Проектная работа на тему:

# Natural language processing

Определение автора по тексту

Отчет по выполненной работе.

Выполнил студент 1 курса, 156 группы

Факультета компьютерных наук НИУ ВШЭ

Алиев М.А

## Основные проблемы.

В настоящей работе задача определения авторства ставится следующим образом. Пусть имеются достаточно длинные фрагменты произведений ряда авторов на русском языке. Про некоторый анонимный фрагмент известно, что он принадлежит одному из этих авторов, но какому - неизвестно. Требуется узнать, кому именно. На основе результатов, полученных в ходе обработки текстов обучающей выборки с достаточно высокой степенью вероятности возможно определить авторство. Избранный метод базируется на учете статистики употребления слов и биграмм слов.

Основная проблема состоит в извлечении из текста статической информации. Требуется очистка текста от всех знаков пунктуации, также при подсчете слов желательно не учитывать местоимения, предлоги, союзы и тд. Главной задачей является определение начальных форм слов для корректного подсчета их частоты использования (слова “дерева”, “дерево” и “деревья” не должны считаться как разные).

## Использованные методы.

Для очищения текстов от знаков пунктуации и стоп-слов использованы регулярные выражения (в качестве стоп-слов были выбраны все слова размера 3 и меньше). Для каждого слова проводилась его лемматизация (был составлен словарь нормальных форм для того, чтобы не тратить лишнее время на обработку встречавшихся ранее элементов) и расчет частоты его использования. Таким образом для каждого текста определен вероятностный вектор использования юниграмм и биграмм слов. Сравнение с проверяемым произведением проводится на основе метода косинусов и метода ближайших k соседей (экспериментально было выбрано значение 50).

## Использованные библиотеки.

* Re – использование регулярных выражений.
* Pymorphy2 – лемматизация слов.
* Nltk – токенизация текста, разбиение на биграммы.
* Os, picke, argparse, sys – работа с файлами и каталогами.

## Литература.

Steven Bird, Ewan Klein, and Edward Loper

Natural Language Processing with Python

<http://www.nltk.org/book/>

## Руководство

Программа поддерживает вызов из командной строки. Единственным аргументом при вызове является путь к файлу с проверяемым текстом.

### python readfile.py –path “Путь к файлу”