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

Отчет по лабораторной работе №2
«Законы Ципфа и Мандельброта»
по курсу
«Обработка естественного языка»

Группа: М80-106М
Выполнил: Забелин М. К.
Преподаватель: Калинин А.Л.

Москва, 2019

Лабораторная работа №2. Законы Ципфа и Мандельброта

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

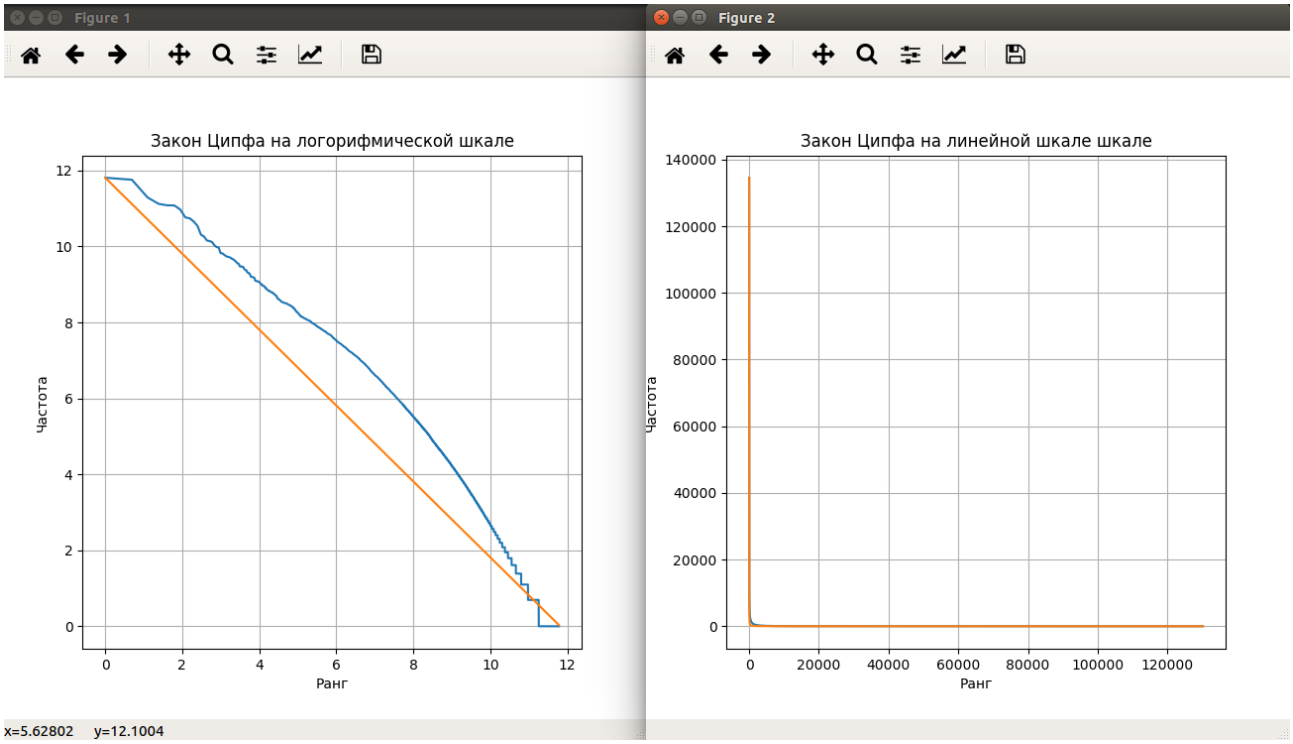
Закон Ципфа – эмпирическая закономерность распределения частот слов естественного языка, которая говорит, что что при расположении слов по частоте всречаемости в документах по убыванию, то частота слова обратно пропорциональна его номеру при расположении, т.е. freq = const/n, где const – это некоторая константа.

Для эксперимента я подставил вместо const – 1. В качестве токенов брал слова из всех документов, суммируя их частоту. В реузльтате чаще всего встречались предлоги и союзы. Слова, которые встретились в начале, середине и конце списка:

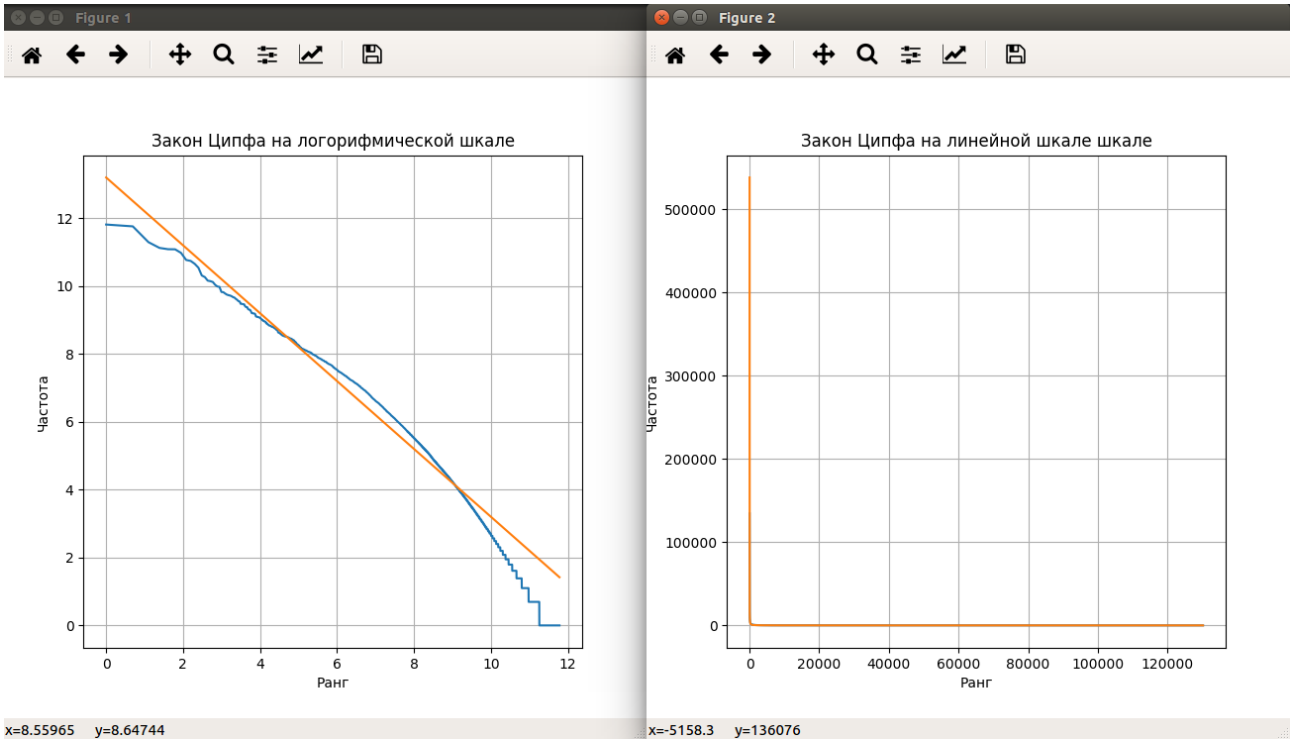
('в', 134552)

('нежелания', 2)

('4G-смартфона', 1)



Выставив const, равный 4, в равенстве Ципфа, получил более близкий к данным, полученным из моих статей, график:



Вывод

В ходе выполнения лабораторной я познакомился законом Ципфа и проверил его выполнение на собственном наборе токенов. В целом, правдивость этого закона подтвердилась, так как график функции, соответствующей закону и набор полученных частот токенов, очень похожи.

Ссылка на Git репозиторий

https://github.com/Mikhail-Z/MAI/tree/master/sem10/Informational%20Search/nlp1