**Московский Авиационный Институт**

**(Национальный исследовательский университет)**

Институт №8 «Информационные технологии и прикладная математика»

Кафедра вычислительной математики и программирования

**Лабораторная работа №1**

**по курсу «Обработка естественно-языковых текстов»**

|  |
| --- |
| Студент: Зайцев Н.В.  группа М8О-208М-20  Преподаватель: Кухтичев А.А. |
|
|
|
|
|
|
|
|
|

Москва, 2021

# Лабораторная работа № 1

Нужно реализовать процесс разбиения текстов документов на токены, который потом будет использоваться при индексации. Для этого потребуется выработать правила, по которым текст делится на токены. Необходимо описать их в отчёте, указать достоинства и недостатки выбранного метода. Привести примеры токенов, которые были выделены неудачно, объяснить, как можно было бы поправить правила, чтобы исправить найденные проблемы.

**Ход работы**

В работе были использованы WinAPI для перекодирования файлов. Папка с документами располагается на два уровня выше файлов программы (по умолчанию в папке docs).

В файлах Tokenizing.h и Tokenizing.cpp происходит разбиение текста на токены. 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.

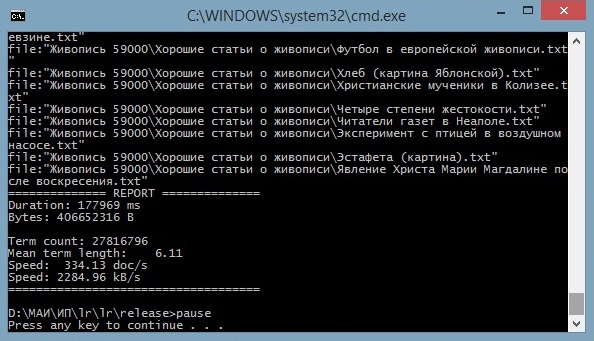
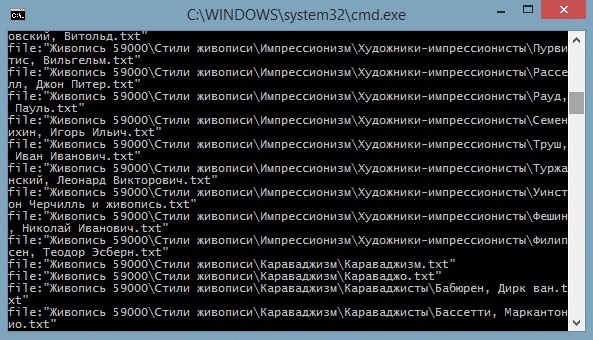
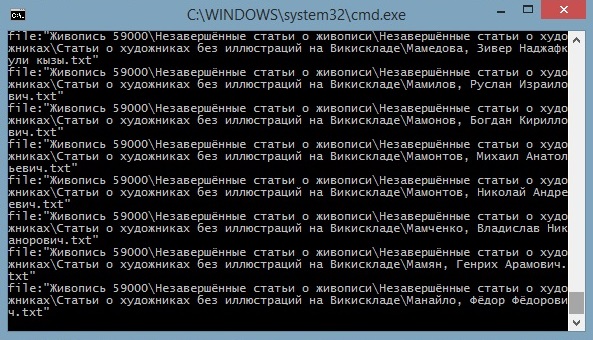
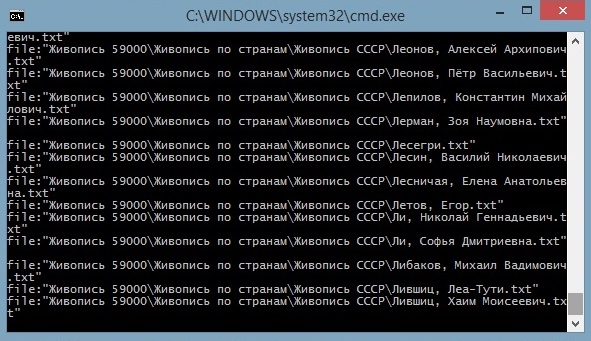
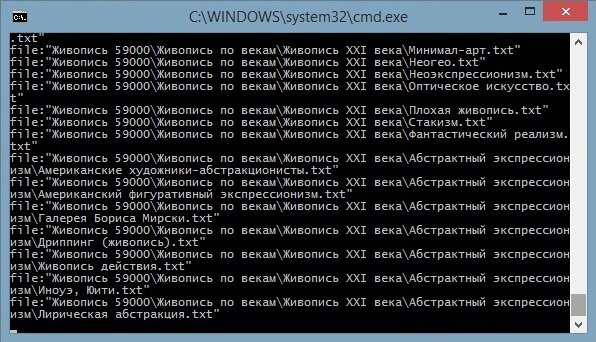
В файле Utils.h модифицируются алгоритмы, активно используемые в STL, под данную задачу – например, функции swap, fill.

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

В 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 Global.h описана структура Location, благодаря которой вся информация по файлу хранится в одном месте – номер файла, номер токена в файле, номер строки и номер символа в строке.

Программа запускается через файл tokenize.bat.



**Статистическая информация**

Продолжительность работы – 177,97 секунд.

Средняя длина токена – 6,11.

Средняя скорость работы – 2284,96 Кб/с.

**Вывод**

В ходе выполнения лабораторной работы была разработана программа, производящая токенизацию корпуса документов с их перекодировкой.