# Метод кластеризации текстов на естественном языке.

#### Обзор предметной области:

Решаемая задача лежит в области компьютерной лингвистики и принадлежит набирающему популярность и актуальность направлению «Data mining» – Автоматическая обработка больших объемов данных.

#### 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, на первый взгляд не пригодной для анализа появляется в сети.

Поскольку эта информация представляет всесторонний интерес, стоит задача получения конкретных сведений из простых и не формализованных предложений, состоящих из простых слов, состоящих из букв, не несущих компьютеру никакой информации.

Возможное применение решаемая задача может найти в вопросах выявления плагиата и самоповторений автора при анализе диссертаций и научных работ.

#### Проблема:

После тщательного изучения программ представленных на рынке обнаружилось, что совсем небольшое их число имеют возможность анализа и кластеризации текстов на русском языке по заданным параметрам.

Так же было не трудно заметить, что в большинстве своем программы оказались огромными комплексами, разработанными профессиональными лингвистами, включающими в себя корпусы языков, словари, морфологически и синтаксические анализаторы.

Именно на основании результатов синтаксического анализа в этих программах и принималось решение о том в какую категорию отнести данный текст.

Однако стандартный метод синтаксического анализа весьма удачно работает на тех естественных языках, в которых есть такое понятие как “прямой порядок” слов в предложении (напр. английский). Проводить синтаксический анализ предложений на русском языка – очень непростая и нетривиальная задача.

#### Описание решения:

#### Для решения поставленной проблемы предлагается создать ПО, которое будет выполнять синтаксический анализ текстов на русском языке и разделять их на заданные кластеры.

#### Вход: набор текстовых файлов

#### Выход: данные о результатах кластеризации

#### **Цель:**

#### Реализация метода кластеризации текста на естественном языке.

#### Задачи:

1. Провести подробный анализ предметной области
2. Проанализировать существующие методы-аналоги
3. Обосновать возможность и необходимость разработки
4. Создать программное обеспечение, реализующее метод
5. Сделать инструментарий работы с текстами как можно более автоматическим
6. Провести эксперименты и обосновать их результаты
7. Сделать выводы относительно корректности работы метода