МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

Институт №8 «Компьютерные науки и прикладная математика»

**Лабораторная работа №5**
**по курсу «Информационный поиск»**

***«Закон Ципфа»***

Выполнил: *Ермаков Ярослав Валерьевич*
Группа: *М8О-407Б-22*
Преподаватели: *А.А. Кухтичев*

**Москва, *2025***

## Условия

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

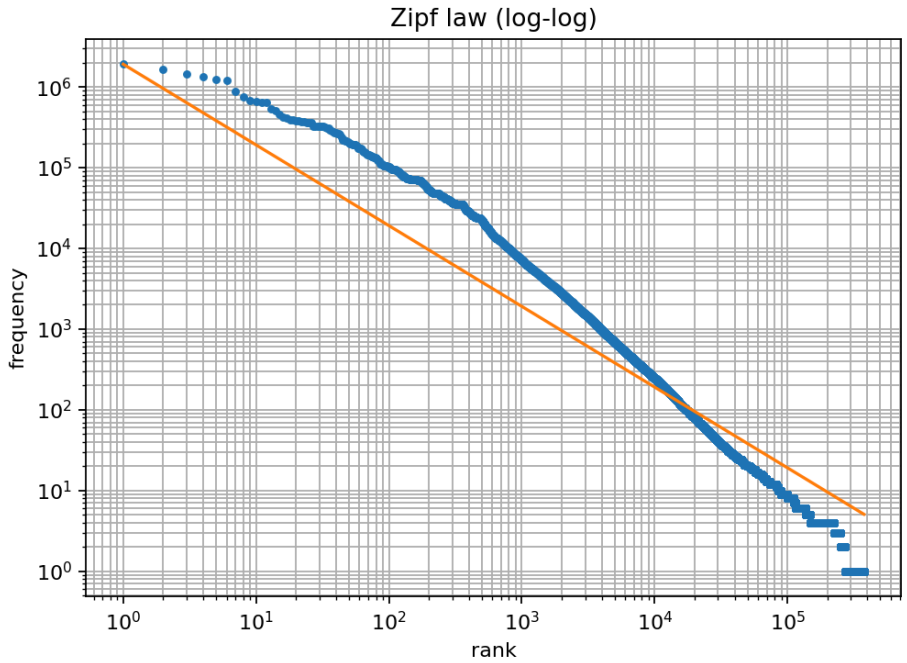
В качестве дополнительного задания, можно (но необязательно) подобрать константы для закона Мандельброта, наложить полученный график на график распределения терминов по частотностям. Привести выбранные константы.

## Описание

Целью лабораторной работы является проверка выполнения закона Ципфа на собранном корпусе. Для этого строится распределение терминов по частотам: термы сортируются по убыванию встречаемости, каждому терму присваивается ранг, после чего строится график зависимости **частоты от ранга** в **логарифмической шкале**. На полученный график накладывается теоретическая кривая по закону Ципфа, чтобы оценить соответствие реального распределения модели

## Журнал выполнения и решение проблем

Для построения распределения по частотам сначала были посчитаны частоты терминов по всему корпусу после токенизации и нормализации (в том числе со стеммингом). Для получения графика далее выполнялась сортировка терминов по убыванию частоты и построение набора точек вида (rank, freq). Чтобы график был информативным на широком диапазоне частот, обе оси были переведены в логарифмическую шкалу. На практике основной сложностью является наличие “шумных” терминов (числовые последовательности, версии, технические идентификаторы, артефакты HTML), которые заметно влияют на хвост распределения и усиливают расхождение с идеальным законом.




**ТОП-10 терминов(term-count):**

1. the 1914451
2. and 1654312
3. type 1465962
4. namepart 1341016
5. to 1245211
6. of 1204498

7. for 891166
8. is 759947
9. in 683674
10. role 656863

## Исходный код

В рамках работы использовались утилиты конвейера корпуса: токенизация и подсчёт частот терминов по документам, затем формирование файла распределения для построения графика. Построение графика выполнялось по данным “терм-частота” с последующей сортировкой и преобразованием в “ранг-частота”, после чего поверх эмпирических точек накладывалась теоретическая кривая Ципфа.

## Результаты и объяснение расхождения с законом Ципфа

График распределения терминов по частотам в логарифмической шкале демонстрирует характерную для текстов картину: небольшое количество самых частых терминов образует “голову” распределения, затем следует почти линейный участок в log-log координатах, и далее - длинный “хвост” редких терминов. Для наложения закона Ципфа использовалась зависимость вида $f(r) = C / r$, где r - ранг терма, а C выбиралась как частота самого частотного терма (то есть так, чтобы кривая начиналась с первой точки распределения).

Расхождения с идеальным законом Ципфа объясняются следующими факторами: корпус является смесью разных источников и стилей текста, поэтому распределение формируется не одним “однородным” языковым процессом; в данных присутствуют шумовые токены (числа, идентификаторы, версии, служебные фрагменты HTML), которые увеличивают долю редких терминов и “утяжеляют хвост”; кроме того, стемминг и правила токенизации объединяют часть словоформ и изменяют частоты некоторых групп терминов. Также влияние оказывает конечный размер корпуса: в реальных данных многие термы встречаются 1-2 раза, что создаёт плотный хвост и отклонение от гладкой теоретической кривой.

## Выводы

В ходе лабораторной работы было построено распределение терминов по частотам в логарифмической шкале и выполнено сравнение с законом Ципфа. Распределение в целом соответствует ожидаемой форме (почти линейное поведение в log-log координатах на среднем диапазоне рангов), а основные расхождения объясняются неоднородностью корпуса, наличием шумовых токенов и влиянием предобработки текста (токенизация и стемминг), а также большим количеством редких терминов в хвосте распределения.