**Белорусский государственный университет**

**Факультет прикладной математики и информатики**

**Кафедра информационных систем управления**

**Расчетно-графическая работа №2**

**Выполнил**

студент 3 курса 11 группы

Яшенок Алина Игоревна

Минск, 2024

**Содержание**

[Задание 1. Закон Ципфа 3](#_Toc181211184)

[Задание 2. Эмпирический закон Ципфа 8](#_Toc181211185)

[Задание 3. Расчет коэффициента D Жуйана 11](#_Toc181211186)

# **ЗАДАНИЕ 1. ЗАКОН ЦИПФА**

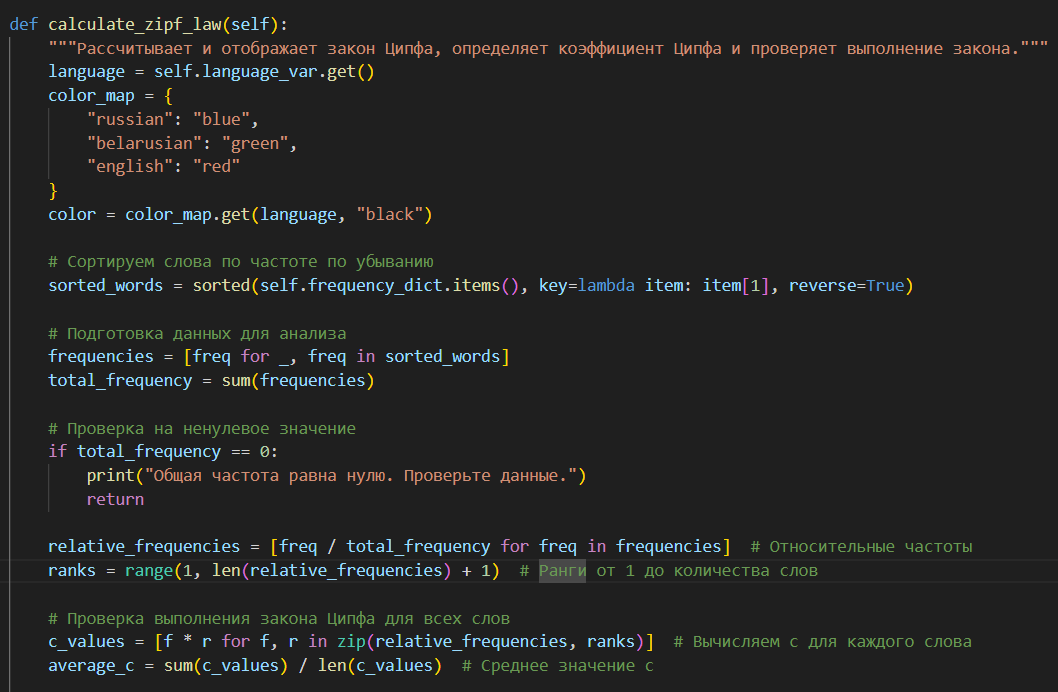
В ходе выполнения задания нужно было рассчитать показатели с построением графика, проверить выполнение формулы (определить коэффициент Ципфа) и сделать вывод о выполнении закона

где - частота встречаемости слова в тексте;

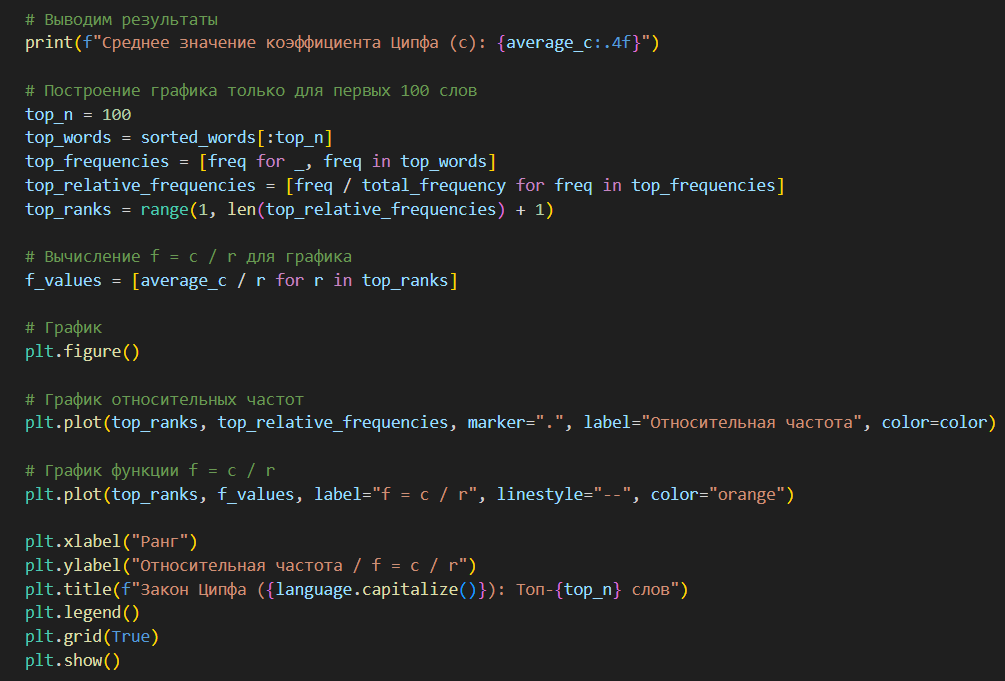
- ранг слова в списке;

- эмпирическая постоянная величина (коэффициент Ципфа).

В рамках выполнения задания был разработан и реализован метод calculate\_zipf\_law на языке Python (Рисунок 1-Рисунок 2) для расчета показателей, построения графика и проверки выполнения закона Ципфа.



**Рисунок 1 - метод calculate\_zipf\_law**



**Рисунок 2 - метод calculate\_zipf\_law (продолжение)**

Метод выполняет следующие шаги:

1. Сбор данных:

* Словарь частоты слов (frequency\_dict) был отсортирован по убыванию частоты для анализа.
* Извлечены относительные частоты слов, вычисленные как отношение частоты слова к общей частоте всех слов.

1. Проверка на ненулевое значение:

* Перед проведением расчетов проверяется, что общая частота не равна нулю. В случае нулевой частоты выводится сообщение об ошибке.

1. Вычисление коэффициента Ципфа:

* Для каждого слова вычисляется значение 𝑐 как произведение относительной частоты и ранга.
* Среднее значение коэффициента Ципфа (𝑐) рассчитывается по всем словам, что позволяет оценить его стабильность.

1. Построение графика:

* Для наглядности только для первых 100 слов строится график, где по оси X откладывается ранг, а по оси Y — относительная частота и функция 𝑓 =𝑐/𝑟.
* На графике представлены две линии:
* Относительная частота слов.
* Значение функции 𝑓=𝑐/𝑟 в виде пунктирной линии.

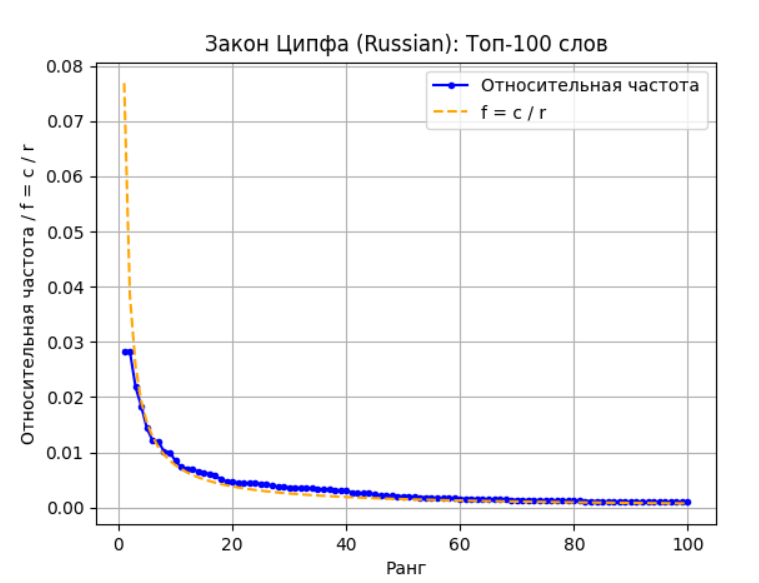
1. Выводы:

* Среднее значение коэффициента Ципфа (𝑐) выводится в консоль.
* График визуализирует, насколько хорошо данные соответствуют закону Ципфа. Если относительная частота близка к линии 𝑓=𝑐/𝑟, это подтверждает выполнение закона.

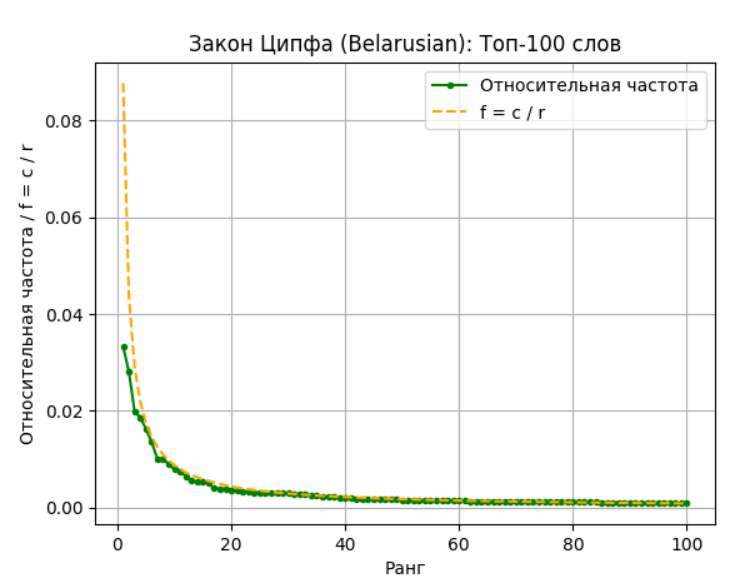
В результате выполнения метода для словарей на 3 различных языках были получены следующие данные:

* Среднее значение коэффициента Ципфа (𝑐) для русского языка - 0.0769
* Среднее значение коэффициента Ципфа (𝑐) для белорусского языка - 0.0878
* Среднее значение коэффициента Ципфа (𝑐) для английского языка - 0.0375

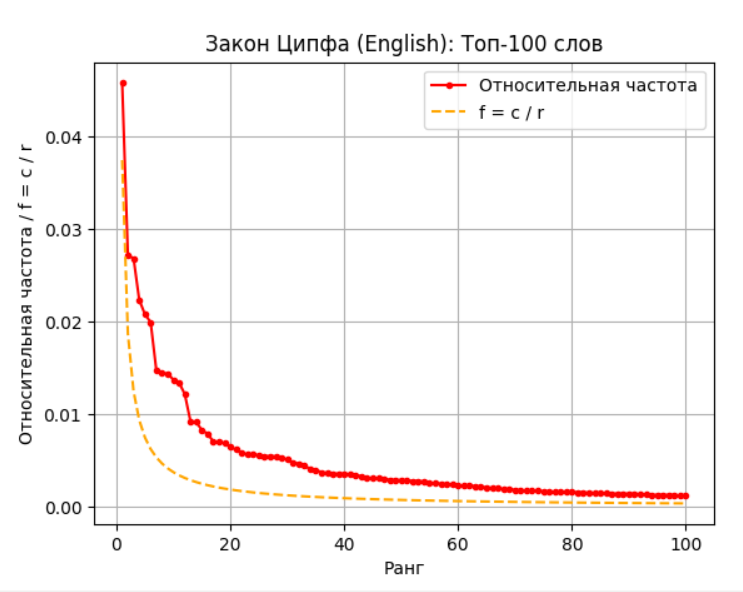
Также были построены графики для русского (Рисунок 3), белорусского (Рисунок 4) и английского (Рисунок 5) языков.



**Рисунок 3 - график для закона Ципфа (русский)**



**Рисунок 4 - график для закона Ципфа (белорусский)**



**Рисунок 5 - график для закона Ципфа (английский)**

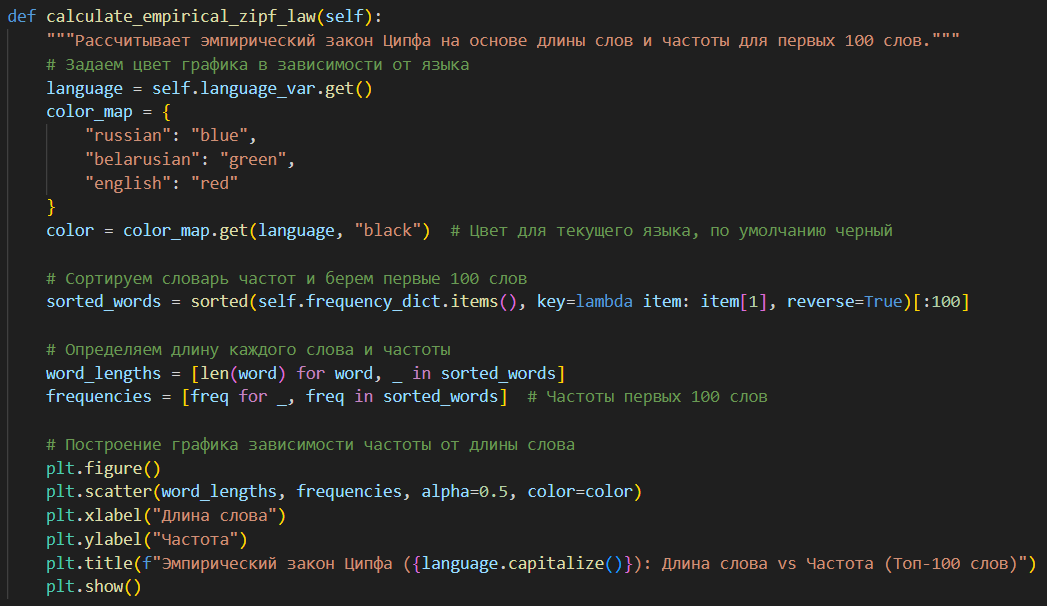
Для русского и белорусского языков графики относительных частот слов демонстрируют высокую степень соответствия ожидаемой линии функции 𝑓=𝑐/𝑟. Это указывает на то, что распределение частот слов в этих языках хорошо согласуется с законом Ципфа. Наблюдаемая близость кривых подтверждает, что небольшое количество слов используется очень часто, в то время как большая часть слов встречается значительно реже.

В начальной версии графиков для английского языка наблюдается небольшое расхождение между графиком относительных частот и линией функции 𝑓=𝑐/𝑟. Это может быть связано с особенностями английского языка, включая наличие большого количества синонимов и разнообразие стилей текста. Однако, при анализе графика для всех слов, это расхождение уменьшается, и кривая становится ближе к ожидаемой функции. Это свидетельствует о том, что, несмотря на индивидуальные вариации, распределение частот в английском языке также в целом соответствует закону Ципфа, особенно при учете более обширного корпуса текста.

# **ЗАДАНИЕ 2. ЭМПИРИЧЕСКИЙ ЗАКОН ЦИПФА**

В рамках второго задания на основе списка служебных слов из созданного частотного словаря нужно было проверить выполнение эмпирического закона Ципфа: длина слова обратно пропорциональна его частоте. Для анализа были использованы первые 100 слов из отсортированного по убыванию частоты частотного словаря, так как именно эта часть охватывает большую долю служебных слов.

В рамках выполнения задания был разработан и реализован метод calculate\_empirical \_zipf\_law на языке Python (Рисунок 6)



**Рисунок 6 – метод calculate\_empirical \_zipf\_law**

Метод выполняет следующие шаги:

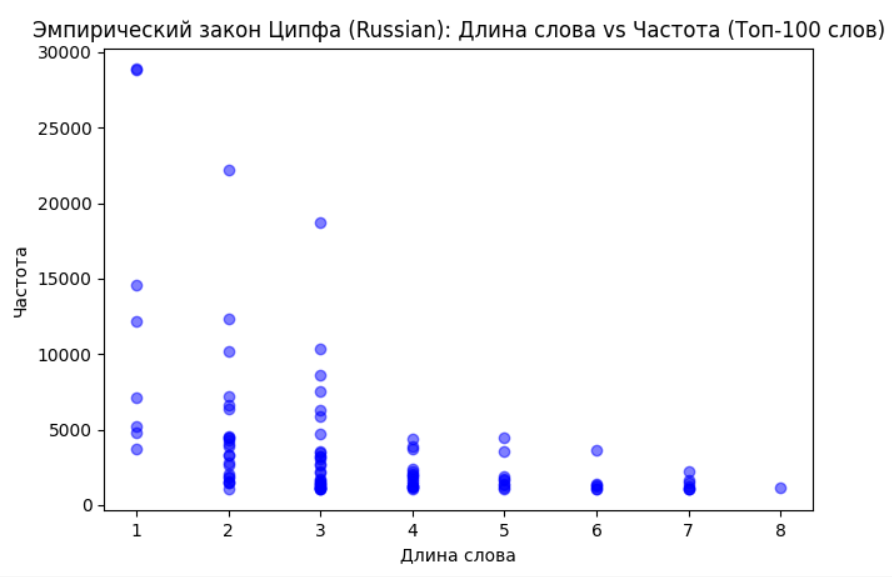
1. Сбор данных:

* Частотный словарь был отсортирован, и выбраны первые 100 слов, что позволяет сосредоточиться на наиболее употребляемых словах в языке.
* Для каждого из выбранных слов была рассчитана его длина и частота.

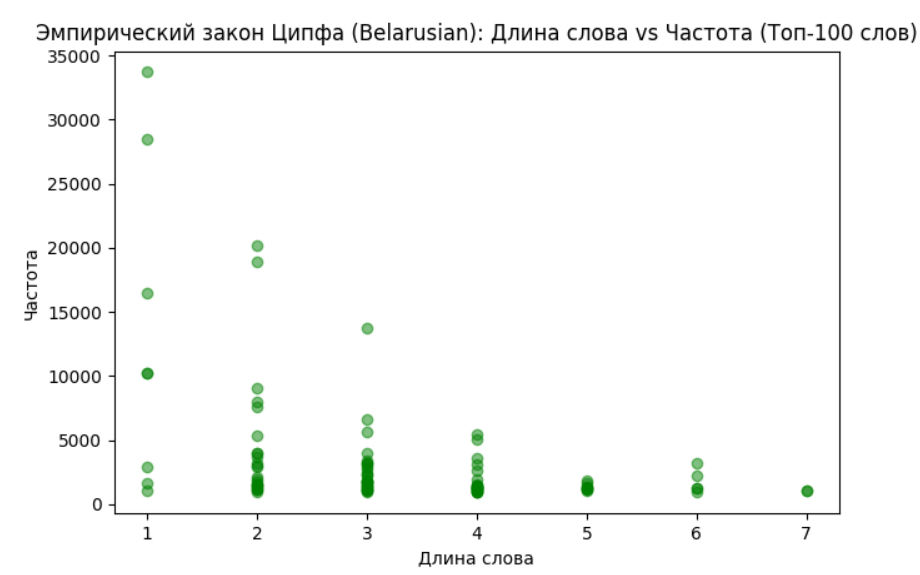
1. Построение графика:

* На графике была представлена зависимость частоты слов от их длины, где по оси X откладывается длина слова, а по оси Y — его частота.

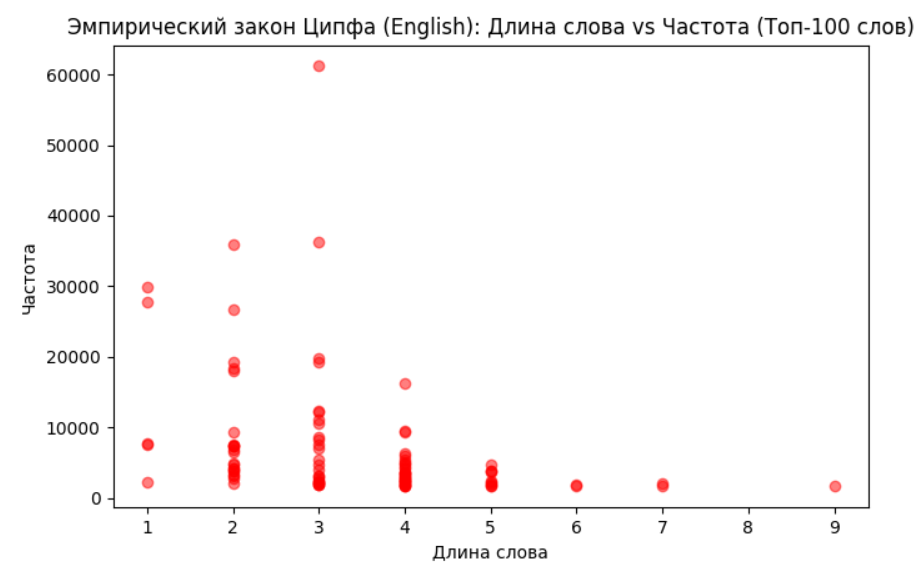
График был построен для трех языков: русского (Рисунок 7), белорусского (Рисунок 8) и английского (Рисунок 9).



**Рисунок 7 – график эмпирического закона Ципфа (русский)**



**Рисунок 8 - график эмпирического закона Ципфа (белорусский)**



**Рисунок 9 - график эмпирического закона Ципфа (английский)**

Для большинства слов наблюдается обратная зависимость: чем больше длина слова, тем ниже его частота. Это подтверждает гипотезу о том, что короткие слова, как правило, встречаются чаще, чем длинные. Первые 100 слов, выбранные для анализа, в значительной степени состоят из служебных слов (предлогов, союзов и т.д.), которые, как правило, являются короткими. Это объясняет, почему на графике видно множество коротких слов с высокой частотой. В русском и белорусском языках можно заметить более выраженную обратную зависимость, чем в английском, где присутствие более длинных слов связано с особенностями языка и его синтаксиса.

Анализ графика зависимости частоты от длины слова подтвердил выполнение эмпирического закона Ципфа для первых 100 слов в языках. Наблюдаемая тенденция, согласно которой длина слова обратно пропорциональна его частоте, согласуется с ожидаемыми результатами, и демонстрирует, что короткие служебные слова используются значительно чаще, чем длинные.

Использование первых 100 слов из частотного словаря является обоснованным, так как именно эта часть охватывает большинство служебных слов, представляющих собой ключевые элементы языка.

# **ЗАДАНИЕ 3. РАСЧЕТ КОЭФФИЦИЕНТА D ЖУЙАНА**

В этом задании нужно было рассчитать коэффициент D Жуйана для полученных частотных словарей. Коэффициент D отражает равномерность распределения частот слов в разных сегментах корпуса и рассчитывается по следующей формуле:

,

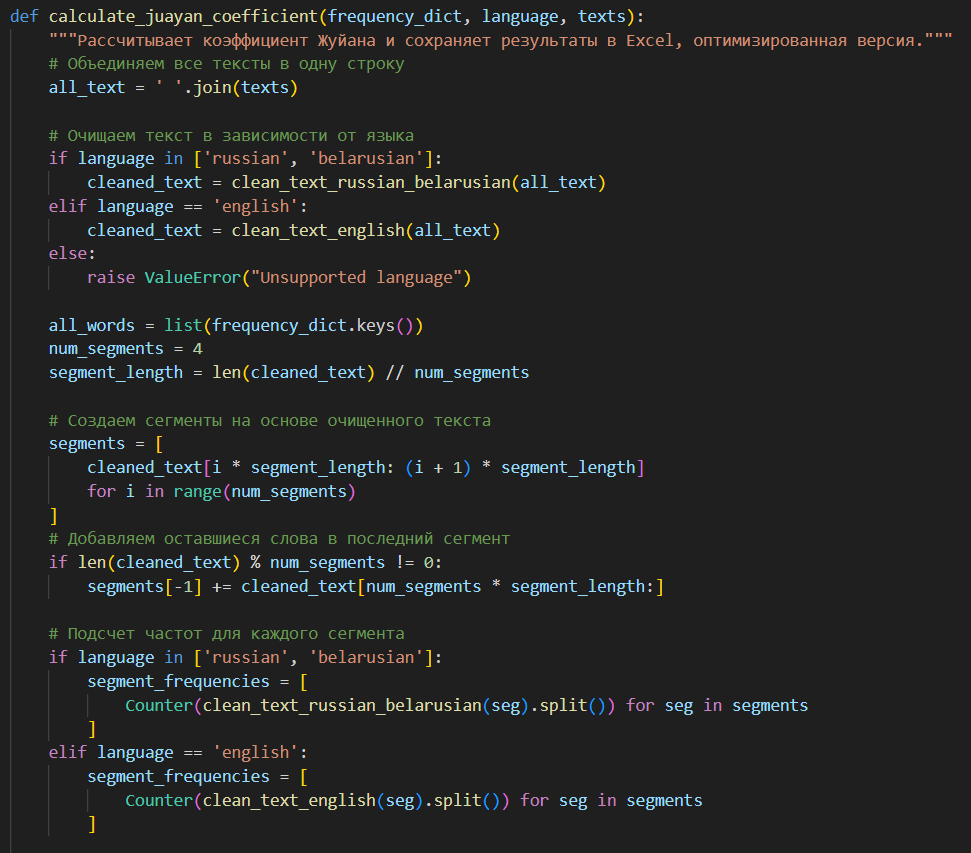
где - средняя частота слова по всему корпусу,

- среднее квадратичное отклонение этой частоты на отдельных документах,

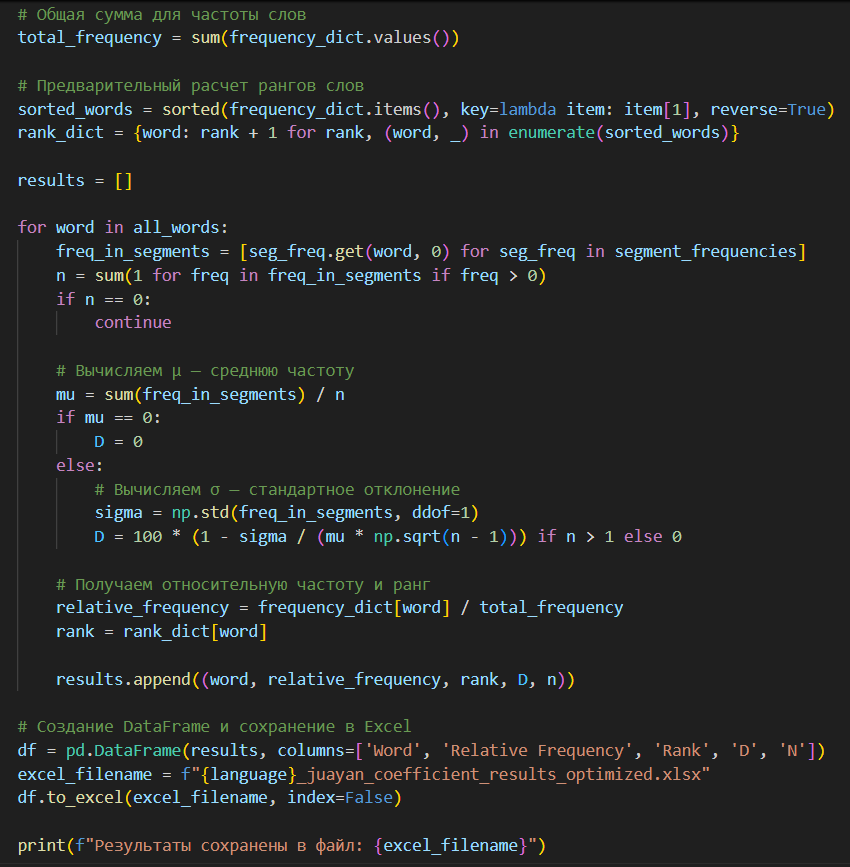
n - число документов, в которых встречается это слово.

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

В рамках выполнения задания был разработан и реализован метод calculate\_juayan\_coefficient на языке Python (Рисунок 10-Рисунок 11), принимающий в качестве параметров частотный словарь, язык и набор текстов.



**Рисунок 10 - calculate\_juayan\_coefficient**



**Рисунок 11 - calculate\_juayan\_coefficient (продолжение)**

Метод выполняет следующие шаги:

1. Сбор данных:

* Все тексты были объединены в одну строку, а затем очищены в зависимости от языка (русский, белорусский или английский).
* Корпус текста был разбит на 4 равных сегмента, что позволяет оценить распределение слов в разных частях текста.

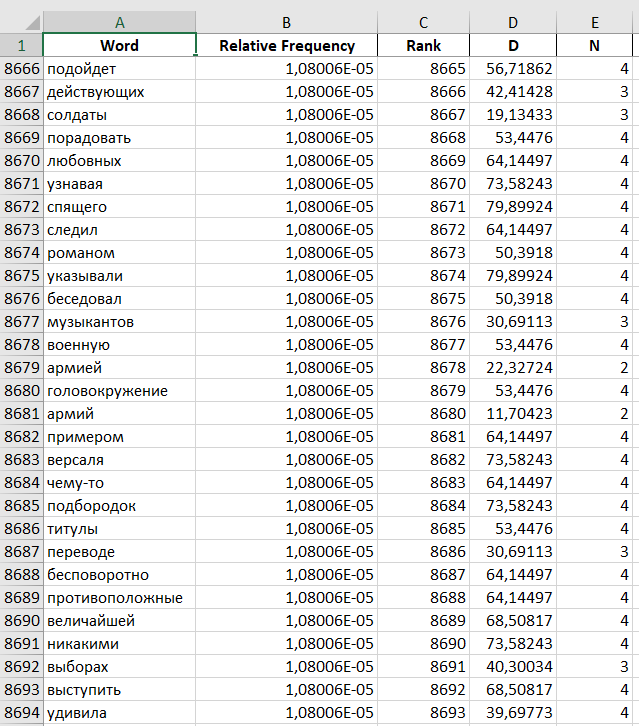
1. Подсчет частот:

* Для каждого сегмента текста были подсчитаны частоты слов с использованием класса Counter.
* На основе общего частотного словаря были рассчитаны относительные частоты, ранги, а также коэффициент D для каждого слова.

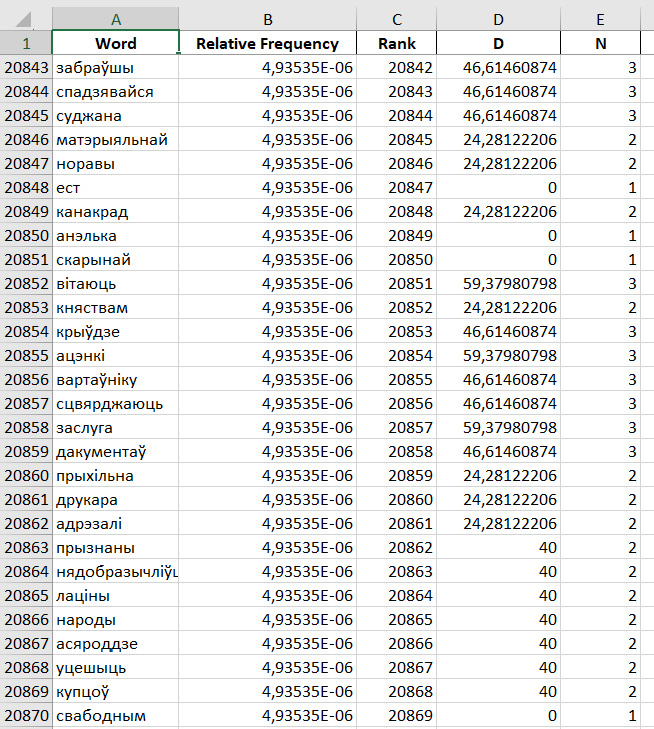
1. Результаты:

* Результаты расчетов были оформлены в виде таблицы, содержащей следующие колонки:
* Word: само слово из частотного словаря.
* Relative Frequency: частота слова относительно общей суммы частот.
* Rank: ранг слова в соответствии с его частотой.
* D: коэффициент Жуйана, отражающий равномерность распределения.
* N: число сегментов, в которых встречается данное слово.

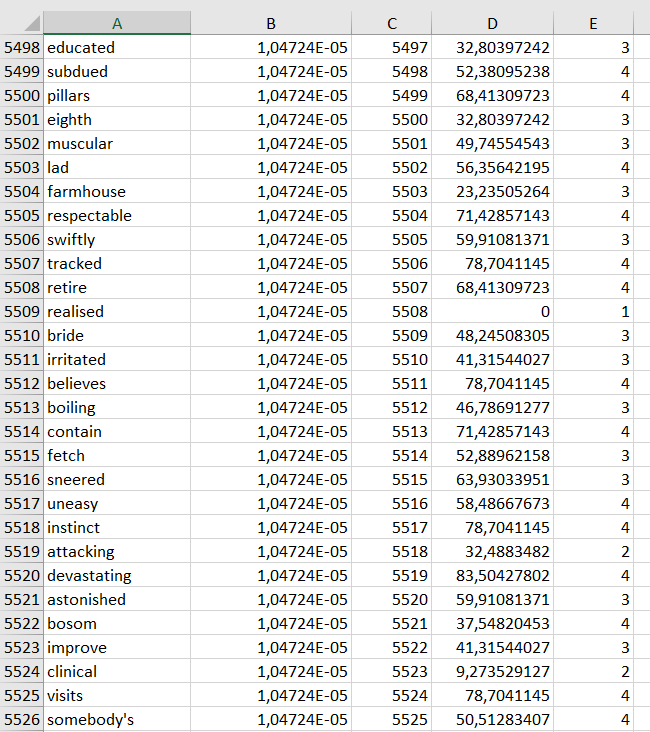
В результате выполнения программы били получены Excel таблицы для русского (Рисунок 12), белорусского (Рисунок 13) и английского (Рисунок 14) текстов. Данные в таблицах отсортированы по рангу.



**Рисунок 12 – фрагмент excel таблицы с коэффициентом Жуйана (русский)**



**Рисунок 13 - фрагмент excel таблицы с коэффициентом Жуйана (белорусский)**



**Рисунок 14 - фрагмент excel таблицы с коэффициентом Жуйана (английский)**

Значения коэффициента D варьировались в зависимости от слова. Высокие значения D (ближе к 100) указывают на более равномерное распределение частоты слова по сегментам, в то время как низкие значения (ближе к 0) свидетельствуют о неравномерности.