

**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 обертки 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 обертки 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.

```
C:\WINDOWS\system32\cmd.exe

.txt"
file:"Живопись 59000\Живопись по векам\Живопись XXI века\Минимал-арт.txt"
file:"Живопись 59000\Живопись по векам\Живопись XXI века\Неогео.txt"
file:"Живопись 59000\Живопись по векам\Живопись XXI века\Неоэкспрессионизм.txt"
file:"Живопись 59000\Живопись по векам\Живопись XXI века\Оптическое искусство.tx
t"
file:"Живопись 59000\Живопись по векам\Живопись XXI века\Плохая живопись.txt"
file:"Живопись 59000\Живопись по векам\Живопись XXI века\Стакизм.txt"
file:"Живопись 59000\Живопись по векам\Живопись XXI века\Фантастический реализм.
txt"
file:"Живопись 59000\Живопись по векам\Живопись XXI века\Абстрактный экспрессион
изм\Американские художники-абстракционисты.txt"
file:"Живопись 59000\Живопись по векам\Живопись XXI века\Абстрактный экспрессион
изм\Американский фигуративный экспрессионизм.txt"
file:"Живопись 59000\Живопись по векам\Живопись XXI века\Абстрактный экспрессион
изм\Галерея Бориса Мирски.txt"
file:"Живопись 59000\Живопись по векам\Живопись XXI века\Абстрактный экспрессион
изм\Дриппинг (живопись).txt"
file:"Живопись 59000\Живопись по векам\Живопись XXI века\Абстрактный экспрессион
изм\Живопись действия.txt"
file:"Живопись 59000\Живопись по векам\Живопись XXI века\Абстрактный экспрессион
изм\Иноуэ, Юити.txt"
file:"Живопись 59000\Живопись по векам\Живопись XXI века\Абстрактный экспрессион
изм\Лирическая абстракция.txt"
```

```
C:\WINDOWS\system32\cmd.exe

евич.txt"
file:"Живопись 59000\Живопись по странам\Живопись СССР\Леонов, Алексей Архипович
.txt"
file:"Живопись 59000\Живопись по странам\Живопись СССР\Леонов, Пётр Васильевич.t
xt"
file:"Живопись 59000\Живопись по странам\Живопись СССР\Лепилов, Константин Михай
лович.txt"
file:"Живопись 59000\Живопись по странам\Живопись СССР\Лерман, Зоя Наумовна.txt"

file:"Живопись 59000\Живопись по странам\Живопись СССР\Лесегри.txt"
file:"Живопись 59000\Живопись по странам\Живопись СССР\Лесин, Василий Николаевич
.txt"
file:"Живопись 59000\Живопись по странам\Живопись СССР\Лесничая, Елена Анатольев
на.txt"
file:"Живопись 59000\Живопись по странам\Живопись СССР\Летов, Егор.txt"
file:"Живопись 59000\Живопись по странам\Живопись СССР\Ли, Николай Геннадьевич.t
xt"
file:"Живопись 59000\Живопись по странам\Живопись СССР\Ли, Софья Дмитриевна.txt"

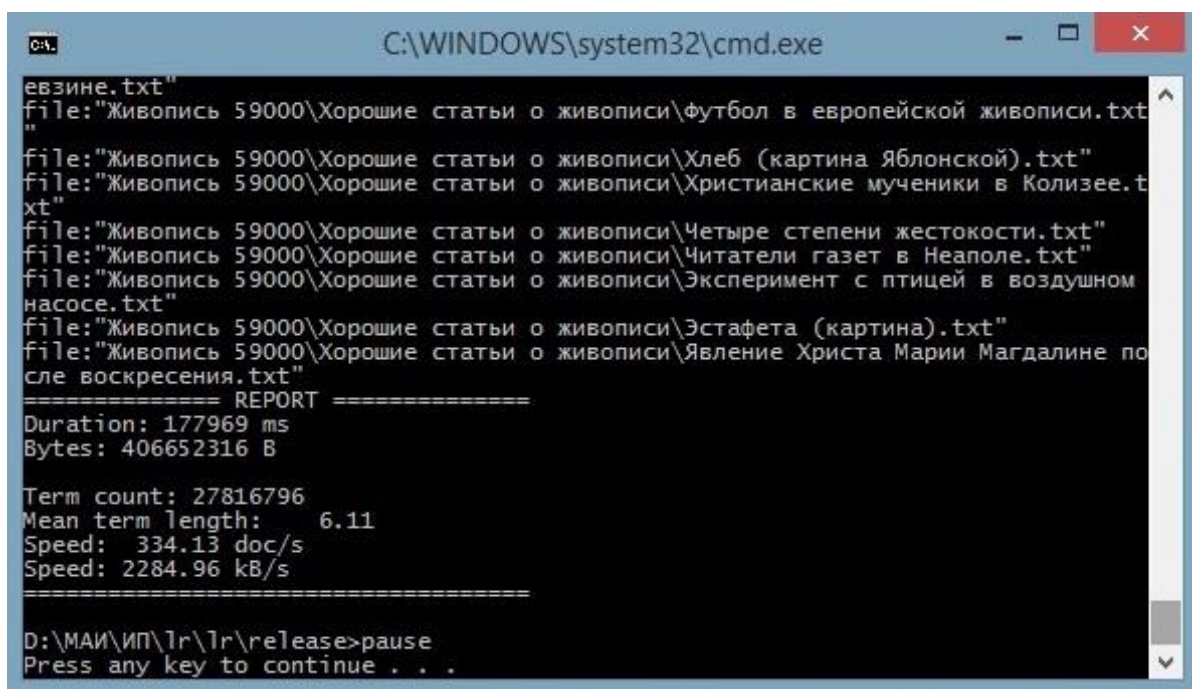
file:"Живопись 59000\Живопись по странам\Живопись СССР\Либакон, Михаил Вадимович
.txt"
file:"Живопись 59000\Живопись по странам\Живопись СССР\Лившиц, Леа-Тути.txt"
file:"Живопись 59000\Живопись по странам\Живопись СССР\Лившиц, Хаим Моисеевич.tx
t"
```

```
C:\WINDOWS\system32\cmd.exe

file:"Живопись 59000\Незавершённые статьи о живописи\Незавершённые статьи о художниках\Статьи о художниках без иллюстраций на Викискладе\Мамедова, Зивер Наджафкули кызы.txt"
file:"Живопись 59000\Незавершённые статьи о живописи\Незавершённые статьи о художниках\Статьи о художниках без иллюстраций на Викискладе\Мамиллов, Руслан Израильевич.txt"
file:"Живопись 59000\Незавершённые статьи о живописи\Незавершённые статьи о художниках\Статьи о художниках без иллюстраций на Викискладе\Мамонов, Богдан Кириллович.txt"
file:"Живопись 59000\Незавершённые статьи о живописи\Незавершённые статьи о художниках\Статьи о художниках без иллюстраций на Викискладе\Мамонтов, Михаил Анатольевич.txt"
file:"Живопись 59000\Незавершённые статьи о живописи\Незавершённые статьи о художниках\Статьи о художниках без иллюстраций на Викискладе\Мамонтов, Николай Андреевич.txt"
file:"Живопись 59000\Незавершённые статьи о живописи\Незавершённые статьи о художниках\Статьи о художниках без иллюстраций на Викискладе\Мамченко, Владислав Николаевич.txt"
file:"Живопись 59000\Незавершённые статьи о живописи\Незавершённые статьи о художниках\Статьи о художниках без иллюстраций на Викискладе\Мамянов, Генрих Арамович.txt"
file:"Живопись 59000\Незавершённые статьи о живописи\Незавершённые статьи о художниках\Статьи о художниках без иллюстраций на Викискладе\Манайло, Фёдор Фёдорович.txt"
```

```
C:\WINDOWS\system32\cmd.exe

овский, Витольд.txt"
file:"Живопись 59000\Стили живописи\Импрессионизм\Художники-импрессионисты\Пурвиртис, Вильгельм.txt"
file:"Живопись 59000\Стили живописи\Импрессионизм\Художники-импрессионисты\Раселл, Джон Питер.txt"
file:"Живопись 59000\Стили живописи\Импрессионизм\Художники-импрессионисты\Рауд, Пауль.txt"
file:"Живопись 59000\Стили живописи\Импрессионизм\Художники-импрессионисты\Семенухин, Игорь Ильич.txt"
file:"Живопись 59000\Стили живописи\Импрессионизм\Художники-импрессионисты\Труш, Иван Иванович.txt"
file:"Живопись 59000\Стили живописи\Импрессионизм\Художники-импрессионисты\Туржанский, Леонард Викторович.txt"
file:"Живопись 59000\Стили живописи\Импрессионизм\Художники-импрессионисты\Уинстон Черчилль и живопись.txt"
file:"Живопись 59000\Стили живописи\Импрессионизм\Художники-импрессионисты\Фешин, Николай Иванович.txt"
file:"Живопись 59000\Стили живописи\Импрессионизм\Художники-импрессионисты\Филиппсен, Теодор Эсберн.txt"
file:"Живопись 59000\Стили живописи\Караваджизм\Караваджизм.txt"
file:"Живопись 59000\Стили живописи\Караваджизм\Караваджо.txt"
file:"Живопись 59000\Стили живописи\Караваджизм\Караваджисты\Бабюррен, Дирк ван.txt"
file:"Живопись 59000\Стили живописи\Караваджизм\Караваджисты\Бассетти, Маркantonio.txt"
```

```
C:\WINDOWS\system32\cmd.exe
евзине.txt"
file:"Живопись 59000\Хорошие статьи о живописи\футбол в европейской живописи.txt"
file:"Живопись 59000\Хорошие статьи о живописи\Хлеб (картина Яблонской).txt"
file:"Живопись 59000\Хорошие статьи о живописи\Христианские мученики в Колизее.txt"
file:"Живопись 59000\Хорошие статьи о живописи\Четыре степени жестокости.txt"
file:"Живопись 59000\Хорошие статьи о живописи\Читатели газет в Неаполе.txt"
file:"Живопись 59000\Хорошие статьи о живописи\Эксперимент с птицей в воздушном
насосе.txt"
file:"Живопись 59000\Хорошие статьи о живописи\Эстафета (картина).txt"
file:"Живопись 59000\Хорошие статьи о живописи\Явление Христа Марии Магдалине по
сле воскресения.txt"
===== REPORT =====
Duration: 177969 ms
Bytes: 406652316 B

Term count: 27816796
Mean term length: 6.11
Speed: 334.13 doc/s
Speed: 2284.96 kB/s
=====
D:\МАИ\ИП\lr\lr\release>pause
Press any key to continue . . .
```

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.