Московский Авиационный Институт
(Национальный исследовательский университет)

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

Лабораторная работа №1
по курсу «Информационный поиск»

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

Преподаватель: Кухтичев А.А.

Москва, 2021

Лабораторная работа № 1

Необходимо подготовить корпус документов, который будет использован при выполнении остальных лабораторных работ. Требуется скачать его к себе на компьютер, ознакомиться с ним и его характеристиками, разбить на документы.

Подготовка к работе

Источником данных является Википедия на русском языке. В качестве корпуса документов была выбрана категория Википедии «Живопись».

Вся категория довольно легко поддалась скачиванию и выделению текста, но процесс занял много времени из-за большого количества статей.

Выделенный текст. Пример статьи

Альфрейная живопись — разновидность интерьерной росписи, имитирующая различные типы и виды отделки внутреннего помещения (например, ценные породы дерева, гипсовую лепнину, шёлковые драпировки, золотые или серебряные поверхности, растительные узоры и т.п.).

Альфрейная живопись отличается от фресковой способом нанесения рисунка на стену: фресковая роспись предполагает нанесение изображения на стену с сырой штукатуркой, а альфрейная, напротив, — на высохшую оштукатуренную поверхность. В настоящее время под термином «альфрейная живопись» подразумеваются следующие виды орнаментально-декоративной росписи:

монохромная (выполняемая в одном цвете и тоновых градациях этого цвета);

полихромная (многоцветные орнаменты, стилизованные цветочные композиции, фантастические фигурки, геральдические элементы и т.д.);

гризайль (раздвигающие пространство картины-обманки, имитация архитектурных элементов, лепнины и т.д.)Альфрейная живопись зародилась в III тысячелетии до н.э. у древних египтян при росписи стен пирамид. Альфрейная живопись по сухой штукатурке стала вытеснять классическую фреску по сырой штукатурке в эпоху Возрождения, когда искусство постепенно возвращалось к идеалам античности и становилось всё более светским. Наибольшего расцвета альфрейная живопись достигла к концу XVIII века в Венеции. В России образцы альфрейной живописи можно увидеть в дворцах Санкт-Петербурга: Зимний, Александровский, Мраморный, Юсуповский и т.д.

Ссылки

Альфрейная живопись // «Карл Павлович Брюллов» : сайт художника (bryullov.ru) (Проверено 10 апреля 2017)

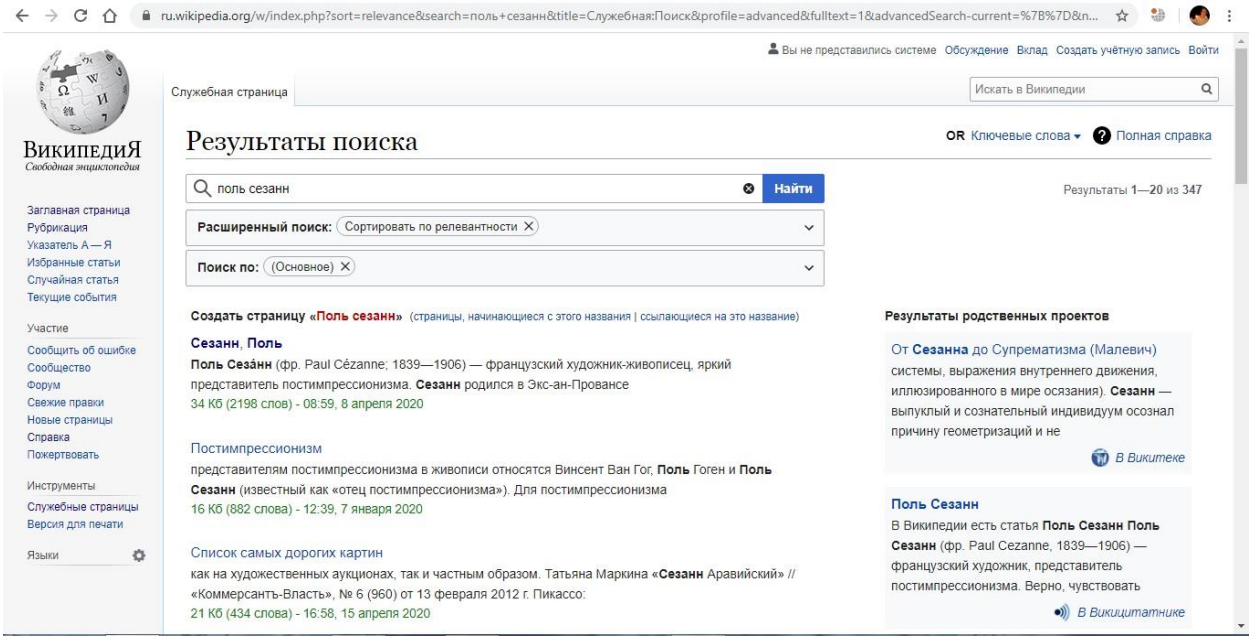
Ход работы

Программа получения корпуса документов была написана на языке Python. Также была использована библиотека wikipedia-api, позволяющая работать с Википедией и средствами языка для записи документов.

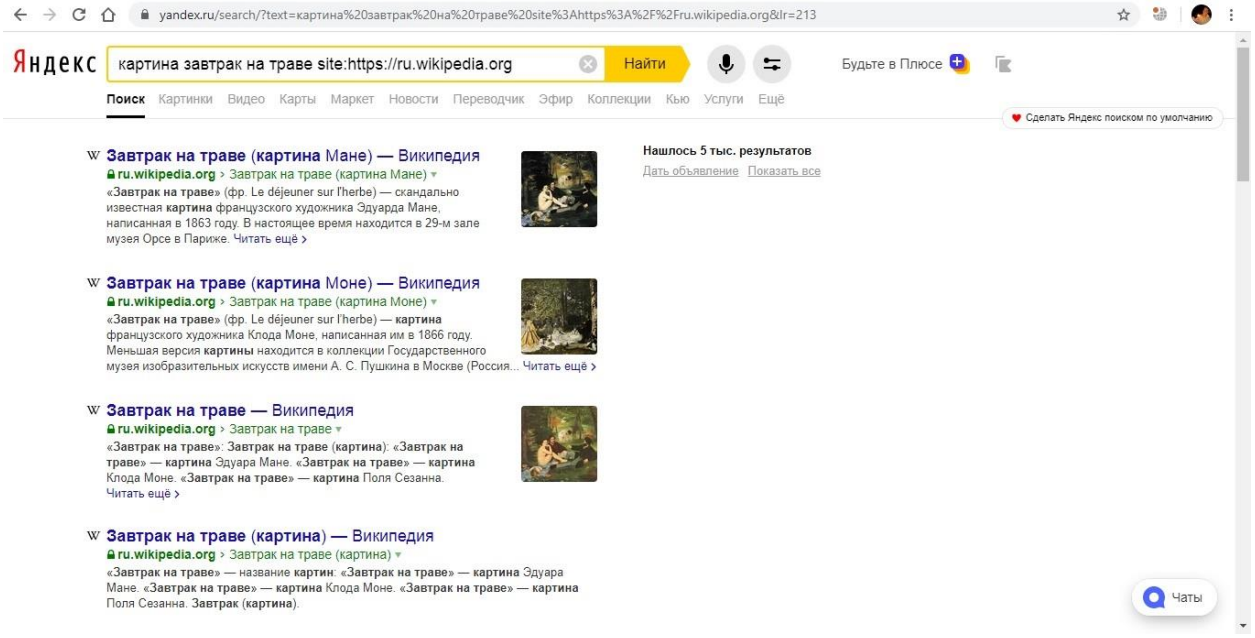
Для поиска по выбранному корпусу документов можно использовать как встроенный поиск Википедии, так и поиски Google, Mail или Яндекс с ограничением на сайт Википедия.

Примеры запросов к существующим поисковикам:

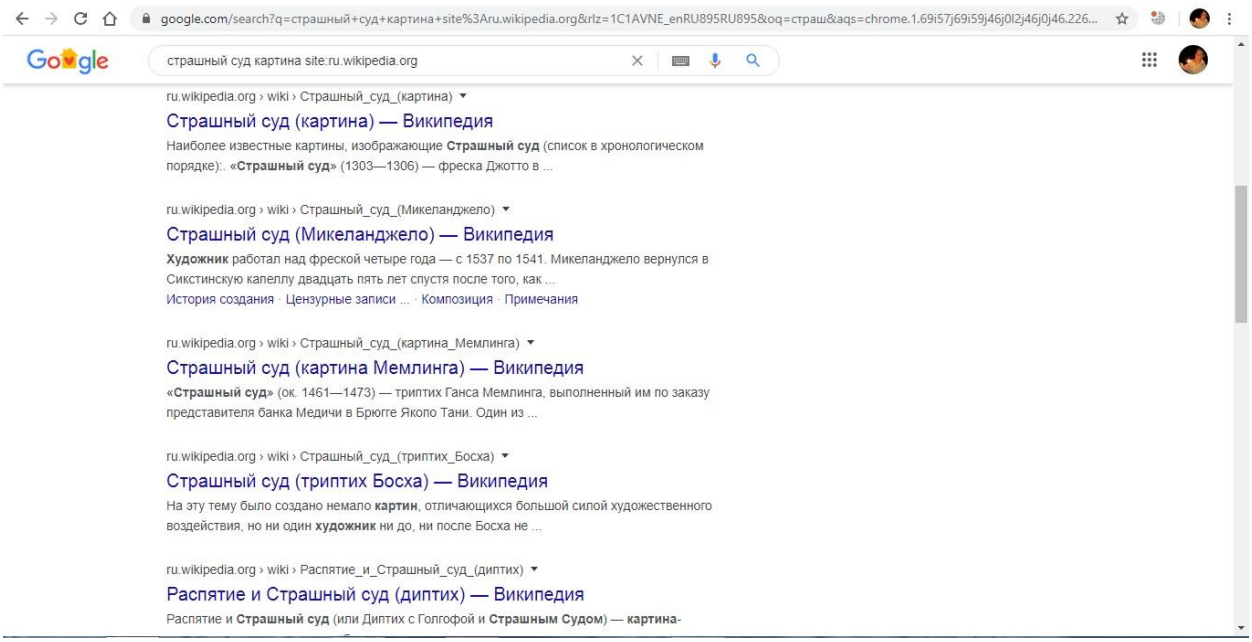
Википедия



Яндекс



Google



Статическая информация о корпусе

«Сырые» данные не сохранялись.

Количество документов – 132938.

Размер «чистого» текста – 975 МБ.

Средний размер документа – 7.3 Кб.

Средний объем текста в документе – 4000 знаков с пробелами.

Вывод

В ходе выполнения лабораторной работы был получен корпус документов по категории «Живопись» с русскоязычной Википедии. Я познакомился с библиотекой Python – wikipedia-api. Для работы с Википедией эта библиотека подходит как можно лучше и позволяет производить импорт категорий и отдельных страниц, их полный текст, разделы страницы и ее перевод на другие языки. В целом время скачивания корпуса документов приблизительно составило 19 часов.

Московский Авиационный Институт
(Национальный исследовательский университет)

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

Лабораторная работа №2
по курсу «Информационный поиск»

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

Преподаватель: Кухтичев А.А.

Москва, 2021

Лабораторная работа № 2

Необходимо оценить качество своего поиска и сравнить их с двумя альтернативами (для Википедии можно собственный поиск по Википедии, поиск Google или Яндекса с ограничением по сайту Википедии). Как минимум, нужно измерить P, DCG, NDCG и ERR уровней @1, @3 и @5, приветствуется использование дополнительных метрик качества.

Для оценки качества необходимо придумать 30 запросов, отражающих интересы пользователей или, если есть доступ к настоящим запросам пользователей, то выбрать репрезентативную подборку.

Ход работы

В качестве исследуемых поисковых систем были выбраны Google и Mail с ограничением по сайту Википедия, а также встроенный поиск Википедия. Далее проводится расчет метрик.

Метрики поиска Google:

Запрос	P@1	P@3	P@5	DCG@1	DCG@3	DCG@5
поль сизан	1	1	1	4	5,5	6,576498
звездная ночь	1	1	1	4	5	6,189645
кто написал завтрак на траве	1	1	0,8	4	6	5,029086
репин не ждали	1	1	0,8	4	4,5	4,642234
богатыри	1	1	1	1	3	3,868528
фрески рафаэля	1	1	1	4	5	6,576498
сколько грачей на картине грачи прилетели	1	0,333333	0,2	4	2	1,547411
кто расписал сикстинскую капеллу	1	1	1	4	5,5	5,802792
лунная ночь над днепром автор	1	1	1	4	4	5,802792
эпоха возрождения	1	1	1	4	4,5	5,029086
потолок исакиевского собора	1	0,666667	0,4	4	2,5	1,934264
страшный суд картина	1	1	1	4	5	5,802792
в каком жанре писал дали	1	0,666667	0,4	4	2,5	1,934264
импрессионизм это	1	0,666667	0,6	4	3,5	3,868528
вторая мировая война в мировой живописи	1	0,666667	0,8	3	3,5	5,029086
монализа особенности полотна	1	0,666667	0,6	4	3,5	3,481675
картина репина приплыли кто автор	1	0,666667	0,6	4	2,5	2,321117

беллерофонт в походе против химеры русский музей сколько эскизов	1	0,666667	0,4	4	2,5	1,934264
волна айвазовского	1	1	1	4	4,5	4,255381
пинакотека ватикана	1	1	1	4	5,5	6,576498
мадона с младенцем	1	1	1	4	6	7,350203
сколько картин написал даВинчи	1	1	1	4	5	5,415939
итальянская живопись ренессанса	1	1	1	4	6	7,737056
картины с венерой	1	1	1	4	5	6,963351
что такое русский авангард	1	1	1	4	5,5	6,963351
руско турецкая война в живописи	1	1	1	3	5	6,189645
коллекция полотен эрмитажа	1	1	1	3	4,5	5,802792
экспозиция лувра	1	1	1	4	5	5,802792
василий поленов биография	1	0,666667	0,8	4	4	3,868528
кто автор витязя на распутье	1	0,333333	0,4	4	2	2,70797

Запрос	**NDCG@1**	**NDCG@3**	**NDCG@5**
поль сизан	1	0,916667	0,85
звездная ночь	1	0,833333	0,842105
кто написал завтрак на траве	1	1	0,722222
репин не ждали	1	0,818182	0,666667
богатыри	0,25	0,545455	0,588235
фрески рафаэля	1	0,909091	1
сколько грачей на картине грачи прилетели	1	0,363636	0,235294
кто расписал сикстинскую капеллу	1	1	0,9375
лунная ночь над днепром автор	1	0,8	0,9375
эпоха возрождения	1	0,9	0,866667
потолок исакиевского собора	1	0,5	0,333333
страшный суд картина	1	1	1
в каком жанре писал дали	1	0,5	0,333333
импрессионизм это	1	0,7	0,666667

вторая мировая война в мировой живописи	0,75	0,777778	0,928571
монализа особенности полотна	1	0,777778	0,692308
картина репина приплыли кто автор	1	0,555556	0,461538
беллерофонт в походе против химеры русский музей сколько эскизов	1	0,555556	0,384615
волна айвазовского	1	1	0,916667
пинакотека ватикана	1	1	1
мадона с младенцем	1	0,909091	0,9375
сколько картин написал даВинчи	1	1	0,85
итальянская живопись ренессанса	1	1	0,722222
картины с венерой	1	0,9	0,85
что такое русский авангард	1	0,818182	0,692308
руско турецкая война в живописи	0,75	1	0,866667
коллекция полотен эрмитажа	1	1	1
экспозиция лувра	1	0,777778	0,85
василий поленов биография	1	0,833333	0,666667
кто автор витязя на распутье	1	1	0,842105

Метрики поиска Mail:

Запрос	**P@1**	**P@3**	**P@5**	**DCG@1**	**DCG@3**	**DCG@5**
поль сизан	1	0,666667	0,4	4	2,5	1,934264
звездная ночь	1	0,666667	0,4	4	3,5	2,70797
кто написал завтрак на траве	1	0,666667	0,6	4	4	3,481675
репин не ждали	1	0,666667	0,4	4	4	3,094822
богатыри	1	1	0,6	4	4	3,094822
фрески рафаэля	1	1	0,8	4	5	5,029086
сколько грачей на картине грачи прилетели	1	0,666667	0,4	4	3	2,321117

кто расписал сикстинскую капеллу	1	1	1	4	5	5,802792
лунная ночь над днепром автор	1	0,333333	0,6	4	2	3,868528
эпоха возрождения	0	0,333333	0,2	0	2	1,547411
потолок исакиевского собора	1	0,666667	0,4	3	3,5	2,70797
страшный суд картина	1	0,666667	0,8	4	4	5,802792
в каком жанре писал дали	0	0	0	0	0	0
импрессионизм это	0	0,333333	0,2	0	2	1,547411
вторая мировая война в мировой живописи	0	0	0	0	0	0
монализа особенности полотна	1	0,666667	0,4	4	3,5	2,70797
картина репина приплыли кто автор	1	1	0,6	2	4,5	3,481675
беллерофонт в походе против химеры русский музей сколько эскизов	1	0,666667	0,4	2	3	2,321117
волна айвазовского	1	0,666667	0,8	1	2,5	5,029086
пинакотека ватикана	1	0,666667	0,8	4	4	3,868528
мадона с младенцем	1	1	1	4	6	7,737056
сколько картин написал даВинчи	1	1	0,8	2	5	4,642234
итальянская живопись ренессанса	1	0,666667	0,6	4	4	4,642234
картины с венерой	1	0,666667	0,6	4	4	4,642234
что такое русский авангард	1	0,666667	0,4	4	3,5	2,70797
руско турецкая война в живописи	1	0,333333	0,6	4	2	3,868528
коллекция полотен эрмитажа	1	0,666667	0,6	3	3,5	4,255381
экспозиция лувра	1	1	1	4	4,5	4,255381
василий поленов биография	1	0,666667	0,6	4	3	3,094822
кто автор витязя на распутье	1	0,333333	0,2	4	2	1,547411

Запрос	**NDCG@1**	**NDCG@3**	**NDCG@5**
поль сизан	1	0,416667	0,25
звездная ночь	1	0,7	0,466667
кто написал завтрак на траве	1	0,8	0,6

репин не ждали	1	0,8	0,615385
богатыри	1	0,888889	0,615385
фрески рафаэля	1	0,888889	0,783333
сколько грачей на картине грачи прилетели	1	0,75	0,5
кто расписал сикстинскую капеллу	1	1	1
лунная ночь над днепром автор	1	0,5	0,909091
эпоха возрождения	0	0,5	0,363636
потолок исакиевского собора	0,75	0,875	0,7
страшный суд картина	1	1	0,75
в каком жанре писал дали	0	0	0
импрессионизм это	0	0,571429	0,444444
вторая мировая война в мировой живописи	0	0	0
монализа особенности полотна	1	1	0,875
картина репина приплыли кто автор	0,5	0,857143	0,909091
беллерофонт в походе против химеры русский музей сколько эскизов	0,5	0,857143	0,75
волна айвазовского	0,25	0,833333	0,980241
пинакотека ватикана	1	0,571429	0,615385
мадона с младенцем	0,666667	0,8	0,980241
сколько картин написал даВинчи	0,666667	1	0,909091
итальянская живопись ренессанса	1	0,6	1
картины с венерой	1	1	1
что такое русский авангард	1	0,75	0,875
руско турецкая война в живописи	1	1	0,909091
коллекция полотен эрмитажа	0	0,75	0,783333
экспозиция лувра	0	0,875	0,783333
василий поленов биография	0	0	0
кто автор витязя на распутье	0	0	0

Метрики поиска Википедии:

Запрос	P@1	P@3	P@5	DCG@1	DCG@3	DCG@5
поль сизан	0	0	0	0	0	0
звездная ночь	1	0,666667	0,4	4	3,5	2,70797
кто написал завтрак на траве	1	0,666667	0,6	4	4	3,481675
репин не ждали	1	0,666667	0,6	4	4	3,868528
богатыри	1	1	0,6	3	5	3,868528
фрески рафаэля	1	1	1	4	6	6,576498
сколько грачей на картине грачи прилетели	1	0,333333	0,2	4	2	1,547411
кто расписал сикстинскую капеллу	1	0,666667	0,8	4	4	5,029086
лунная ночь над днепром автор	1	1	0,6	4	4,5	3,481675
эпоха возрождения	1	0,666667	0,4	4	4	3,094822
потолок исакиевского собора	0	0,333333	0,2	0	0,5	0,386853
страшный суд картина	1	1	1	3	5,5	6,576498
в каком жанре писал дали	0	0	0	0	0	0
импрессионизм это	0	0,333333	0,4	0	1	2,321117
вторая мировая война в мировой живописи	0	0	0,2	0	0	1,160558
монализа особенности полотна	0	0	0	0	0	0
картина репина приплыли кто автор	0	0	0	0	0	0
беллерофонт в походе против химеры русский музей сколько эскизов	1	0,333333	0,2	4	2	1,547411
волна айвазовского	1	1	1	4	4,5	4,255381
пинакотека ватикана	1	0,666667	0,6	4	3,5	3,868528
мадона с младенцем	1	1	1	2	5	6,963351
сколько картин написал даВинчи	0	0	0	0	0	0
итальянская живопись ренессанса	1	1	1	4	5	6,963351

картины с венерой	1	1	1	4	6	7,737056
что такое русский авангард	1	1	0,8	4	5,5	5,415939
руско турецкая война в живописи	0	0	0	0	0	0
коллекция полотен эрмитажа	1	1	0,6	4	5	3,868528
экспозиция лувра	1	0,333333	0,4	4	2	2,321117
василий поленов биография	1	0,666667	0,4	4	3,5	2,70797
кто автор витязя на распутье	0	0	0	0	0	0

Запрос	**NDCG@1**	**NDCG@3**	**NDCG@5**
поль сизан	0	0	0
звездная ночь	1	0,583333	0,388889
кто написал завтрак на траве	1	0,727273	0,5
репин не ждали	1	0,727273	0,588235
богатыри	0,75	1	0,588235
фрески рафаэля	1	0,25	0,9
сколько грачей на картине грачи прилетели	1	0,4	0,307692
кто расписал сикстинскую капеллу	1	0,8	0,888889
лунная ночь над днепром автор	1	1	0,9
эпоха возрождения	1	0,888889	0,8
потолок исакиевского собора	0	0,125	0,1
страшный суд картина	0,75	1	1
в каком жанре писал дали	0	0	0
импрессионизм это	0	0,25	0,666667
вторая мировая война в мировой живописи	0	0	0,375
монализа особенности полотна	0	0	0
картина репина приплыли кто автор	0	0	0
беллерофонт в походе против химеры русский музей сколько эскизов	0,666667	1	0,666667
волна айвазовского	0,666667	0,8	1

пинакотека ватикана	1	1	0,727273
мадона с младенцем	1	0	1
сколько картин написал даВинчи	0	0	0
итальянская живопись ренессанса	1	1	0,8
картины с венерой	0	0	0,984132
что такое русский авангард	1	1	0,888889
руско турецкая война в живописи	0	0	0
коллекция полотен эрмитажа	1	0,583333	1
экспозиция лувра	1	0,583333	0,388889
василий поленов биография	1	1	0
кто автор витязя на распутье	0	0	0

Подводя итог, можно сделать следующие выводы по метрикам для поисковых систем:

	Google	**Mail**	**Википедия**
P@1	1	0,866667	0,666667
P@3	0,866667	0,644444	0,544444
P@5	0,826667	0,54	0,466667
DCG@1	3,8	3,1	2,533333
DCG@3	4,283333	3,316667	2,866667
DCG@5	4,900136	3,39141	2,991662
NDCG@1	0,958333	0,644444	0,594444
NDCG@3	0,823047	0,686164	0,490614
NDCG@5	0,75469	0,645592	0,515315

Вывод

Проанализировав метрики качества поиска для систем Google, Mail и внутреннего поиска Википедии, можно сделать вывод о том, что внутренний поиск Википедии во многом уступает другим изученным системам. Особенно это видно на запросах, требующих понимания того, что именно пользователь хочет видеть в выдаче. Поиск Википедии в этом случае выдает статьи, основываясь на вхождении слов запроса, из-за чего результат часто не является релевантным.

Однако, в целом, когда запрос нацелен на поиск чего-то конкретного, например, «Прогноз погоды» или «Торрент», метрики показывают незначительное отличие между поисковыми системами. Но при сложных запросах поиск Google выигрывает.

Московский Авиационный Институт
(Национальный исследовательский университет)

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

Лабораторная работа №3

по курсу «Информационный поиск»

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

Преподаватель: Кухтичев А.А.

Москва, 2021

Лабораторная работа № 3

Требуется построить поисковый индекс, пригодный для булева поиска, по подготовленному в ЛР1 корпусу документов.

Требования к индексу:

- самостоятельно разработанный, бинарный формат представления данных. Формат необходимо описать в отчёте, в побайтовом представлении;
- формат должен предполагать расширение, т.к. в следующих работах он будет меняться под требования новых лабораторных работ;
- использование текстового представления или готовых баз данных не допускается;
- кроме обратного индекса, должен быть создан «прямой» индекс, содержащий в себе как минимум заголовки документов и ссылки на них (понадобятся для выполнения ЛР4, при генерации страницы поисковой выдачи);
- для термов должна быть как минимум понижена капитализация.

Ход работы

В работе были использованы WinAPI для поиска файлов и их перекодирования и хеш-таблица. 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. Благодаря структуре Location вся информация по файлу хранится в одном месте – номер файла в общем списке, номер токена в файле, номер строки и номер символа в строке. 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 Storage.h и Storage.cpp создается класс хранилища, где будут храниться индексы, токены и их положение в файлах статей. Используются процессы сериализации и десериализации для обработки файлов, а также сниппеты для создания описания результатов поиска.

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

В файлах UserVector.h, UserString.h и UserList.h создаются рукописные контейнеры вектора, строки и списка, используемые затем при создании хеш-таблицы в файле UserHashTable.h. В качестве хеш-функции была выбрана функция murmurhash2 – простая и быстрая хеш-функция с хорошим распределением, возвращающая 32-разрядное беззнаковое число. Она описана в файлах murmur_hash2.h и murmur_hash2.cpp.

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.

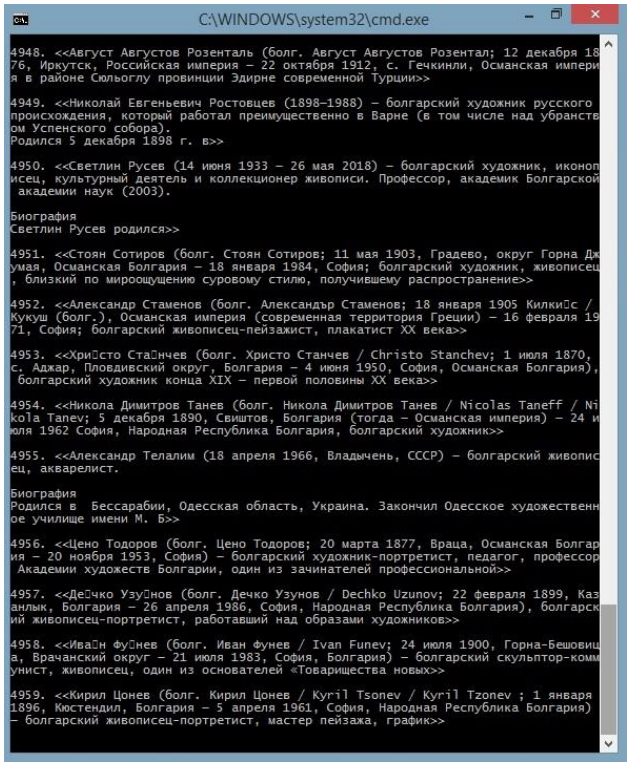
В файле Global.h прописанные шаблоны вектора, строки и списка сравниваются с STL-контейнерами. Это сделано для проверки работы рукописных контейнеров.

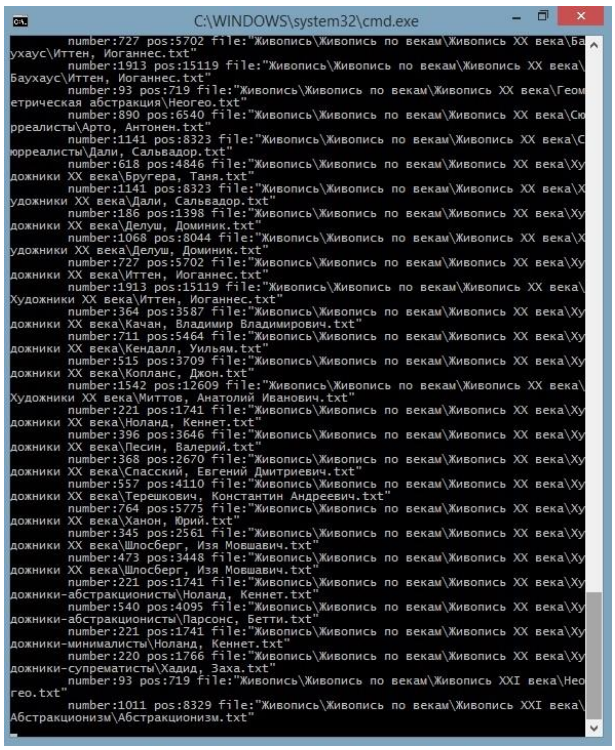
В результате вся работа программы сводится к следующему:

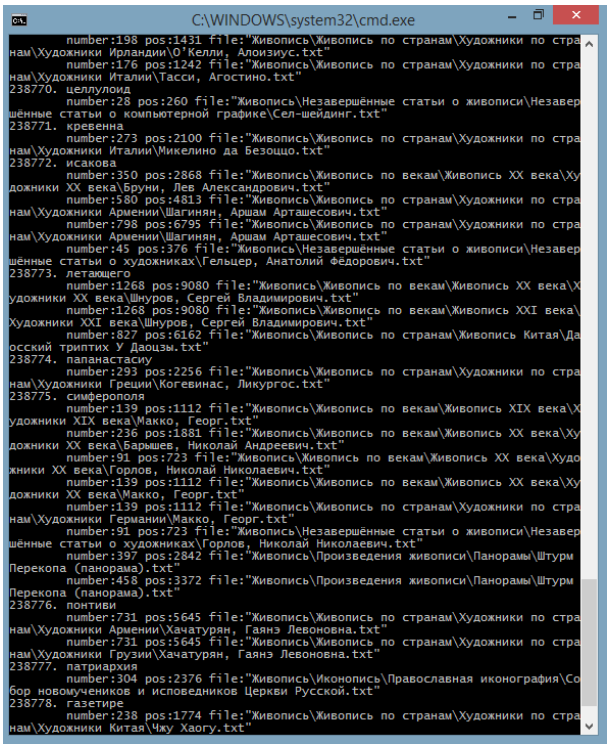
- меняется кодировка;
- начинается поиск файлов, происходит получение полного пути до документов;
- данные файлов разбиваются на токены и сохраняются в бинарный файл индекса;
- далее происходит загрузка полученных данных из файла и вывод токенов в консоль.

Программа запускается в двух режимах через файлы формата bat:

- «Обработка и сохранение» – происходит анализ файлов, строится база данных по поиску и сохраняется в файл index.binary.
- «Подгрузка из файла и вывод» – из полученной базы данных происходит выгрузка найденных слов в консоль.







Статистическая информация

Размер индекса – 69724 Кб.

Время создания индекса базы данных – 1252 секунды.

Время загрузки индекса базы данных – 934 секунды.

Для ускорения работы можно попробовать добавить в хеш-таблицу вектор с указателями на все пары типа «ключ-значение» (в данной лабораторной номер токена и сам токен), чтобы при сохранении полученной базы данных не проверять все ячейки таблицы на пустоту.

Описание файла бинарного формата

Файл состоит из трех последовательных частей – имен документов, их описания и токенов. На один элемент каждой части выделяется по 4 байта. Списки документов, их описания и токены хранятся в контейнерах, сами названия файлов, сниппетов и токенов сохранены в виде строк. Кроме того, для информации по каждому токену используется структура, в которую входят индекс файла, номер токена и его позиция в файле; на каждую часть структуры выделяется по 4 байта.

Московский Авиационный Институт
(Национальный исследовательский университет)

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

Лабораторная работа №4

по курсу «Информационный поиск»

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

Преподаватель: Кухтичев А.А.

Москва, 2021

Лабораторная работа № 4

Нужно реализовать ввод поисковых запросов и их выполнение над индексом, получение поисковой выдачи.

Синтаксис поисковых запросов:

- пробел или два амперсанда, «&&», соответствуют логической операции «И»;
- две вертикальных «палочки», «||» – логическая операция «ИЛИ»;
- восклицательный знак, «!» – логическая операция «НЕТ»;
- могут использоваться скобки.

Парсер поисковых запросов должен быть устойчив к переменному числу пробелов, максимально толерантен к введённому поисковому запросу.

Для демонстрации работы поисковой системы должен быть реализован веб-сервис.

Ход работы

Работа выполнена на основе предыдущей лабораторной работы №3. В работе были использованы WinAPI для поиска файлов и их перекодирования и хеш-таблица. 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 файлах Storage.h и Storage.cpp создается класс хранилища, где будут храниться индексы, токены и их положение в файлах статей. Используются процессы сериализации и десериализации (файлы Serialization.h и Serialization.cpp) для обработки файлов, а также сниппеты для создания описания результатов поиска.

В файле Utils.h модифицируются алгоритмы, активно используемые в STL, под данную задачу – например, функции swap, fill, а также логические операции «&&» и «||».

В файлах UserVector.h, UserString.h и UserList.h создаются рукописные контейнеры вектора, строки и списка, используемые затем при создании хеш-таблицы в файле UserHashTable.h. В качестве хеш-функции была выбрана функция murmurhash2 – простая и быстрая хеш-функция с хорошим распределением, возвращающая 32-разрядное беззнаковое число. Она описана в файлах murmur_hash2.h и murmur_hash2.cpp.

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 прописанные шаблоны вектора, строки и списка сравниваются с STL-контейнерами. Это сделано для проверки работы рукописных контейнеров. Здесь же описана структура Location, благодаря которой вся информация по файлу хранится в одном месте – номер файла в общем списке, номер токена в файле, номер строки и номер символа

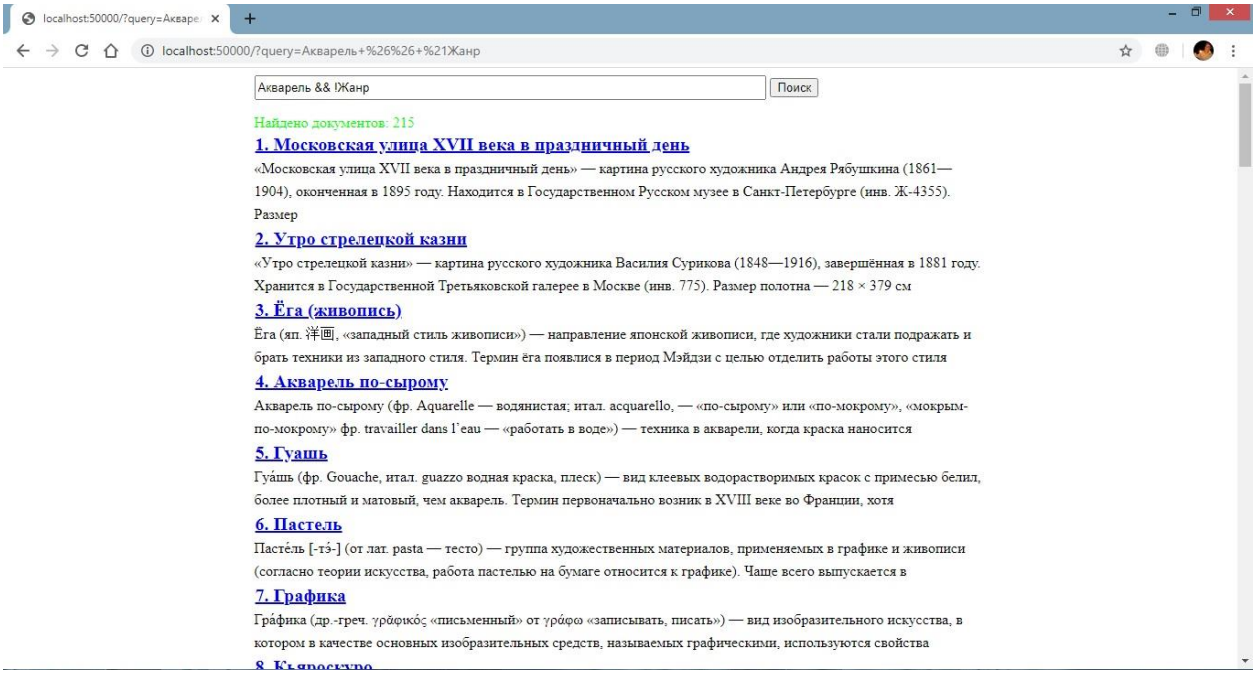
в строке. Повторения токенов игнорируются, а все их упоминания отсортированы по возрастанию индекса документа.

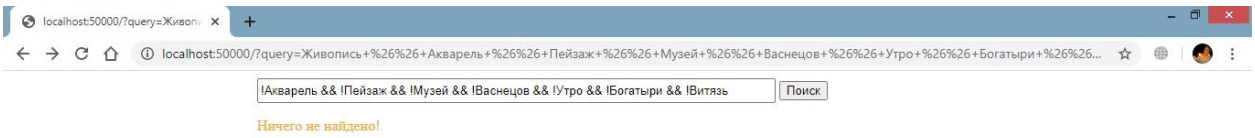
В файлах Query.h и Query.cpp описан соответствующий класс, отвечающий за анализ поискового запроса, его разбор по словам и выполнение логических операций «&&», «||», «!».

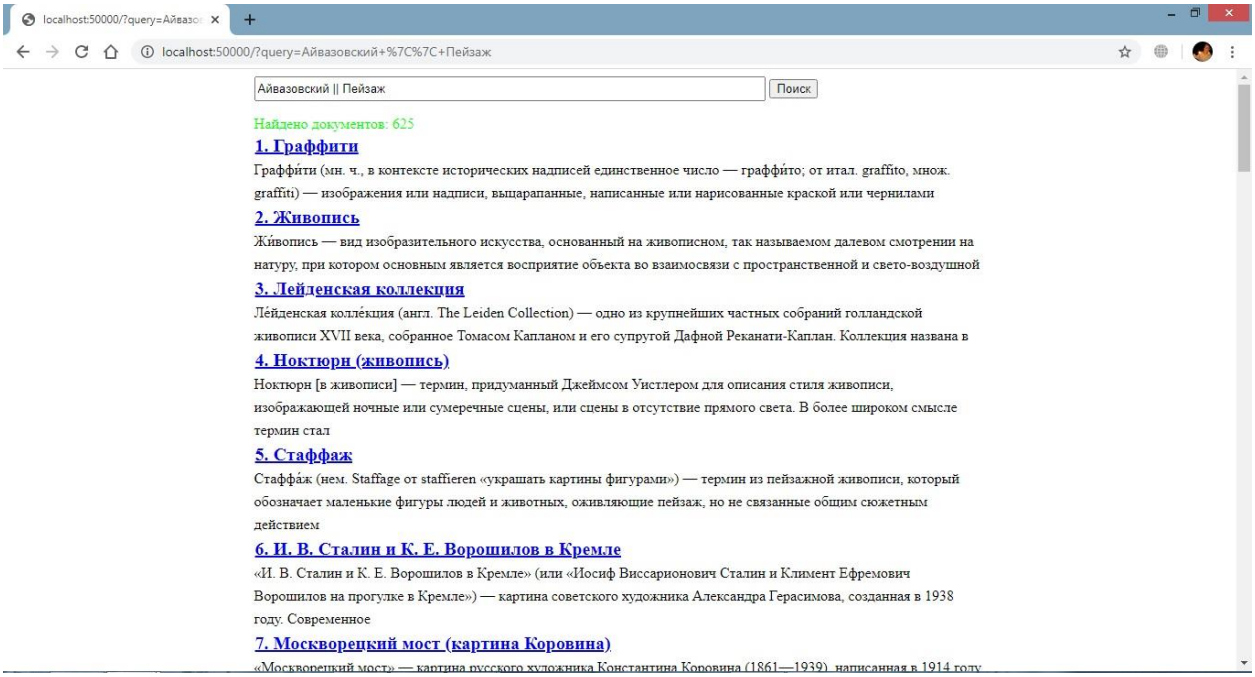
В файлах WsaException.h WsaException.cpp создан класс для обработки информации Winsock. Далее в файлах TcpClient.h, TcpClient.cpp, TcpServer.h, и TcpServer.cpp были описаны классы типа клиент-сервер и на их основе построен HTTP-сервер (файлы HttpServer.h и HttpServer.cpp). Также на html был написан шаблон главной страницы поиска, куда веб-сервер подставляет значения запроса и передает ответ клиенту.

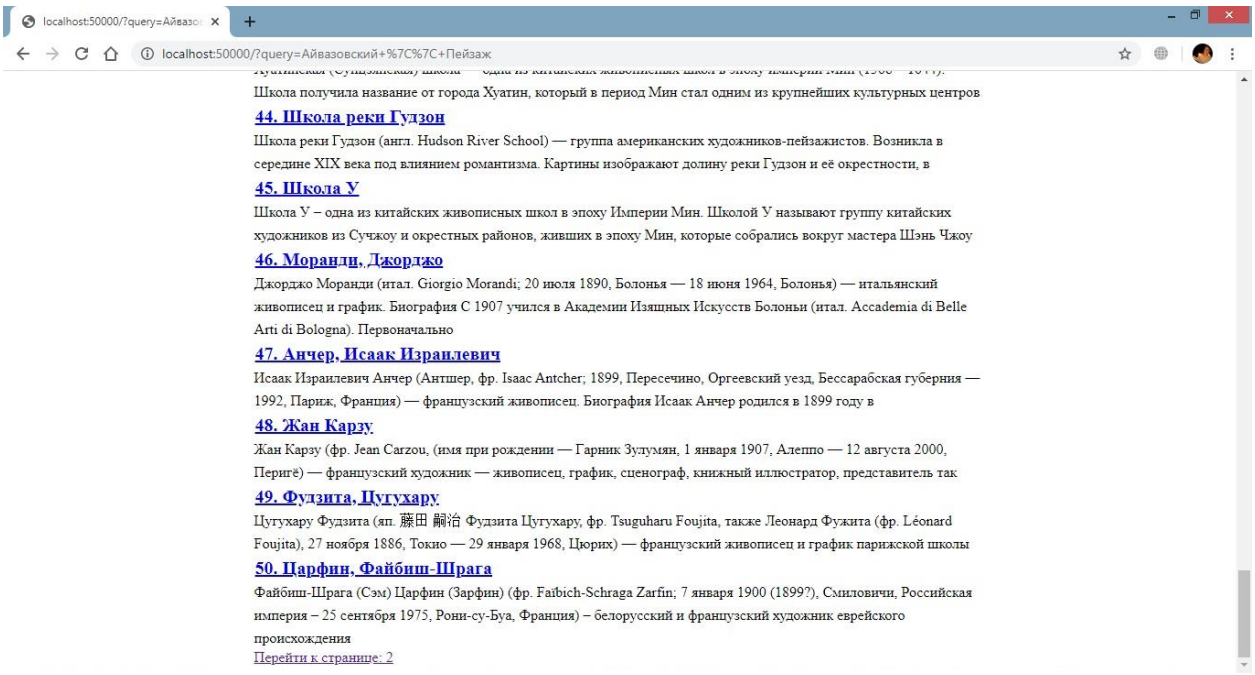
Помимо анализа файлов и создания базы данных токенов в двух режимах («Обработка и сохранение» и «Подгрузка из файла и вывод») предусмотрены еще два режима поиска файлов, запускаемые через файлы формата bat:

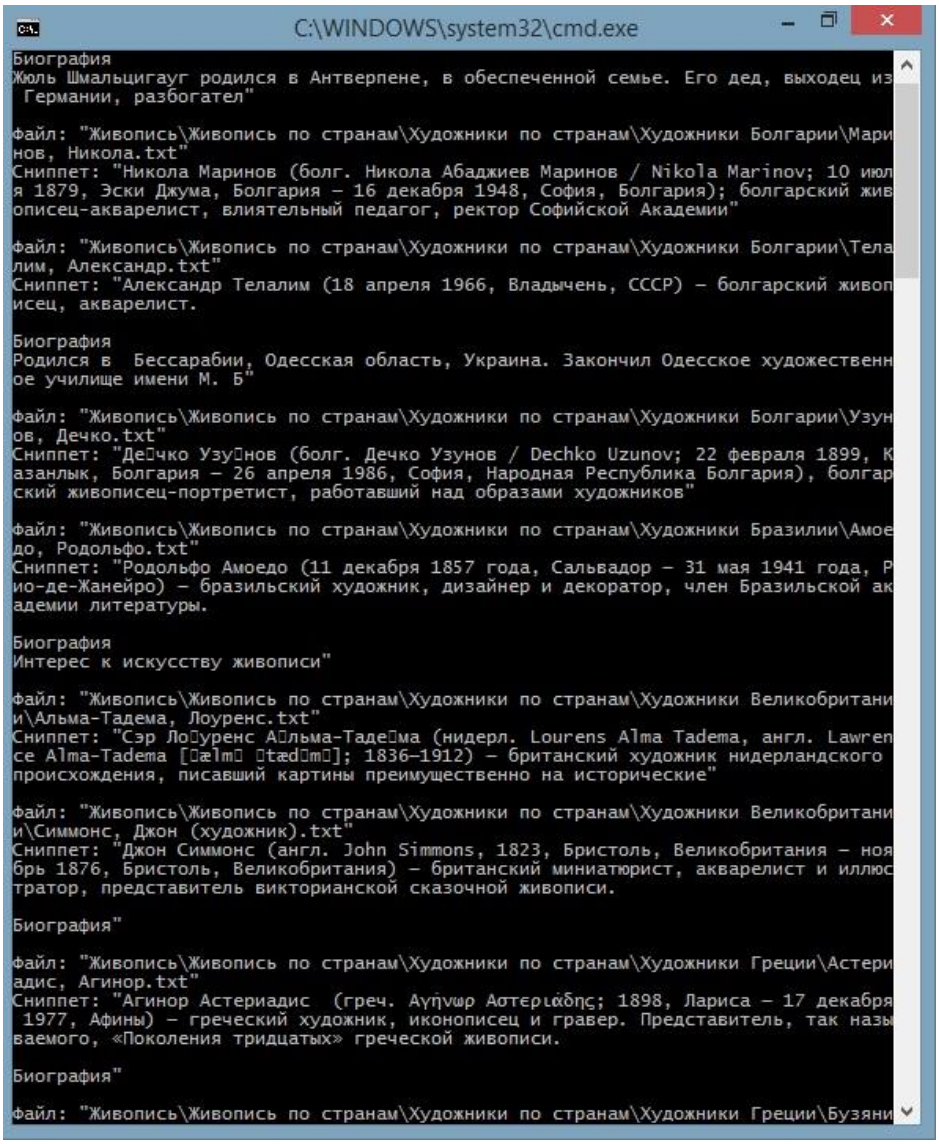
- «Обработка запросов из файла» – поиск происходит в консоли, а запросы передаются из файла queries.txt.
- «Запуск веб-сервера» – поиск происходит в странице браузера после запуска веб-сервера с портом localhost:50000.




localhost:50000/?query=Живопись+%26%26+Акварель+%26%26+Пейзаж+%26%26+Музей+%26%26+Васнецов+%26%26+Утро+%26%26+Богатыри+%26%26...
!Акварель && !Пейзаж && !Музей && !Васнецов && !Утро && !Богатыри && !Витязь
Поиск
Ничего не найдено!


localhost:50000/?query=Айвазовский+%7C%7C+Пейзаж
Айвазовский || Пейзаж
Поиск
Найдено документов: 625
1. Граффити
Граффи́ти (мн. ч., в контексте исторических надписей единственное число — граффи́то; от итал. graffito, множ. graffiti) — изображения или надписи, вышарапанные, написанные или нарисованные краской или чернилами
2. Живопись
Жи́вопись — вид изобразительного искусства, основанный на живописном, так называемом далевом смотрении на натуру, при котором основным является восприятие объекта во взаимосвязи с пространственной и свето-воздушной
3. Лейденская коллекция
Ле́йденская колле́кция (англ. The Leiden Collection) — одно из крупнейших частных собраний голландской живописи XVII века, собранное Томасом Капланом и его супругой Дафной Реканати-Каплан. Коллекция названа в
4. Ноктюрн (живопись)
Ноктюрн [в живописи] — термин, придуманный Джеймсом Уистлером для описания стиля живописи, изображающей ночные или сумеречные сцены, или сцены в отсутствие прямого света. В более широком смысле термин стал
5. Стаффаж
Стаффа́ж (нем. Staffage от staffieren «украшать картины фигурами») — термин из пейзажной живописи, который обозначает маленькие фигуры людей и животных, оживляющие пейзаж, но не связанные общим сюжетным действием
6. И. В. Сталин и К. Е. Ворошилов в Кремле
«И. В. Сталин и К. Е. Ворошилов в Кремле» (или «Иосиф Виссарионович Сталин и Климент Ефремович Ворошилов на прогулке в Кремле») — картина советского художника Александра Герасимова, созданная в 1938 году. Современное
7. Москворецкий мост (картина Коровина)
«Москворецкий мост» — картина русского художника Константина Коровина (1861—1939), написанная в 1914 году





Статистическая информация

Была проведена оценка быстродействия полученной поисковой программы. По одиночным терминам время поиска почти всегда близко к 0 мс. В случае же сложных запросов особого внимания заслуживают запросы с логической операцией «!». С точки зрения работы программы — это самые сложные запросы, поскольку для их поиска приходится перебирать практически все файлы, что занимает время. Например, поиск запроса «Живопись && !Акварель && !Пейзаж && !Музей && !Васнецов && !Утро && !Богатыри && !Витязь» занял уже 15 мс.

Оценка качества поиска

Была произведена оценка качества поиска и ее сравнение с полученными оценками в ходе лабораторной работы №2. Список запросов был взят из той же лабораторной работы.

Запрос	**P@1**	**P@3**	**P@5**	**DCG@1**	**DCG@3**	**DCG@5**
поль сизан	0	0	0	0	0	0
звездная ночь	1	1	1	4	5	4,642234
кто написал завтрак на траве	1	0,333333	0,2	1	0,5	0,386853
репин не ждали	0	0,333333	0,2	0	2	1,547411
богатыри	1	1	1	1	2,5	3,868528
фрески рафаэля	1	1	0,6	3	5	3,868528
сколько грачей на картине грачи прилетели	1	0,333333	0,2	4	2	1,547411
кто расписал сикстинскую капеллу	1	0,666667	0,4	4	4	3,094822
лунная ночь над днепром автор	1	0,333333	0,2	4	2	1,547411
эпоха возрождения	1	0,666667	0,8	2	2	3,094822
потолок исакиевского собора	0	0	0	0	0	0
страшный суд картина	0	0,666667	0,8	0	2,5	3,868528
в каком жанре писал дали	0	0	0	0	0	0
импрессионизм это	1	1	0,6	3	3,5	2,70797

вторая мировая война в мировой живописи	1	1	0,6	2	4	3,094822
монализа особенности полотна	0	0	0	0	0	0
картина репина приплыли кто автор	0	0	0	0	0	0
беллерофонт в походе против химеры русский музей сколько эскизов	0	0	0	0	0	0
волна айвазовского	1	0,666667	0,4	2	2,5	1,934264
пинакотека ватикана	1	1	1	3	4	5,029086
мадона с младенцем	1	0,666667	0,4	2	1,5	1,160558
сколько картин написал даВинчи	1	1	0,6	2	3,5	2,70797
итальянская живопись ренессанса	1	1	1	2	5	6,189645
картины с венерой	0	0,666667	0,8	0	2,5	3,868528
что такое русский авангард	1	1	1	3	4,5	5,029086
руско турецкая война в живописи	0	0	0	0	0	0
коллекция полотен эрмитажа	1	1	0,8	1	3,5	3,094822
экспозиция лувра	1	0,666667	0,4	4	3,5	2,70797
василий поленов биография	1	0,666667	0,4	3	3,5	2,70797
кто автор витязя на распутье	1	1	0,2	2	1	0,773706

Запрос	**NDCG@1**	**NDCG@3**	**NDCG@5**
поль сизан	0	0	0
звездная ночь	1	1	0,923077

кто написал завтрак на траве	0,25	0,1	0,076923
репин не ждали	0	0,444444	0,333333
богатыри	0,25	0,625	1
фрески рафаэля	1	0,714286	1
сколько грачей на картине грачи прилетели	1	0,5	0,4
кто расписал сикстинскую капеллу	1	0,8	0,888889
лунная ночь над днепром автор	1	0,571429	0,5
эпоха возрождения	0,666667	0,571429	1
потолок исакиевского собора	0	0	0
страшный суд картина	0	0,714286	0,666667
в каком жанре писал дали	0	0	0
импрессионизм это	0,75	1	1
вторая мировая война в мировой живописи	1	1	0,923077
монализа особенности полотна	0	0	0
картина репина приплыли кто автор	0	0	0
беллерофонт в походе против химеры русский музей сколько эскизов	0	0	0
волна айвазовского	1	0,714286	0,5
пинакотека ватикана	1	1	0,888889
мадона с младенцем	0,75	1	1
сколько картин написал даВинчи	0,666667	0,5	0,4
итальянская живопись ренессанса	1	1	0,923077
картины с венерой	0	0,8	1
что такое русский авангард	1	0,625	1
руско турецкая война в живописи	0	0	0
коллекция полотен эрмитажа	0	0	0
экспозиция лувра	0	0	0
василий поленов биография	0	0	0

кто автор витязя на распутье	0	0	0

В ходе сравнения были получены следующие выводы:

	Google	**Mail**	**Википедия**	**Полученная поисковая система**
P@1	1	0,866667	0,666667	0,666667
P@3	0,866667	0,644444	0,544444	0,588889
P@5	0,826667	0,54	0,466667	0,453333
DCG@1	3,8	3,1	2,533333	1,733333
DCG@3	4,283333	3,316667	2,866667	2,333333
DCG@5	4,900136	3,39141	2,991662	2,282432
NDCG@1	0,958333	0,644444	0,594444	0,444444
NDCG@3	0,823047	0,686164	0,490614	0,456005
NDCG@5	0,75469	0,645592	0,515315	0,480798

Вывод

В ходе выполнения лабораторной работы была разработана программа, производящая булев поиск по корпусу документов, а также веб-интерфейс для взаимодействия клиента с поисковой программой.

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.

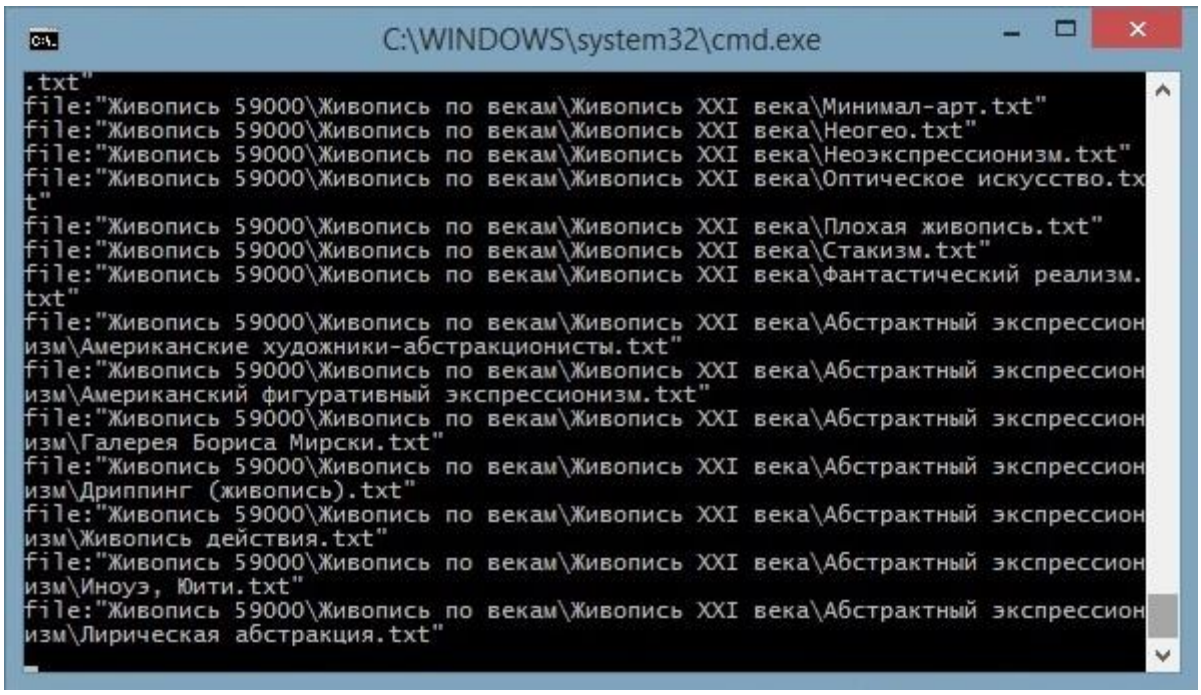
В файле Utils.h модифицируются алгоритмы, активно используемые в STL, под данную задачу – например, функции swap, fill.

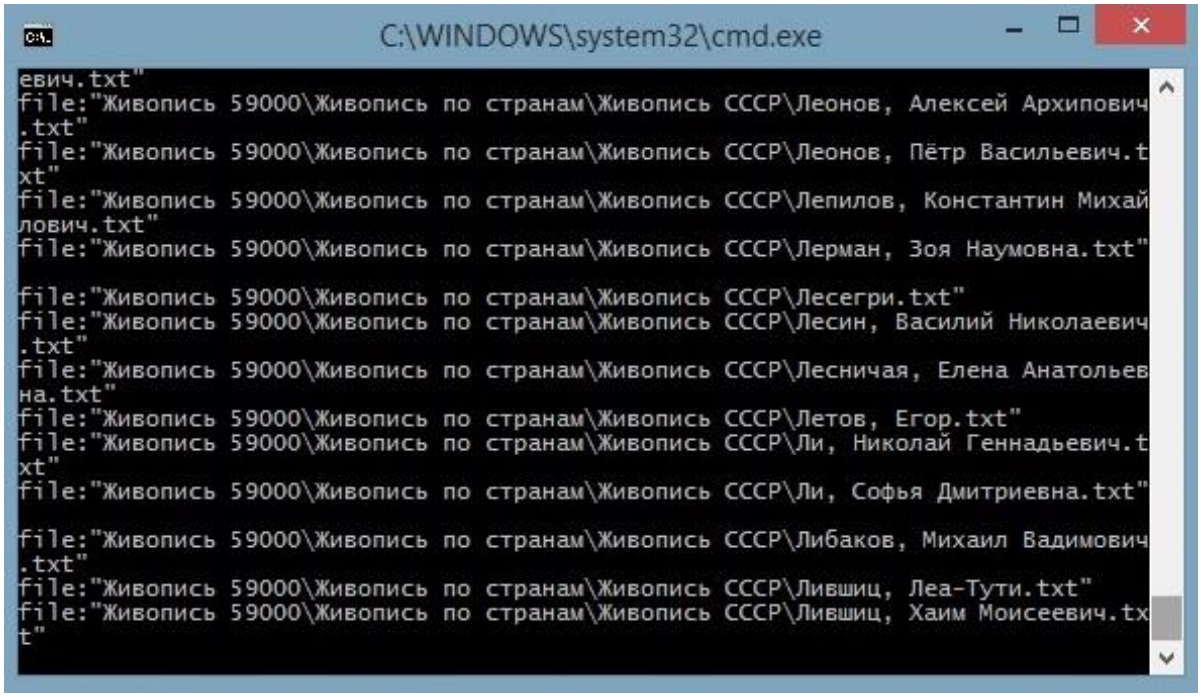
В файлах UserVector.h и UserString.h создаются рукописные контейнеры вектора и строки.

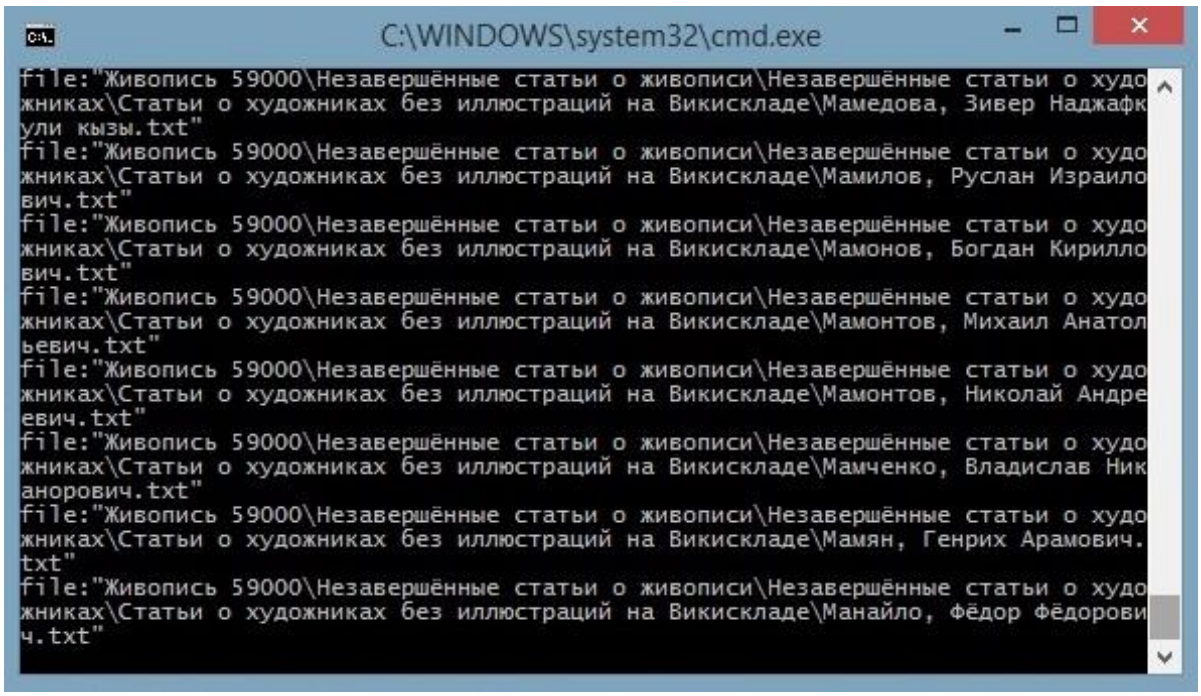
В файлах FS.h и FS.cpp прописаны рукописные обвертки над функциями WinAPI, а именно флаги поиска, поиск файлов в папке, объединение путей, абсолютный путь, существование файла или папки, чтение файла, вывод информации, а также класс файла для чтения или записи. В файлах Exception.h и Exception.cpp прописаны все возможные исключения и ошибки с их кодами.

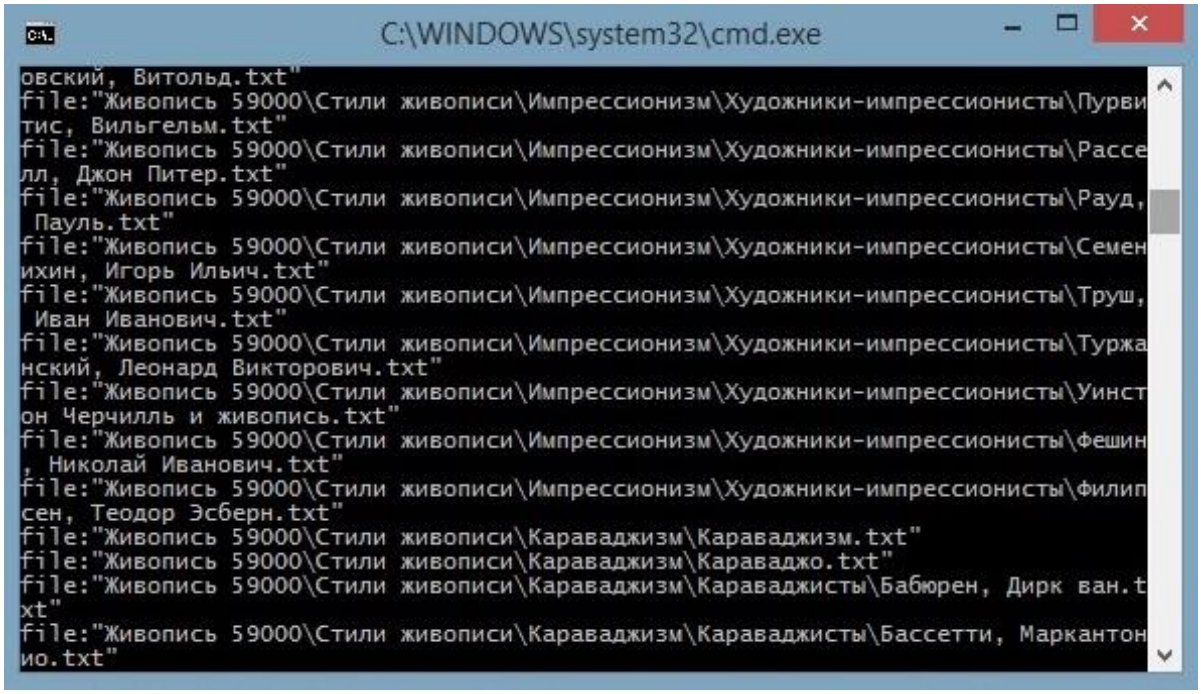
В файле Global.h описана структура Location, благодаря которой вся информация по файлу хранится в одном месте – номер файла, номер токена в файле, номер строки и номер символа в строке.

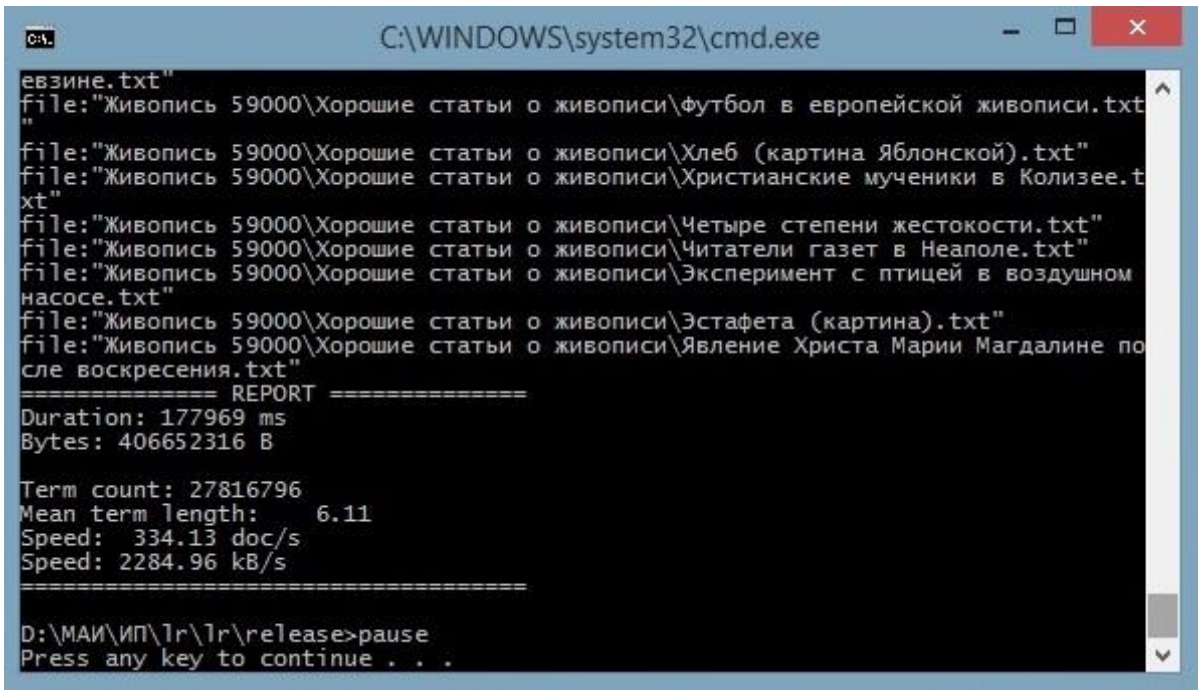
Программа запускается через файл tokenize.bat.


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\Либаков, 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\Мамилов, Руслан Израилович.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 Никанорович.txt"
file:"Живопись 59000\Незавершённые статьи о живописи\Незавершённые статьи о художниках\Статьи о художниках без иллюстраций на Викискладе\Мамян, 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\Пурвитис, Вильгельм.txt"
file:"Живопись 59000\Стили живописи\Импрессионизм\Художники-импрессионисты\Рассел, Джон Питер.txt"
file:"Живопись 59000\Стили живописи\Импрессионизм\Художники-импрессионисты\Рауд, Пауль.txt"
file:"Живопись 59000\Стили живописи\Импрессионизм\Художники-импрессионисты\Семенихин, 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\Филипсен, 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\Бабюрен, Дирк ван.txt"
file:"Живопись 59000\Стили живописи\Караваджизм\Караваджисты\Бассетти, Маркантонио.txt"



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.

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.

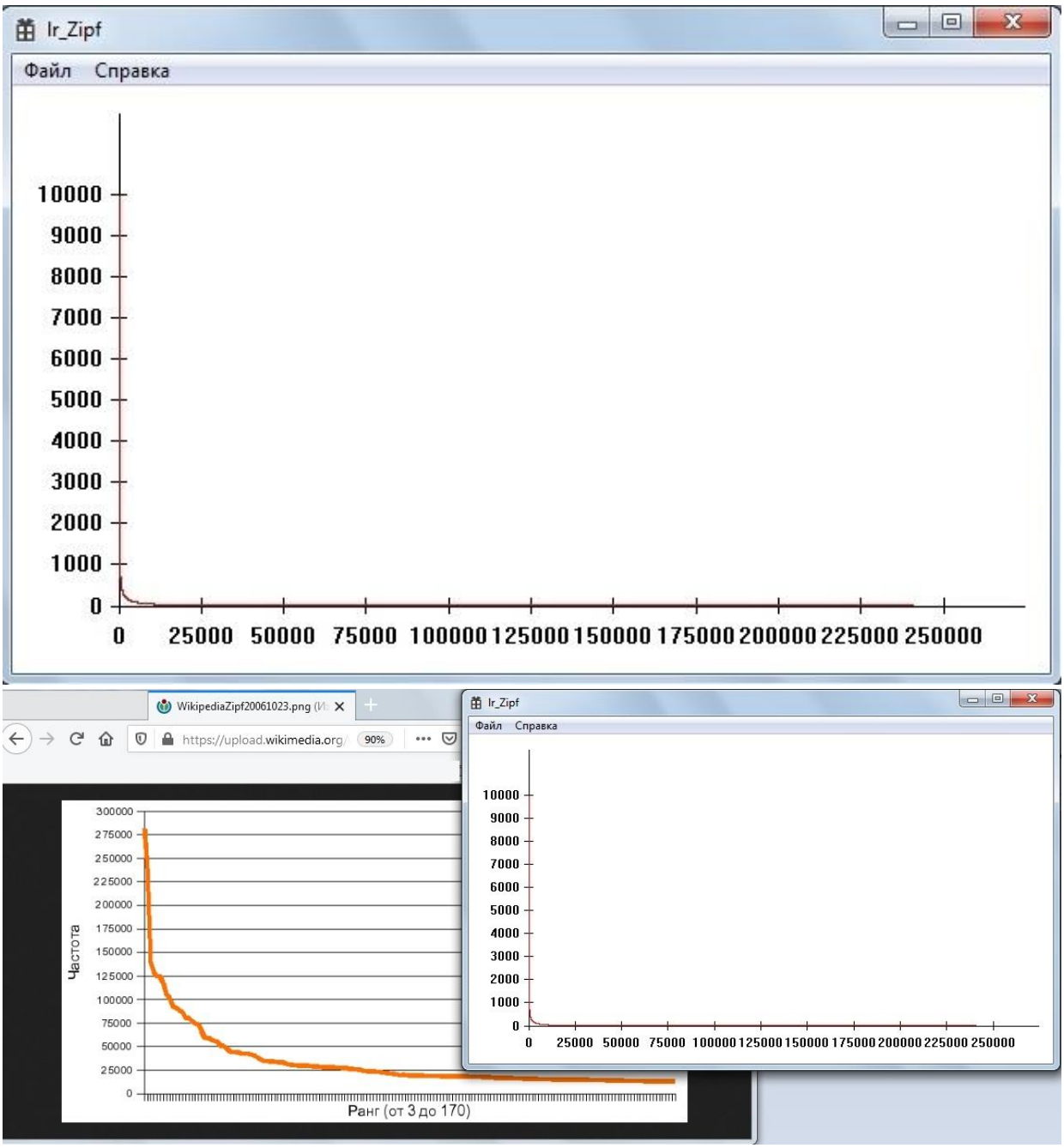
В работе использовались результаты лабораторной работы №3 по курсу «Информационный поиск». Были использованы WinAPI для перекодирования файлов. Поскольку основная часть текста статей на русском, используются только русские слова, а прочие символы – цифры, латиница, греческий и символы кодировки – считаются 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.