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

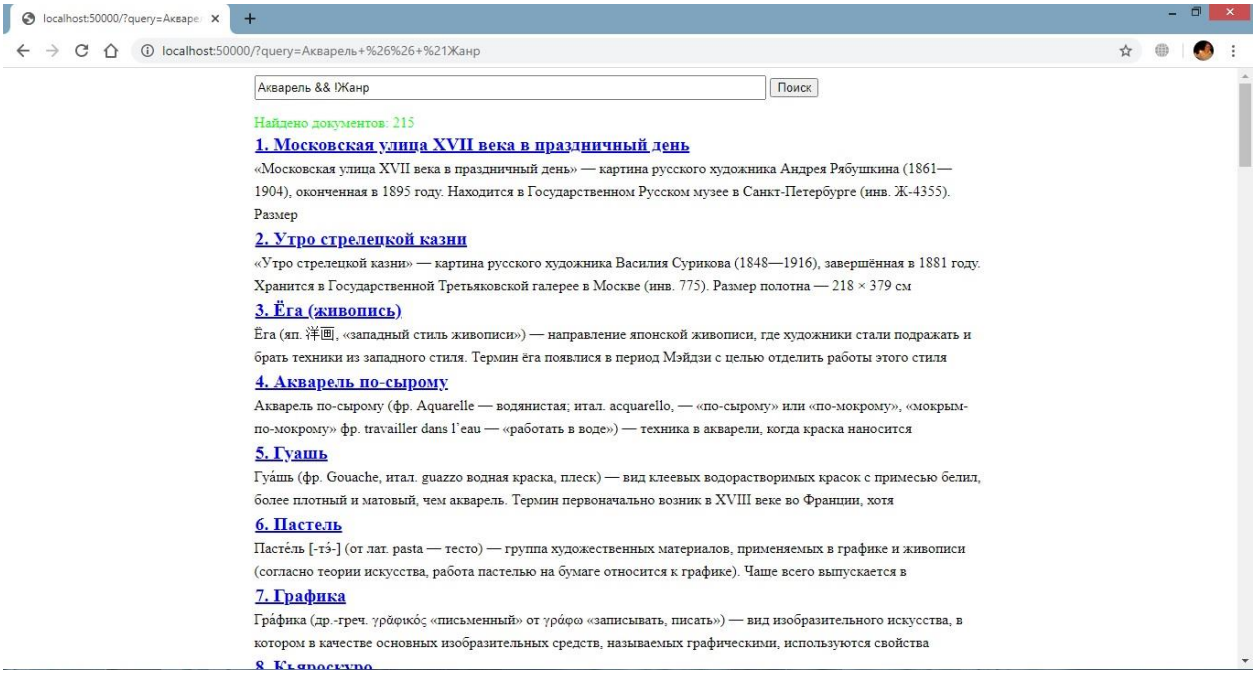
символа в строке. Повторения токенов игнорируются, а все их упоминания отсортированы по возрастанию индекса документа.

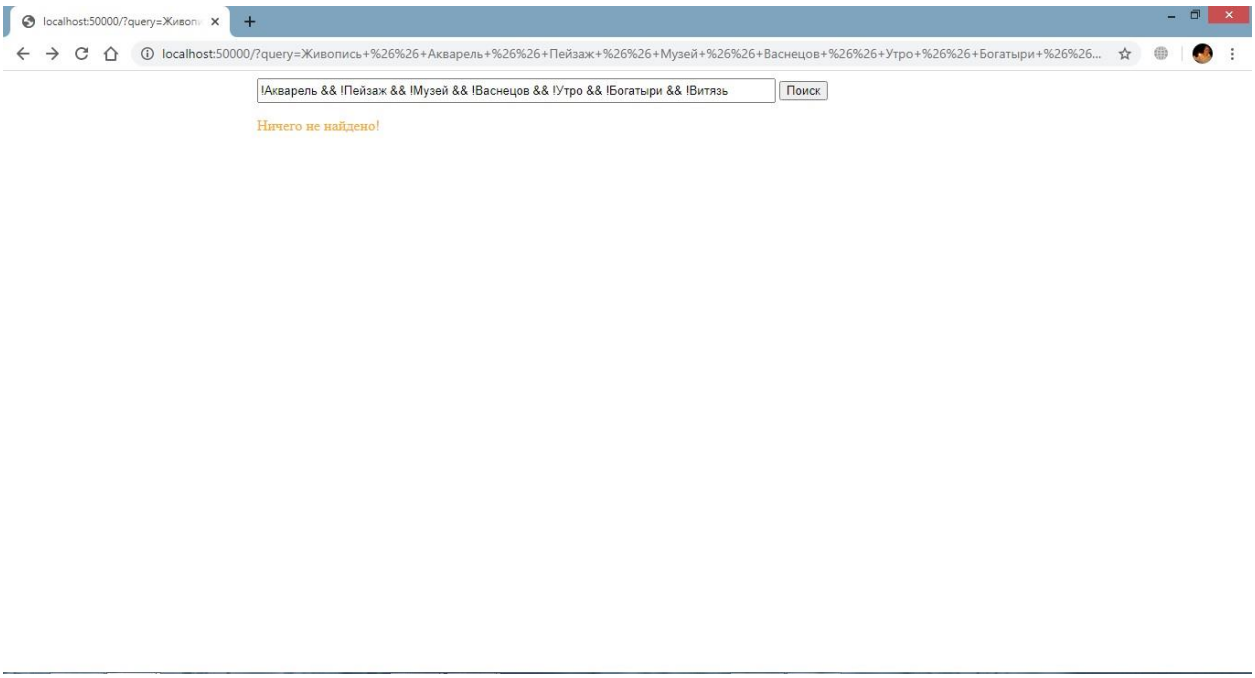
В файлах Query.h и Query.cpp описан соответствующий класс, отвечающий за анализ поискового запроса, его разбор по словам и выполнение логических операций «&&», «||», «!».

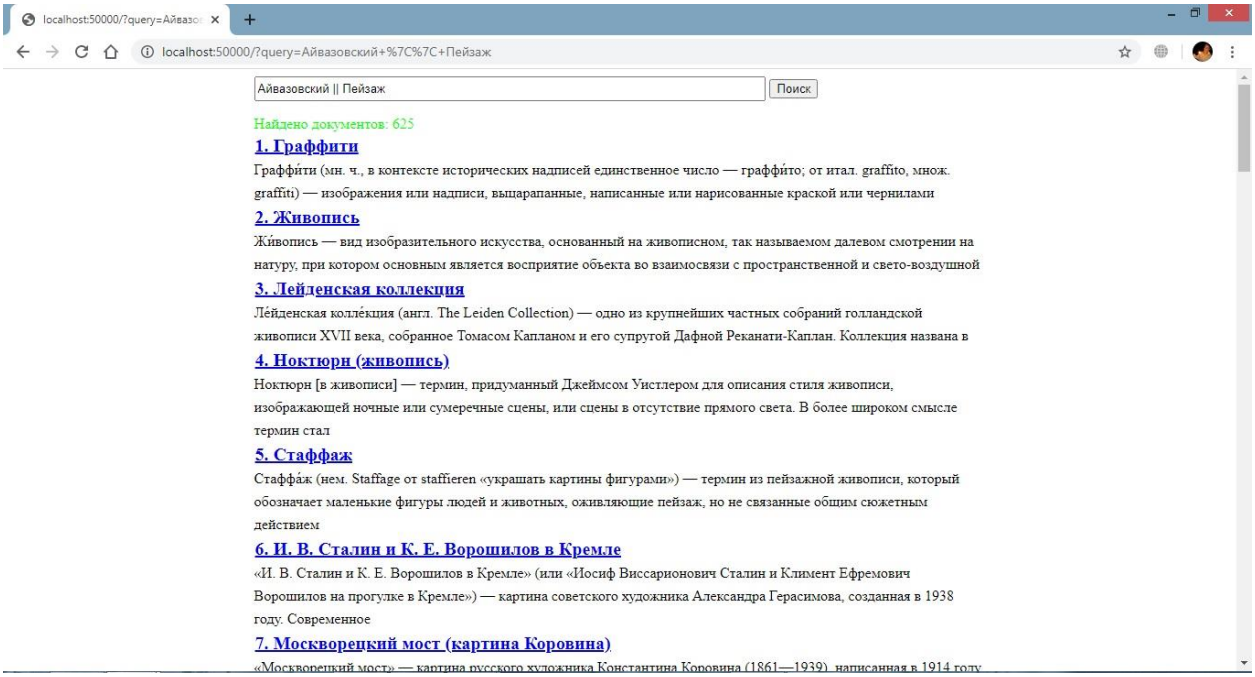
В файлах WsaException.h WsaException.cpp создан класс для обработки информации Winsock. Далее в файлах TcpClient.h, TcpClient.cpp, TcpServer.h, и TcpServer.cpp были описаны классы типа клиент-сервер и на их основе построен HTTP-сервер (файлы HttpServer.h и HttpServer.cpp). Также на html был написан шаблон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

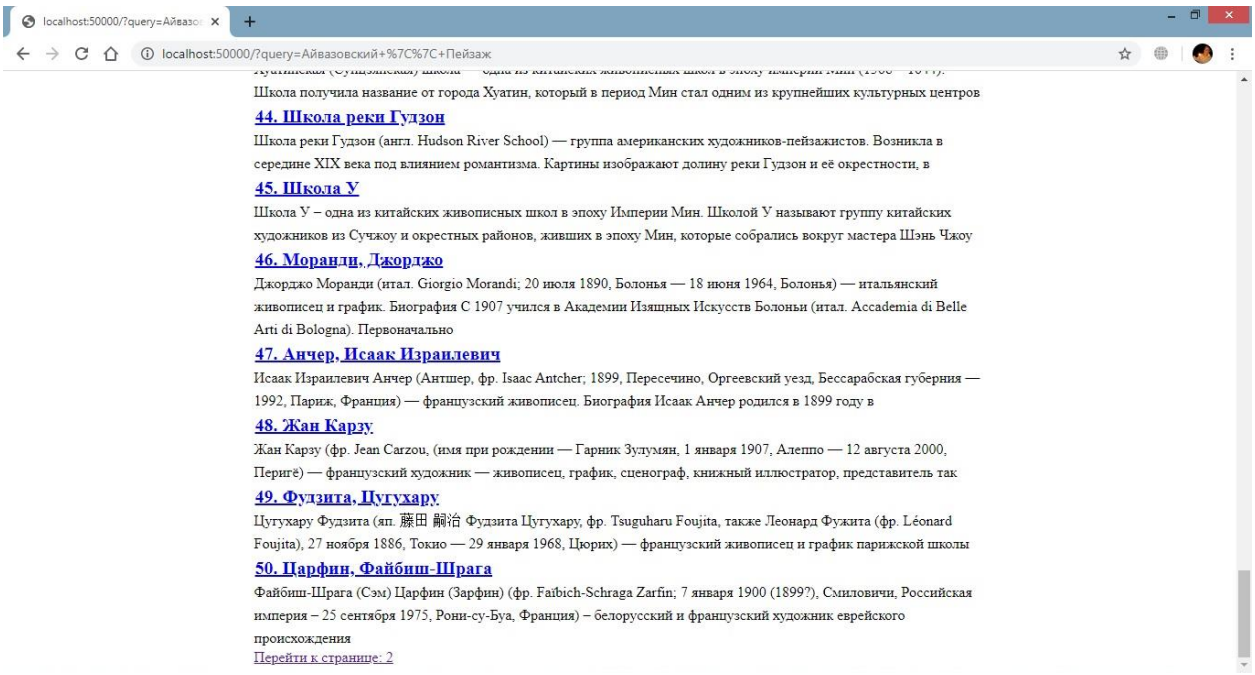
Помимо анализа файлов и создания базы данных токенов в двух режимах («Обработка и сохранение» и «Подгрузка из файла и вывод») предусмотрены еще два режима поиска файлов, запускаемые через файлы формата bat:

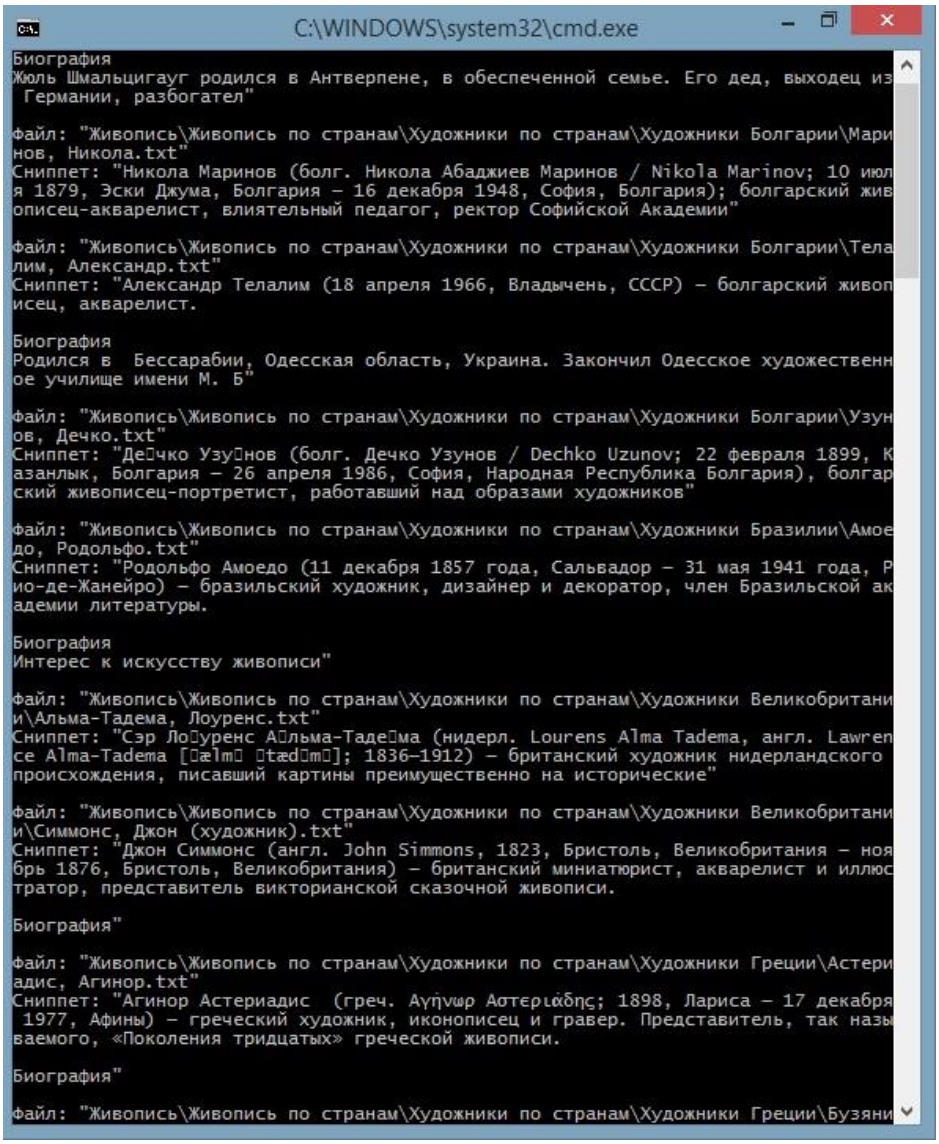
- «Обработка запросов из файла» – поиск происходит в консоли, а запросы передаются из файла queries.txt.
- «Запуск веб-сервера» – поиск происходит в странице браузера после запуска веб-сервера с портом localhost:50000.




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


localhost:50000/?query=Айвазо
localhost:50000/?query=Айвазовский+%7C%7C+Пейзаж
Школа получила название от города Хуатин, который в период Мин стал одним из крупнейших культурных центров
44. Школа реки Гудзон
Школа реки Гудзон (англ. Hudson River School) — группа американских художников-пейзажистов. Возникла в середине XIX века под влиянием романтизма. Картины изображают долину реки Гудзон и её окрестности, в
45. Школа У
Школа У – одна из китайских живописных школ в эпоху Империи Мин. Школой У называют группу китайских художников из Сучжоу и окрестных районов, живших в эпоху Мин, которые собрались вокруг мастера Шэнь Чжоу
46. Моранди, Джорджо
Джорджо Моранди (итал. Giorgio Morandi; 20 июля 1890, Болонья — 18 июня 1964, Болонья) — итальянский живописец и график. Биография С 1907 учился в Академии Изящных Искусств Болоньи (итал. Accademia di Belle Arti di Bologna). Первоначально
47. Анчер, Исаак Израилевич
Исаак Израилевич Анчер (Антшер, фр. Isaac Antcher; 1899, Пересечино, Оргеевский уезд, Бессарабская губерния — 1992, Париж, Франция) — французский живописец. Биография Исаак Анчер родился в 1899 году в
48. Жан Карзу
Жан Карзу (фр. Jean Carzou, (имя при рождении — Гарник Зулумян, 1 января 1907, Алеппо — 12 августа 2000, Периге) — французский художник — живописец, график, сценограф, книжный иллюстратор, представитель так
49. Фудзита, Цугухару
Цугухару Фудзита (яп. 藤田 嗣治 Фудзита Цугухару, фр. Tsuguharu Foujita, также Леонард Фужита (фр. Léonard Foujita), 27 ноября 1886, Токио — 29 января 1968, Цюрих) — французский живописец и график парижской школы
50. Царфин, Файбиш-Шрага
Файбиш-Шрага (Сэм) Царфин (Зарфин) (фр. Faïbich-Schraga Zarfin; 7 января 1900 (1899?), Смиловичи, Российская империя – 25 сентября 1975, Рони-су-Буа, Франция) – белорусский и французский художник еврейского происхождения
Перейти к странице: 2


C:\WINDOWS\system32\cmd.exe
Биография
Жоль Шмальцигауг родился в Антверпене, в обеспеченной семье. Его дед, выходец из
Германии, разбогател"
файл: "Живопись\Живопись по странам\Художники по странам\Художники Болгарии\Мари
нов, Никола.txt"
Сниппет: "Никола Маринов (болг. Никола Абаджиев Маринов / Nikola Marinov; 10 июл
я 1879, Эски Джума, Болгария – 16 декабря 1948, София, Болгария); болгарский жив
описец-акварелист, влиятельный педагог, ректор Софийской Академии"
файл: "Живопись\Живопись по странам\Художники по странам\Художники Болгарии\Тела
лим, Александр.txt"
Сниппет: "Александр Телалим (18 апреля 1966, Владычень, СССР) – болгарский живоп
исец, акварелист.
Биография
Родился в Бессарабии, Одесская область, Украина. Закончил Одесское художественн
ое училище имени М. Б"
файл: "Живопись\Живопись по странам\Художники по странам\Художники Болгарии\Узун
ов, Дечко.txt"
Сниппет: "Де□чко Узу□нов (болг. Дечко Узунов / Dechko Uzunov; 22 февраля 1899, К
азанлык, Болгария – 26 апреля 1986, София, Народная Республика Болгария), болгар
ский живописец-портретист, работавший над образами художников"
файл: "Живопись\Живопись по странам\Художники по странам\Художники Бразилии\Амое
до, Родольфо.txt"
Сниппет: "Родольфо Амоедо (11 декабря 1857 года, Сальвадор – 31 мая 1941 года, Р
ио-де-Жанейро) – бразильский художник, дизайнер и декоратор, член Бразильской ак
адемии литературы.
Биография
Интерес к искусству живописи"
файл: "Живопись\Живопись по странам\Художники по странам\Художники Великобритани
и\Альма-Тадема, Лоуренс.txt"
Сниппет: "Сэр Ло□уренс А□льма-Таде□ма (нидерл. Lourens Alma Tadema, англ. Lawren
ce Alma-Tadema [□ælm□ □tæd□m□]; 1836–1912) – британский художник нидерландского
происхождения, писавший картины преимущественно на исторические"
файл: "Живопись\Живопись по странам\Художники по странам\Художники Великобритани
и\Симмонс, Джон (художник).txt"
Сниппет: "Джон Симмонс (англ. John Simmons, 1823, Бристоль, Великобритания – ноя
брь 1876, Бристоль, Великобритания) – британский миниатюрист, акварелист и иллюс
тратор, представитель викторианской сказочной живописи.
Биография"
файл: "Живопись\Живопись по странам\Художники по странам\Художники Греции\Астери
адис, Агинор.txt"
Сниппет: "Агинор Астериадис (греч. Αγήνωρ Αστεριάδης; 1898, Лариса – 17 декабря
1977, Афины) – греческий художник, иконописец и гравер. Представитель, так назы
ваемого, «Поколения тридцатых» греческой живописи.
Биография"
файл: "Живопись\Живопись по странам\Художники по странам\Художники Греции\Бузяни

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.