**Московский авиационный институт**

**(национальный исследовательский университет)**

**Институт информационные технологии и прикладной математики**

**Кафедра вычислительной математики и программирования**

**Лабораторная работа №2 по курсу**

**«Обработка текстов на естественном языке»**

Студент: Е.М. Стифеев
Преподаватель: А.А. Кухтичев
Группа: М8О-109М-21
Дата: 28.10.21
Оценка:
Подпись:

**Москва, 2021**

## Лабораторная работа №2 «Закон Ципфа»

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

В качестве дополнительного задания, можно (но необязательно) подобрать константы для закона Мандельброта, наложить полученный график на график распределения терминов по частотностям. Привести выбранные константы.

# 1. Описание

В ходе ЛР3 по «Информационному поиску» был построен булев индекс, пригодный для подсчёта терминов и частот для выполнения данной ЛР. Этот индекс хранится в бинарном формате на локальном компьютере и был создан средствами языка C++. Для построения графиков будет использован язык Python, в котором предусмотрены библиотеки struct и ctypes для в том числе чтения бинарных файлов, созданных функцией языка C++ fwrite.

Закон Ципфа можно записать в виде:

$$f = \frac{k}{r} + c,$$

где $f$ – частота термина в корпусе, $r$ – позиция символа в отсортированном по частоте массиве в порядке убывания частот (начиная с 1), $k, r$ – неизвестные константы.

Для поиска констант $k$ и $c$ можно составить и минимизировать квадратичный функционал:

$$S(k,c) = \frac{1}{2}\sum_i \left(\frac{k}{r_i} + c - f_i\right)^2 \to \min_{k,c}.$$

Найдём частные производные:

$$\frac{\partial S}{\partial k} = \sum_i \left(\frac{k}{r_i} + c - f_i\right)\frac{1}{r_i},$$

$$\frac{\partial S}{\partial c} = \sum_i \left(\frac{k}{r_i} + c - f_i\right).$$

И приравняем их к нулю:

$$\begin{cases} \sum_i \left(\frac{k}{r_i} + c - f_i\right)\frac{1}{r_i} = 0 \\ \sum_i \left(\frac{k}{r_i} + c - f_i\right) = 0 \end{cases}$$

$$\begin{cases} k\sum_i \frac{1}{r_i^2} + c\sum_i \frac{1}{r_i} = \sum_i \frac{f_i}{r_i} \\ k\sum_i \frac{1}{r_i} + cN = \sum_i f_i \end{cases}$$

Решив, линейную систему, получим константы $k$ и $c$ для закона Ципфа.

Закон Мандельброта записывается в виде:

$$f = P(r+\rho)^{-B}.$$

Здесь $P, \rho$ и $B$ – неизвестные константы. Будем их искать из условия минимизации следующего функционала:

$$S(P,\rho,B) = \sum_i |P(r_i+\rho)^{-B} - f_i| \to \min_{P,\rho,B}$$

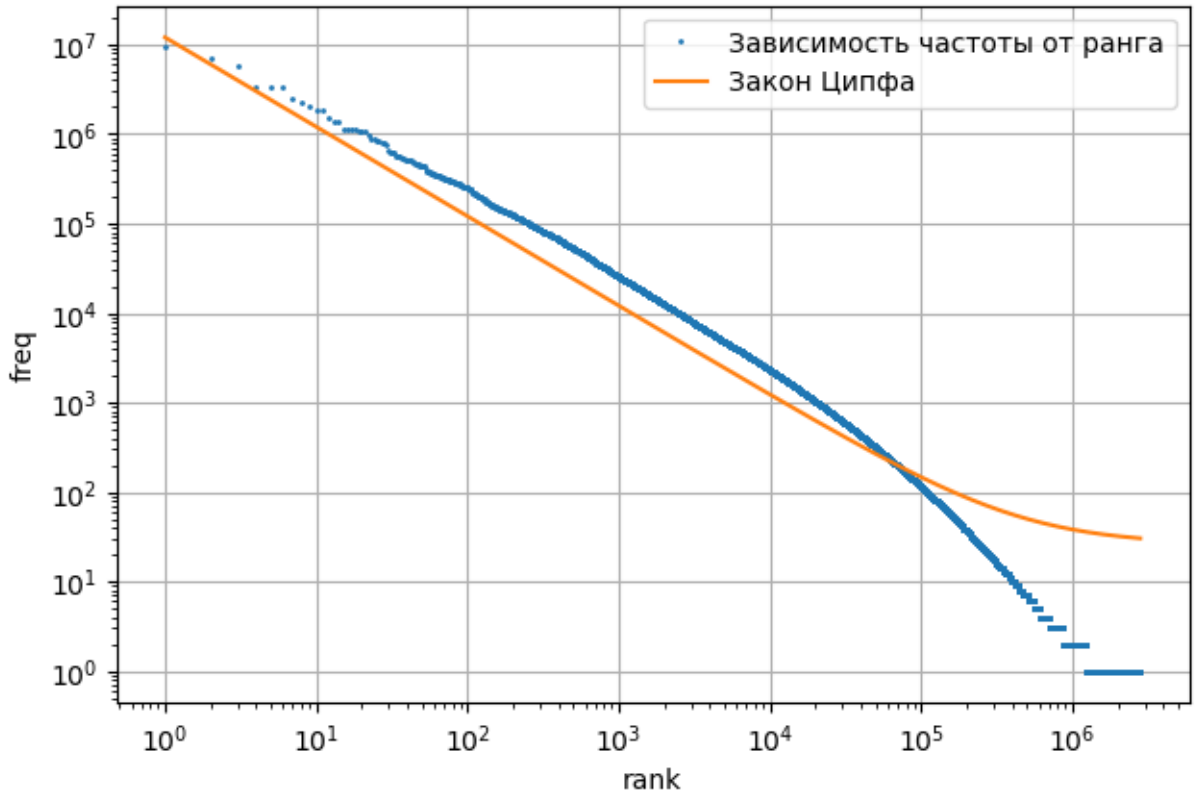
с помощью метода Нелдера-Мида, который не требует градиента, который в данном случае не существует в некоторых точках из-за модуля. Вариант с квадратичным функционалом в этом случае не пройдёт, т.к. мы не сможем аналитически найти значения параметров, а численный поиск (например, градиентный спуск) немного затруднён угрозой переполнения при возведение больших чисел в квадрат.

## 2. Исходный код

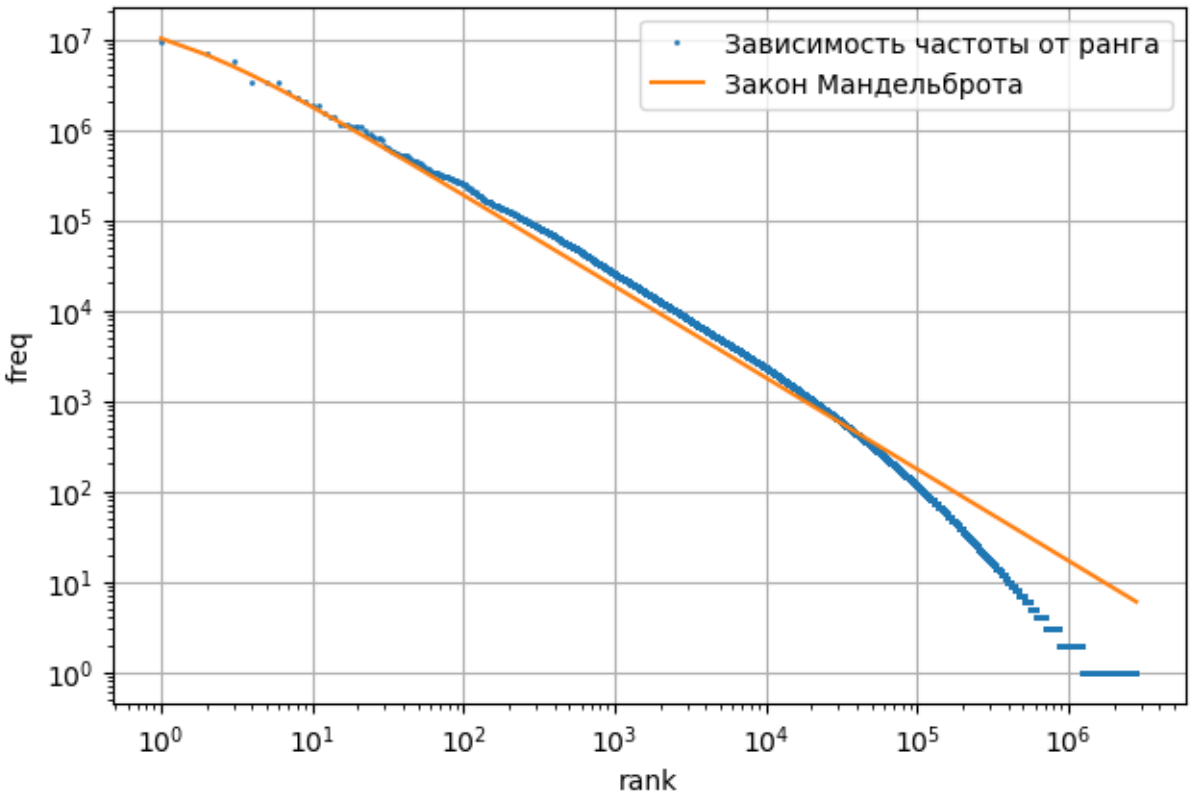
Написана программа на языке Python, которая считывает термы и частоты из бинарных файлов (более подробно описанных в ЛР3 по курсу «Информационного поиска») с помощью библиотек struct и ctypes, выполняет сортировку по частотам (стандартная функций sort), вычисляет константы для законов Ципфа и Мандельброта (с применением numpy и scipy) и наконец – строит графики (matplotlib, pyplot). Соответствующий Python-ноутбук (в формате Spyder) находится в директории с отчётом.

## 3. Выводы

После завершения получаем следующие графики:




$k = 12047804.33, c = 26.1$




$$P = 20110520.42, \rho = 0.9515, B = 1.012.$$

Как видно из графиков, закон Мандельброта точнее показывает распределение частот. Закон Ципфа ошибается для малых частот.

В ходе выполнения ЛР я повторил обработку бинарных файлов на Python’е, познакомился с моделями, моделирующими распределение частот терминов по их рангу.