Белевцев Никита ФКН ПМИ 104

**Итоговый отчет**

# Утилита для очистки текстов от обсценной лексики (проект)

Программа, которая обрабатывает тексты, удаляя предложения с нецензурными словами.

Она представляет  собой пользовательское приложение, обрабатывающее текстовый документ, и впоследствии может использоваться, как часть web-сервиса SketchEngine - сервиса обработки текстовых корпусов.

SketchEngine позволяет людям, изучающим языковое поведение (лексикографам, исследователям в корпусной лингвистике, переводчикам или учащимся) проводить поиск больших текстовых коллекций в соответствии со сложными и лингвистически мотивированными запросами.

## Цели

Текстовые корпуса - подобранная и обработанная по определённым правилам совокупность текстов, используемых в качестве базы для исследования языка.

Целью стало очищение текстовых корпусов от обсценной лексики, чтобы ими могли пользоваться лица, не достигшие 18 лет.

## Задача

Организовать очистку текста от обсценной лекскики.

В рамках данного проекта используется язык программирования Python, а также дополнительные его библиотеки NLTK, pymorphy2 и gensim.

- NLTK (Natural Language Toolkit) - библиотека для обработки текстов на естественных языках, к которой прилагаются готовые корпуса (в том числе необходимые для данного проекта), а также программные интерфейсы для удобного доступа к ним. Библиотека позволяет токенизировать исходные тексты - разбивать их на отдельные слова.

- Pymorphy2 - морфологический анализатор, который позволяет приводить слова к нормальной форме (Для существительных - им. п., ед. число. Для глаголов - форма инфинитив. И т.п.), а также работать с формами слова, ставить его в нужный падеж, число, род и т.д.

- gensim - используется для расчета статистических показателей текста, его индексирования.

Итоговая программа получает текст, разбивает его на предложения, а предложения на слова, приведенные к начальной форме. Это позволяет быстро сравнивать их с запрещенными словами и удалять предложения с оными.

Она использует дополненный список запрещенных доменных имен сегмента .рф в качеств "стоп-слов".

Конечно, программа не удаляет 100% обсценной лексики. Основным препятсвием является богатство и сложность строения слов русского языка.

Чтобы уменьшить процент нефиксируемой лексики список "стоп-слов" пришлось пополнять. Были найдены и добавлены в список наиболее частые матерные слова, которые программа не могла привести к начальной форме.

Итогом стала высокоэффективная программа, которая может использоваться обычными пользователями, а может быть переработана под очистку больших текстовых корпусов.