Белорусский государственный университет

Факультет прикладной математики и информатики

Кафедра информационных систем управления

**Лабораторная работа №2**

**Выполнил**

Студент 3 курса 12 группы

Веренич Владислав Николаевич

**Минск 2022**

***Задание 1. Закон Ципфа***

– рассчитать показатели с построением графика

(график построен по частотам первых 500 наиболее часто встречающихся слов)

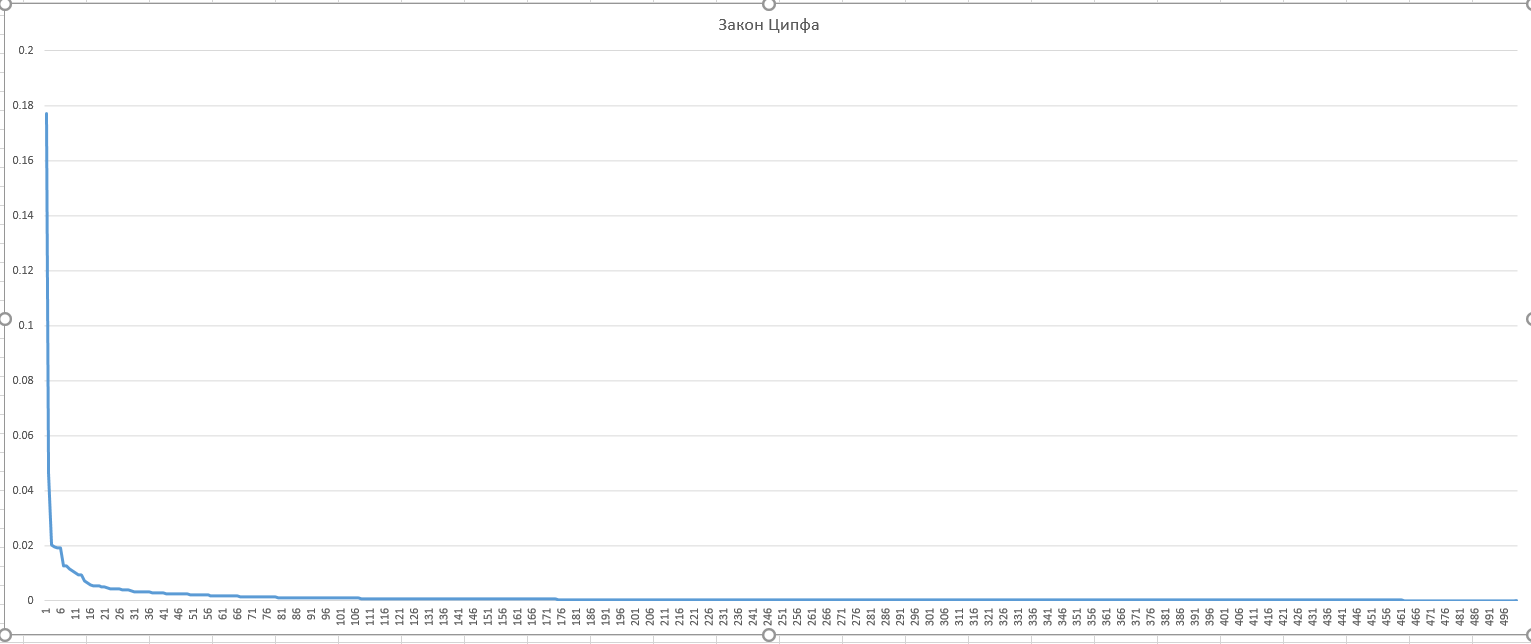


Рисунок . График частот

– проверить выполнение формулы (определить коэффициент Ципфа) и сделать вывод о выполнении закона

***f·r = c***

где f – частота встречаемости слова в тексте;

r – ранг слова в списке;

с – эмпирическая постоянная величина (коэффициент Ципфа).

Рассмотрим произвольные точки на графике:

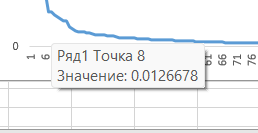


Рисунок . Точка 1 на графике частот

r = 8, f = 0.0126678

c = 0.1013424

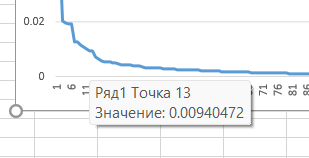


Рисунок . Точка 13 на графике частот

r = 13, f = 0.0094

c = 0.122261

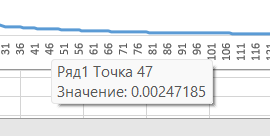


Рисунок . Точка 47 на графике частот

r = 47, f = 0.00247185

c = 0.11617

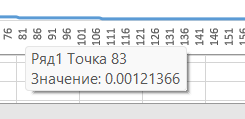


Рисунок . Точка 83 на графике частот

r = 83, f = 0.00121366

c = 0.10073

Также я построил график зависимости коэффициента Ципфа от ранга для первых 500 наиболее встречающихся слов(полученный коэффициент Ципфа равен 0.1):

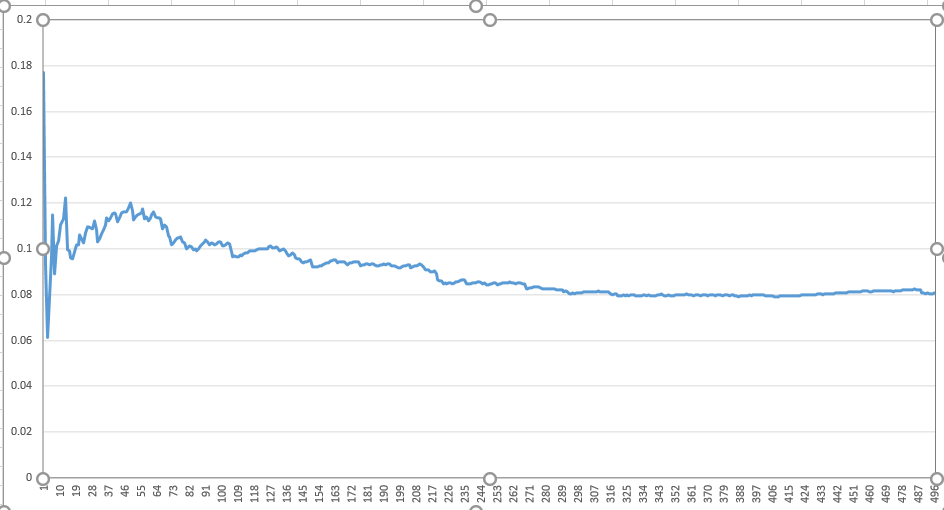


Рисунок . График зависимости коэффициента Ципфа от ранга(500 слов)

И для первых 10000(полученный коэффициент Ципфа равен 0.06):

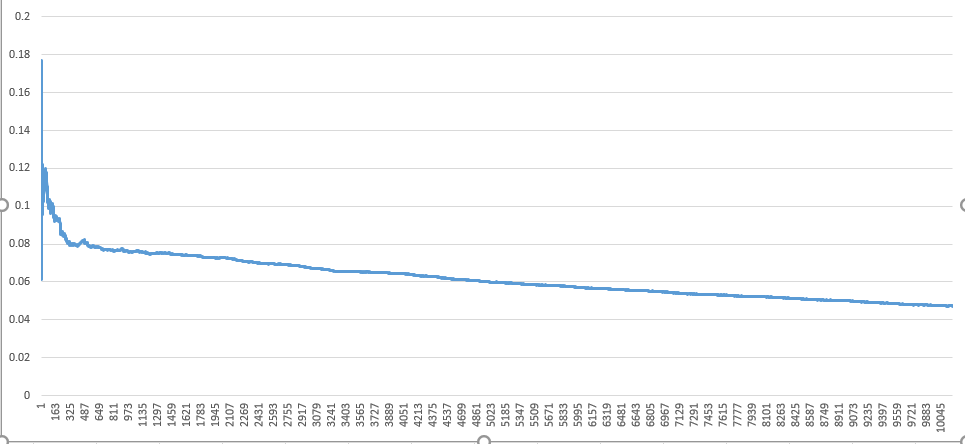


Рисунок . График зависимости коэффициента Ципфа от ранга(10000 слов)

Вывод: как видно из результатов выше, коэффициент Ципфа для первых 500 слов практически одинаков и закон Ципфа для них выполняется, однако, рассмотрев первые 10000 слов, можно увидеть, что закон Ципфа в данном случае не выполняется, так как график не представлен прямой, параллельной оси с номерами ранга, в отличие от случая для 500 первых точек.

***Задание 2. 2-й закон Ципфа***

– рассчитать константу нормирования B в соответствии с формулой для первых 10 максимальных значений частоты:

***N(f) = B / (f)^b***

где N(f) – количество различных слов, каждое из которых используется в тексте f раз,

В – константа нормирования;

b — количество слов.

Я понял задание следующим образом: пусть слово w1 и слово w2 встречаются в тексте с одинаковой частотой, тогда (f)^b – эта частота. N(f) = 2, если это единственные слова с такой частотой, соответственно, B = (f)^b \* N(f). Первые 10 максимальных значений частоты уникальны, значит, N(f) = 1 для каждого из них. Получается, что значением B, будут просто значения частот этих слов.

Рассмотрим 10 максимальных частот:

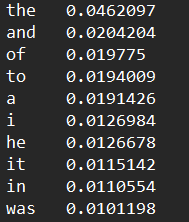


Рисунок . 10 слов с максимальным значением частоты

Среднее арифметическое частот равно 0.0183 – константа нормирования

– построить график, где Х - частота слова, Y – число слов данной частоты, проверить, будет форма кривой будет неизменной.

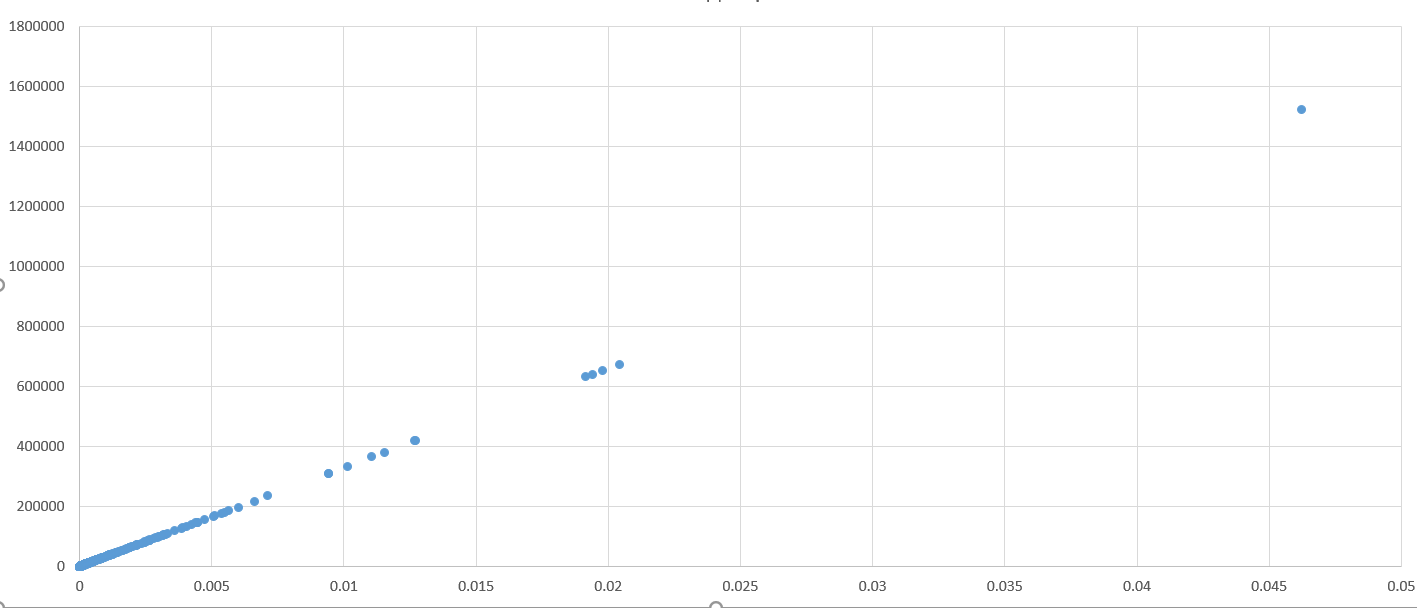


Рисунок . График зависимости числа слов данной частоты от частоты

***Задание 3. Эмпирический закон Ципфа***

– на основе списка служебных слов из созданного частотного словаря проверить выполнение эмпирического закона Ципфа: длина слова обратно пропорциональна его частоте.

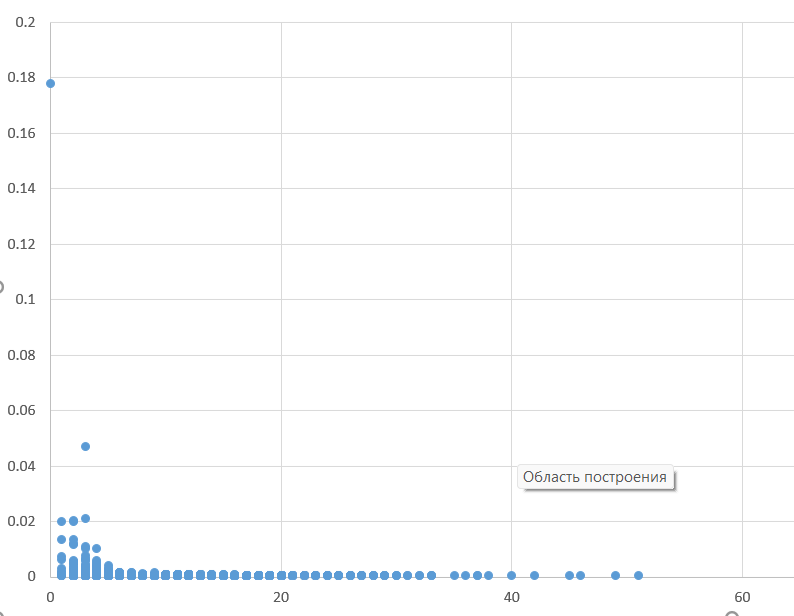


Рисунок . График зависимости частоты слова от длины слова

Вывод: как видно на графике, точки, имеющие наибольшую частоту, имеют наименьшую длину, а значит, эмпирический закон Ципфа выполняется и длина слова обратно пропорциональна его частоте.

***Задание 4. Расчет коэффициента D Жуйана***

Рассчитать коэффициент для полученного частотного словаря и результат оформить в виде таблицы (слово, относительная частота, ранг, D).

Коэффициент D отражает равномерность распределения частот в разных сегментах корпуса и вычисляется по следующей формуле:

D = 100 \* (1 – (σ / (μ \* sqrt(n - 1))))

где μ – средняя частота слова по всему корпусу,

σ – среднее квадратичное отклонение этой частоты на отдельных документах,

n – число документов, в которых встречается это слово.

Для подсчета коэффициента Жуйана корпус разбивается на n равных сегментов (например, на 4 части, размером приблизительно в 250 тыс. слов каждый).

Для выполнения этого задания я разбил исходный корпус слов на 4 части. И получил следующие значения частот:

Слова из 1-й части:

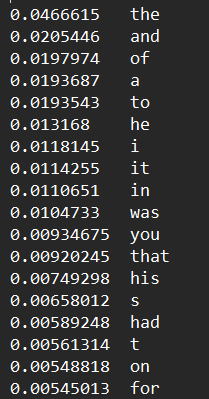


Рисунок . Слова с наибольшим значением частот из 1-го сегмента

Слова из 2-й части:

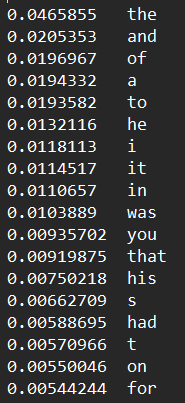


Рисунок . Слова с наибольшим значением частот из 2-го сегмента

Слова из 3-й части:

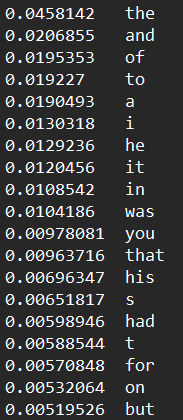


Рисунок . Слова с наибольшим значением частот из 3-го сегмента

Слова из 4-й части:

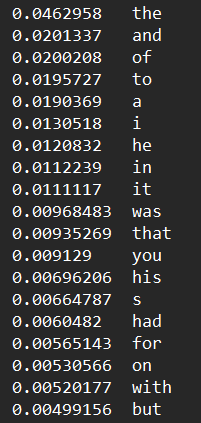


Рисунок . Слова с наибольшим значением частот из 4-го сегмента

Рассчитал коэффициент D Жуйана для первых 10 слов и записал результаты в таблицу:

Таблица . Коэффициент D Жуйана для 10 слов с наибольшей частотой

