Московский Авиационный Институт
(Научный Исследовательский Институт)

Факультет прикладной математики и физики
Кафедра вычислительной математики и программирования

Лабораторная работа №2 по курсу «Обработка текстов на естественном языке»

Выполнил: Ефименко Н.А.

Преподаватель: Калинин. А. Л.

Москва, 2020 г.

ЛР2: Закон Ципфа

Задание

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

Метод решения

1. Изучение материалов по построению графика анализа частоты терминов текста в логарифмической шкале.
2. Изучение закона Ципфа.
3. Написание и отладка кода, выполняющего построение требуемых графиков и вывод данных характерных частотах терминов (самый частотный, средней частоты и малой).

Журнал выполнения

№	Действие	Проблема	Решение
1	Сбор данных для построения графиков	Не выявлена	

Результаты выполнения

Общее количество терминов без удаления стоп слов = 113.303.

Общее количество терминов с удалением стоп слов = 84.434

«Чистая» токенизация

Токен	Частота	Позиция
the	96345	1
of	48058	2
to	38030	3
source	1989	100
p	2954	1000

С удалением стоп-слов

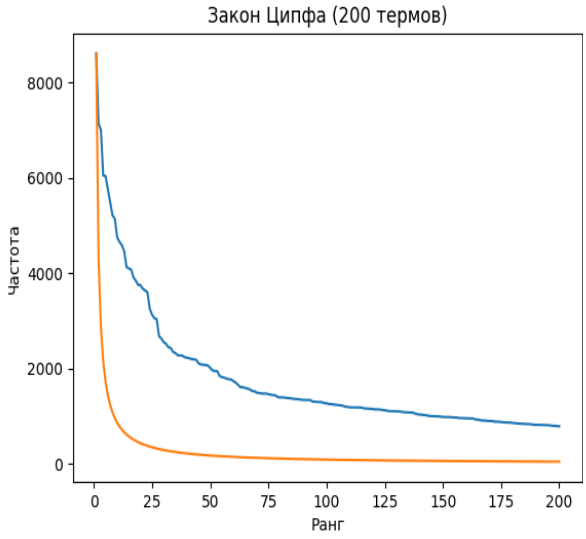
Токен	Частота	Позиция
language	9803	1
code	8113	2
programming	7808	3
end	1251	100
produced	190	1000

Графики, представленный на рис. 1, показывает зависимость частоты термина от ранга термина для корпуса документов и закон Ципфа. Зависимость близка к гиперболической.

(рыжий — закон Ципфа, синий — зависимость для корпуса)

Графики на рис. 2 – отображает закон Ципфа и зависимость частоты от ранга термина в логарифмической шкале.

Без стоп слов



со стоп словами

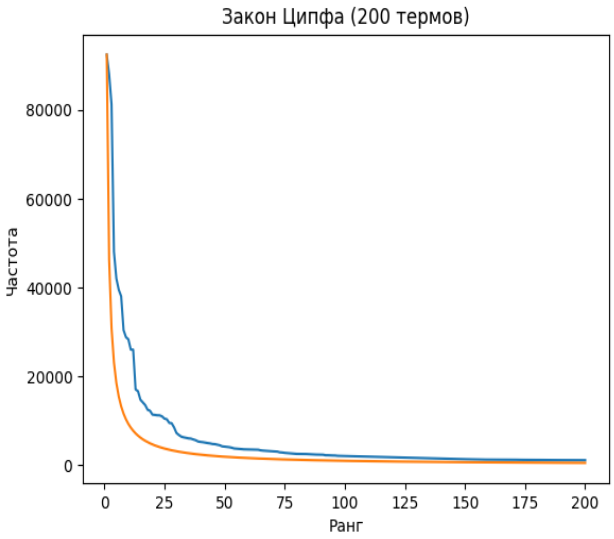
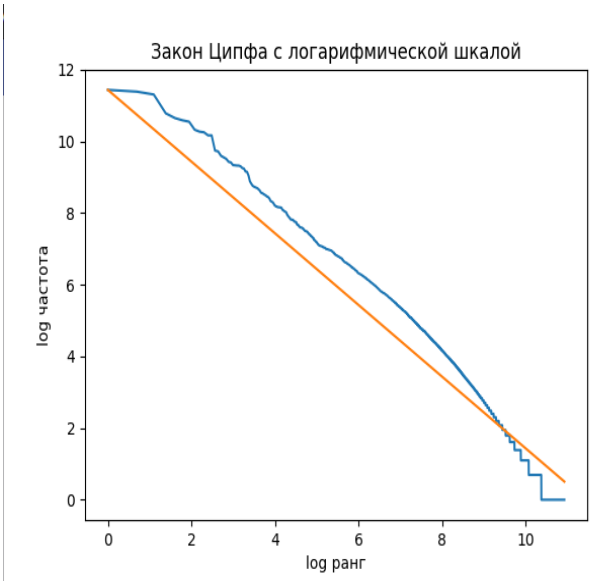


Рис. 1



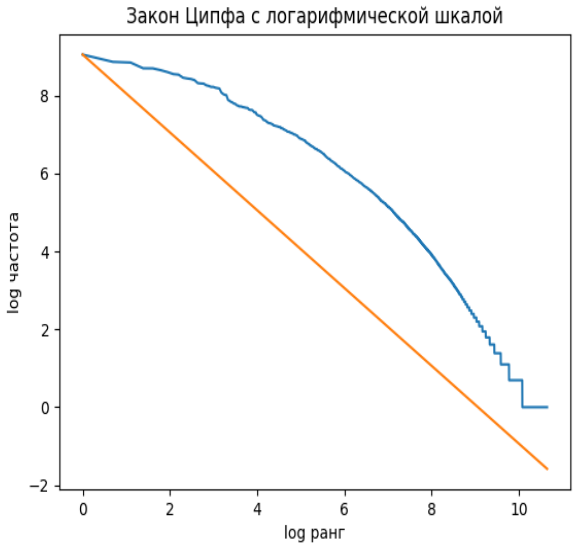


Рис. 2

Выводы

В процессе выполнения лабораторной работы были построены графики распределения терминов по частотностям и в логарифмической шкале в сравнении с законом Ципфа, можно заметить отклонения от закона в разрезе корпуса документов, которые можно объяснить стилистикой текста и наличием общей тематической линии текстов, также видно, даление стоп-слов уменьшает различие в изменения частот при переходе от большего ранга к меньшему.