**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.

# 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 С++. 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:

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

Для поиска констант и можно составить и минимизировать квадратичный функционал:

Найдём частные производные:

И приравняем их к нулю:

Решив, линейную систему, получим константы и для закона Ципфа.

Закон Мандельброта записывается в виде:

Здесь и – неизвестные константы. Будем их искать из условия минимизации следующего функционала:

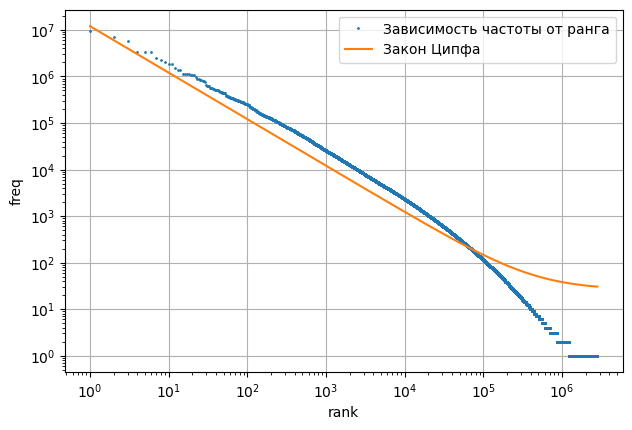
с помощью метода Нелдера-Мида, который не требует градиента, который в данном случае не существует в некоторых точках из-за модуля. Вариант с квадратичным функционалом в этом случае не пройдёт, т.к. мы не сможем аналитически найти значения параметров, а численный поиск (например, градиентный спуск) немного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

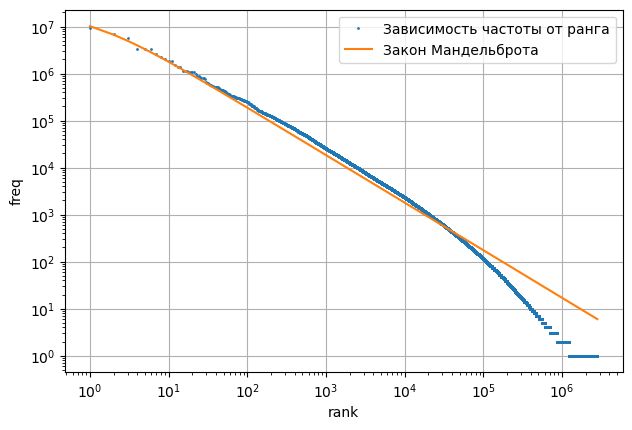
# Исходный код

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.

# Выводы

После завершения получаем следующие графики:





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.