Московский Авиационный Институт
(Национальный исследовательский университет)

Институт №8 «Информационные технологии и прикладная математика»
Кафедра вычислительной математики и программирования

Лабораторная работа №2

по курсу «Обработка естественно-языковых текстов»

Студент: Зайцев Н.В.
группа М8О-208М-20

Преподаватель: Кухтичев А.А.

Москва, 2021

Лабораторная работа № 2

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

Ход работы

Закон Ципфа – закономерность распределения частоты слов естественного языка: если все слова языка упорядочить по убыванию частоты их использования, то частота n-го слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру n.

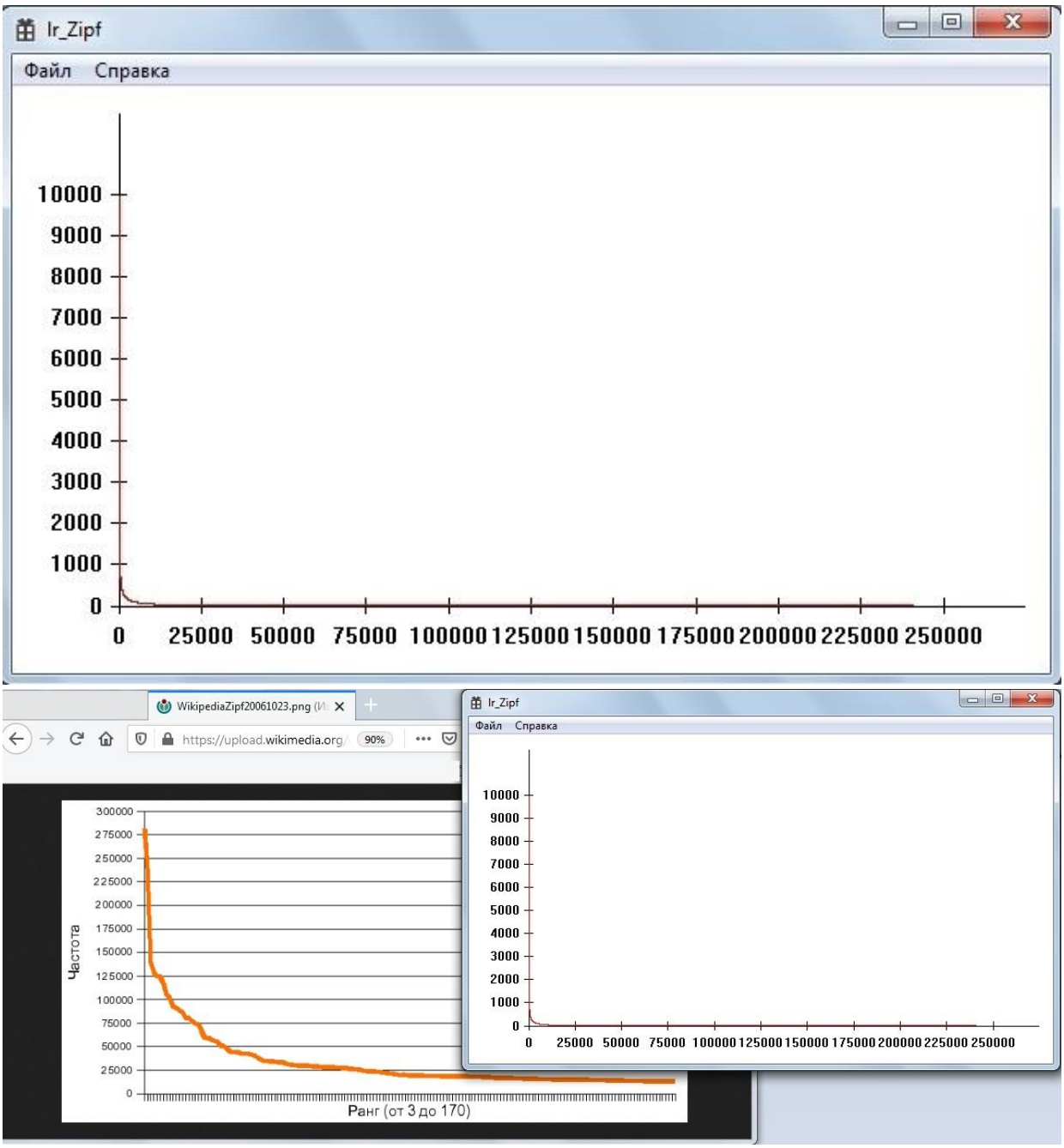
В работе использовались результаты лабораторной работы №3 по курсу «Информационный поиск». Были использованы WinAPI для перекодирования файлов. Поскольку основная часть текста статей на русском, используются только русские слова, а прочие символы – цифры, латиница, греческий и символы кодировки – считаются разделителями. Для удобства анализа исходная кодировка UTF-8 переводится в UCS-2 и вводится функция понижения регистра и определения символа как кириллицы в соответствии с таблицей Юникод-символов. Перекодировка описана в файлах Encoding.h и Encoding.cpp с использованием обвертки над функциями WinAPI. В файлах FS.h и FS.cpp прописаны рукописные обвертки над функциями WinAPI, а именно флаги поиска, поиск файлов в папке, объединение путей, абсолютный путь, существование файла или папки, чтение файла, вывод информации, а также класс файла для чтения или записи. В файлах Exception.h и Exception.cpp прописаны все возможные исключения и ошибки с их кодами.

В файле Utils.h модифицируются алгоритмы, активно используемые в STL, под данную задачу – например, функции swap, fill.

В файлах UserVector.h и UserString.h создаются рукописные контейнеры вектора и строки.

В файле Global.h прописанные шаблоны вектора и строки сравниваются с STL-контейнерами.

В файле Resourse.h описаны все используемые строки. В файлах Chart.h и Chart.cpp описан класс графика, загрузка данных для его построения и его прорисовка; здесь же токены упорядочиваются по частотности появления в тексте. Файлы lr_Zipf.h и lr_Zipf.cpp – это файлы, где создается окно графика и обрабатываются сообщения. Кроме того, в lr_Zipf.cpp указан файл index.binary, из которого берутся данные о токенах для создания графика.



При сравнении с идеальным графиком закона Ципфа из Википедии видно, что есть отклонения. Это можно объяснить наличием слов на английском языке, а также большим количеством форм слов (склонением или спряжением).

Вывод

В ходе выполнения лабораторной работы был построен график распределения терминов по частотностям, а также проведено его сравнение с графиком закона Ципфа.