МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ Н. Э. БАУМАНА
Факультет информатики и систем управления
Кафедра теоретической информатики и компьютерных технологий

Лабораторная работа №1
по курсу «Автоматическая обработка текстов»

«Закон Ципфа»

Выполнил:
студент группы ИУ9-11М
Беляев А. В.

Проверила:
Лукашевич Н. В.

Москва 2018

1 Цель работы

Проверить выполнение закона Ципфа для художественного произведения

2 Текст программы

```
import re
import pymorphy2
import matplotlib
import numpy as np
from collections import Counter

matplotlib.use('TkAgg') # macos fix

morph = pymorphy2.MorphAnalyzer()


def draw_plot(x_values, y_values):
    from matplotlib import pyplot as plt
    plt.plot(np.array(x_values), np.array(y_values))
    plt.show()


def main():
    with open("a_dance_with_dragons.txt", encoding="utf-8", mode="r") as f:
        contents = f.read().lower()

        clean_text = re.sub("[^a-z0-9]", " ", contents)
        splitted = re.compile("\s+").split(clean_text)
        normalized = list(map(lambda word: morph.parse(word)[0].normal_form
            , splitted))
        sorted_by_rank = Counter(normalized).most_common()

        with open('out.txt', 'w+') as out:
            for word, count in sorted_by_rank[:1000]:
                print(f'{word}\t{count}', file=out)

        frequencies_list = list(map(lambda word_count_tuple:
            word_count_tuple[1], sorted_by_rank))
        draw_plot(list(range(1, len(frequencies_list) + 1)),
            frequencies_list)
```

Листинг 1: Исходный код программы

3 Результаты тестирования

В качестве художественного произведения был взят роман «Танец с драконами» из цикла «Песнь Льда и Пламени» Джорджа Р. Р. Мартина. В результате тестирования был построен график распределения частоты использования слов. Как видно из графика на Рис. 1, закон Ципфа выполняется и частота встречаемости слова обратно пропорциональна рангу слова.

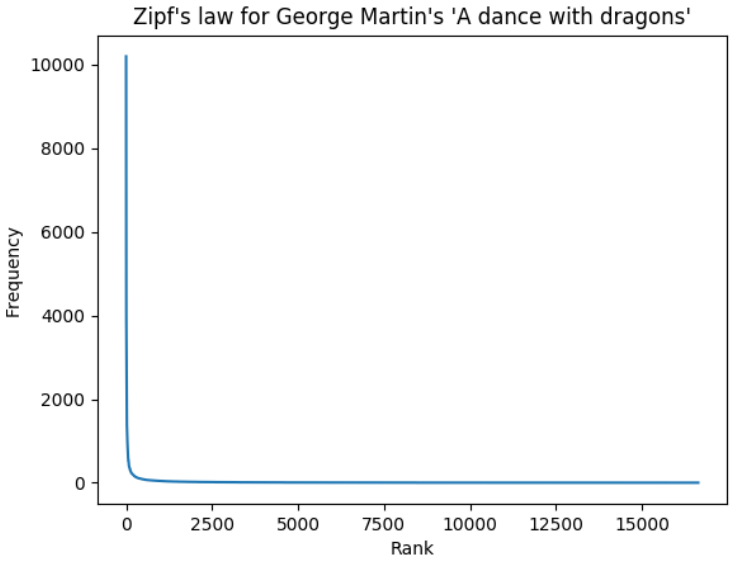

Рис. 1:

Таблица 1: Частоты некоторых слов с рангом от 30 до 50

слово	частота
один	1169
лорд	1049
человек	1021
так	973
сказать	956
джон	871
рука	835
тирион	696
кто	680
только	677
большой	606
королева	593

Частоты некоторых слов в ранге от 30 до 50 представлены в Таблице 1

4 Выводы

В ходе работы было проверено выполнение закона Ципфа для романа «Танец с драконами». Тестирование показало, что закон выполняется.