**Московский авиационный институт  
(национальный исследовательский университет)**

Факультет прикладной математики и физики  
Кафедра вычислительной математики и программирования

**Лабораторная работа №2  
по курсу «Обработка естественно-языковых текстов»**

Законы Ципфа и Мандельброта

|  |  |  |
| --- | --- | --- |
|  | **Студент:** | Рожлейс И.А. |
| **Группа:** | М8О-108М-18 |
| **Преподаватель:** | Калинин А.Л. |
| **Дата:** | 19.05.2019 |

Москва, 2019

**Цель работы**

Для своего корпуса необходимо построить график распределения терминов 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.

**Ход работы**

Для подсчёта частоты встречаемости токенов была написана отдельная программа на языке программирования Си++, которая парсит входной XML файл, содержащий токены корпуса документов. На выходе программа формирует текстовый файл с расширением txt в котором через пробел указаны в порядке убывания частоты токенов. Токены в выходном файле не указаны.

Внутри программы токены и их частоты хранятся в стандартном контейнере std::map, ключом которого является контрольная сумма байтов токена, а значением его частота. В качестве функции контрольной суммы используется CRC-32.



После успешного подсчёта частотностей токенов происходит их запись в файл, предварительно преобразуя контейнер std::map в последовательный контейнер std::vector сортируя частоты на убывание. Токены в файл не выводятся, так как задача предполагает только наличие статистики. Кроме того, в связи с использованием контрольной суммы вместо строкового представления токена эта информация теряется в процессе вычисления.



Закон Ципфа описывает закономерность частоты использования слов естественного языка. Так закон гласит что в любом достаточно объёмном тексте самое употребляемое слово будет встречаться примерно в два раза чаще чем втрое, в три раза чаще чем третье и т.д. – т.е. обратно пропорционально его рангу.

В общем случае закон Ципфа можно записать формулой:

где r – ранг слова, а p – частота самого употребляемого слова.

Для построения графиков был составлен скрипт в математическом пакете MATLAB на вход которого подаётся ранее сформированный файл с частотами.



Скрипт открывает диалоговое окно с предложением указать путь к файлу частот. Если файл указан верно, то происходит считывание значений из файла и отображение их на графике. Параллельно рисуется идеальный график (1/r) и находятся константы Мандельброта.

Корпус документов состоит примерно из 1,2 млн. токенов. Самый частотный токен встречается примерно 60,5 тыс. раз.

|  |  |
| --- | --- |
| C:\Users\sven4500\AppData\Local\Microsoft\Windows\INetCache\Content.Word\linear2.png | C:\Users\sven4500\AppData\Local\Microsoft\Windows\INetCache\Content.Word\log2.png |
| **Рис.1.** Закон Ципфа в линейных координатах. | **Рис.2.** Закон Ципфа в логарифмических координатах. |

**Заключение**

На полученных графиках видно, что корпус документов удовлетворяет закону Ципфа не полностью. Среди наиболее вероятных причин такого расхождения можно отметить:

1. наличие большого количество склонений слов в русском языке. Таким образом итоговый график оказывается «смазан» по горизонтальной оси;
2. наличие слов на английском языке.

Коэффициенты Мандельброта: p = 60446 и b = 0.3.

**Источники**

[1] Филиппович А.Ю. Старые записи и рассуждения о законе Ципфа-Мандельброта [Электронный ресурс] // URL: <http://blogs.it-claim.ru/andrey/2013/03/24/zipf/>