МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)
Факультет информационных технологий и прикладной математики
Кафедра вычислительной математики и программирования

Отчет по лабораторной работе №1
«Токенизация»
по курсу
«Обработка текстов на естественном языке»

Группа: М80-108М-17
Выполнил: Забарин Н.И.
Преподаватель: А.Л. Калинин

Москва, 2018

Задание

Реализовать процесс разбиения текстов документов на токены, который потом будет использоваться при индексации. Выработать правила, по которым текст делится на токены. Указать достоинства и недостатки выбранного метода. Привести примеры токенов, которые были выделены неудачно, объяснить, как можно было бы поправить правила, чтобы исправить найденные проблемы.
В результатах выполнения работы нужно указать следующие статистические данные:

- Количество токенов.
- Среднюю длину токена.

Кроме того, нужно привести время выполнения программы, указать зависимость времени от объёма входных данных. Указать скорость токенизации в расчёте на килобайт входного текста.

Решение

Лабораторная работа написана на языке Python3, сначала из текста удаляются(заменяются на пробелы) все «плохие» символы, далее текст разбивается по пробелам.

Результат работы:

- Средняя длина термина 5.593 символов.
- Среднее время обработки документа 0.0004 секунды.
- Суммарный объем текста 20.7 Gb.
- Время обработки всего корпуса документов 63 минуты.
- Средняя скорость обработки 5,5 Mb/sec.
- Число уникальных терминов 11141536.
- Число терминов 2467964568.
- Топ-10 высокочастотных терминов:
    1. the
    2. and
    3. was
    4. for
    5. that
    6. with
    7. http
    8. from
    9. this
    10. utc

Выводы

Поскольку исходная кодировка текста UTF-8, нет необходимости обрабатывать огромное число специальных символов, с другой стороны мы теряем поиск по различным нетривиальным символам.

Токенизация — очень гибкая и нетривиальная задача, код хорошо работающий для одной задачи не подойдет для другой. Решение сильно зависит от реализации поиска и направленности текста.