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 текстов на естественном языке»

Группа: М80-108М-17
Выполнил: Забарин Н.И.
Преподаватель: А.Л. Калинин

Москва, 2018

Задание

Построить график распределения терминов корпуса документов по частотностям в логарифмической шкале, наложить на этот график закон Ципфа. Объяснить причины расхождения.

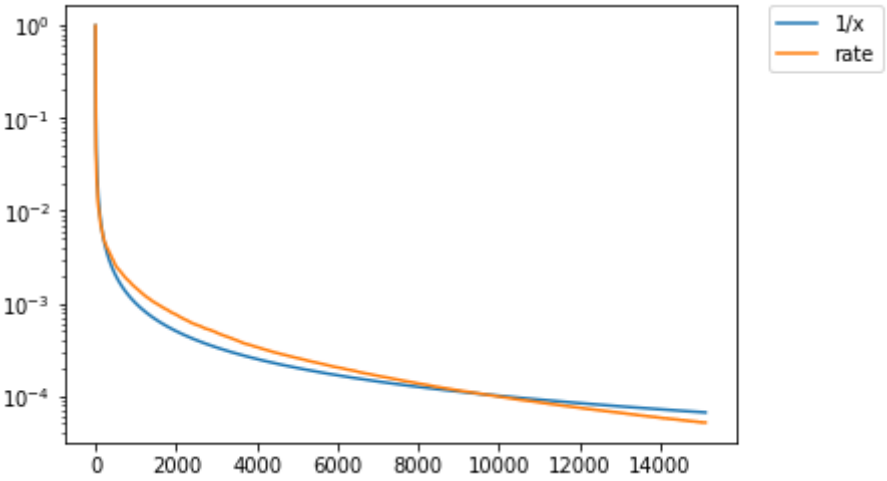
Подобрать константы для закона Мандельброта, наложить полученный график на график распределения терминов по частотностям. Привести выбранные константы.

Решение

Графики строились в Ipython с помощью библиотеки matplotlib, данные импортировались из файлов.

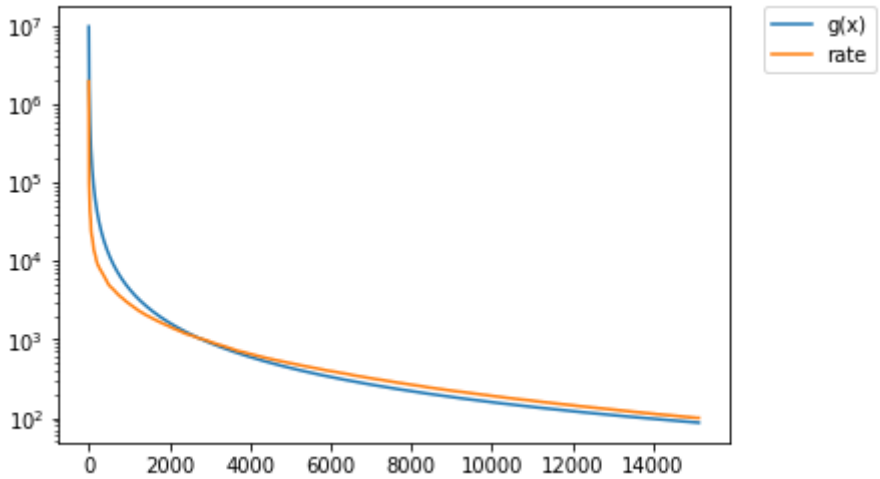
Результаты

Закон Ципфа:



Если при вычислении частот делить на максимальный rate, то получится такой график. Мы можем позволить себе такую операцию, тк закон Ципфа говорит о пропорциональности.

Закон Мандельброта, со следующими параметрами P = 10**8, q = 5, b = -1.45:



Выводы

На подобранных мной параметрах можно увидеть что закон Ципфа ошибается на низко- и среднечастотных терминах, закон Мандельброта же только на части среднечастотных.

В целом закон Мандельброта лучше аппроксимирует данную функцию.