

1. Изучить предметную область
2. Провести классификацию метод по использующимся подходам и способам представления данных
3. Изучить особенности и условия работы
4. Изучить возможность комбинирования методов
5. Создать программное обеспечение, реализующее метод
6. Сделать инструментарий для автоматизации извлечения ключевых слов
7. Провести эксперименты и обосновать их результаты
8. Сделать выводы относительно корректности работы методы