

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

**Отчет по лабораторной работе №2**  
**«Закон Ципфа»**  
**по курсу**  
**«Обработка текстов на естественном языке»**

Группа: 80-106М

Выполнил: Демин И.А.

Преподаватель: Калинин А.Л.

Москва, 2019

## Задание

Для своего корпуса необходимо построить график распределения терминов по частотностям в логарифмической шкале, наложить на этот график закон Ципфа. Объяснить причины расхождения.

## Выполнение лабораторной работы

Закон Ципфа («ранг—частота») — эмпирическая закономерность распределения частоты слов естественного языка: если все слова языка (или просто достаточно длинного текста) упорядочить по убыванию частоты их использования, то частота  $n$ -го слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру  $n$  (так называемому рангу этого слова, см. шкала порядка).

Иными словами, если составить список из всех слов текста и отсортировать его в порядке убывания частоты( $F$ ) использования слов, то для любого слова произведение его порядкового номера( $R$ ) в этом списке на частоту использования будет равно постоянной величине( $C$ ).

$$FR = C$$

$F$  — частота использования слова;

$R$  — порядковый номер;

$C$  — постоянная величина;

Для выполнения этого был использован Counter из библиотеки collections. Сначала были получены Counter-ы для каждой статьи, а затем создан общий, после чего значения были отсортированы по порядку убыванию частоты и построены графики.

Расчет значений для закона Ципфа для точки из начала, середины и конца набора точек:

$a[0] = 230254$ , слово «В», частота - 230254

$a[n/2] = 220982$ , слово «титанов», частота - 2

$a[n] = 220980$ , слово «пулидор» (французский велогонщик), частота - 1

Как видно, самые часто встречающиеся слова — предлоги. Они достаточно сильно выделяются от определения, о котором говорится в законе, но затем частотность падает, значения усредняются и мы начинаем наблюдать почти одинаковые значения.

## Графики

График зависимости частоты от порядка в отсортированном массиве.

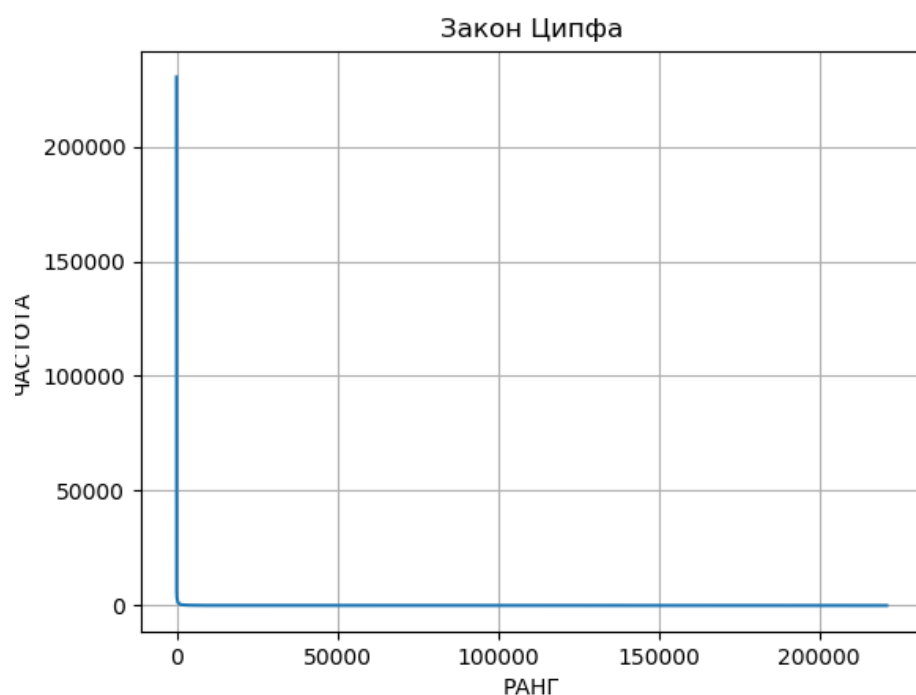
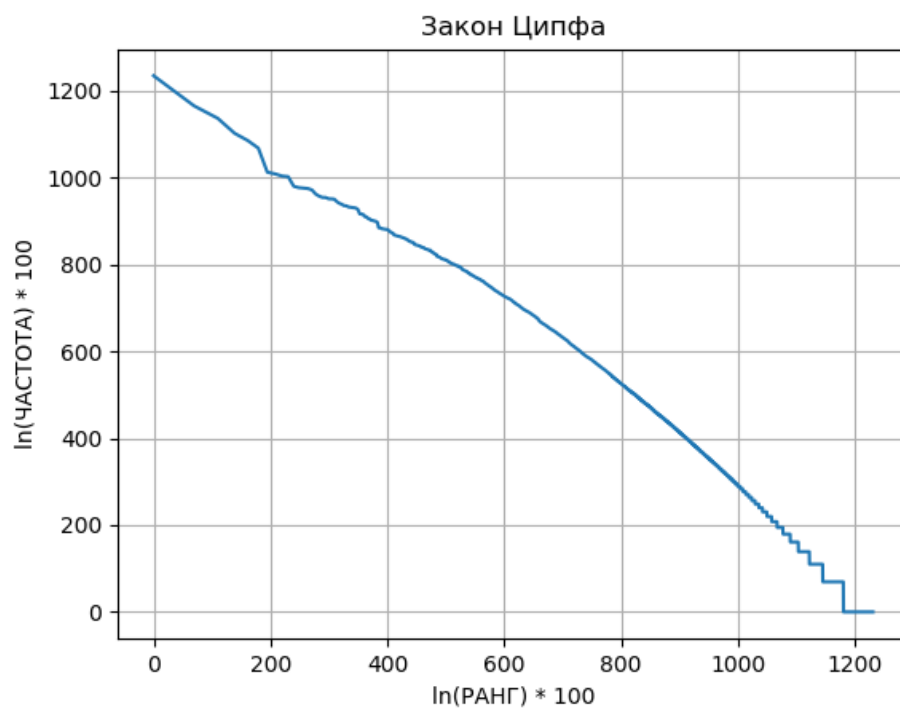


График в логарифмическом масштабе



## **Вывод**

В ходе лабораторной работы был рассмотрен закон Ципфа и применен к моему корпусу документов. Как видно из расчета закона по 3м точкам и графику в логарифмическом масштабе, то на данном корпусе документов и для данного разбиения текста на токены закон выполняется.