

**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 обертки 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 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 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.

```
C:\WINDOWS\system32\cmd.exe
4948. <<Август Августов Розенталь (болг. Август Августов Розентал; 12 декабря 1876, Иркутск, Российская империя – 22 октября 1912, с. Гечкили, Османская империя в районе Сильоглу провинции Зидрие современной Турции>>
4949. <<Николай Евгеньевич Ростовцев (1898-1988) – болгарский художник русского происхождения, который работал преимущественно в Варне (в том числе над убранством Успенского собора). Родился 5 декабря 1898 г. в>>
4950. <<Светлин Русев (14 июня 1933 – 26 мая 2018) – болгарский художник, иконописец, культурный деятель и коллекционер живописи. Профессор, академик Болгарской академии наук (2003).
Биография
Светлин Русев родился>>
4951. <<Стоян Сотиров (болг. Стоян Сотиров; 11 мая 1903, Градево, округ Горна Дюма, Османская Болгария – 18 января 1984, София; болгарский художник, живописец, близкий по мироощущению суровому стилю, получившему распространение>>
4952. <<Александр Стаменов (болг. Александър Стаменов; 18 января 1905 Килкис / Кукуш (болг.), Османская империя (современная территория Греции) – 16 февраля 1971, София; болгарский живописец-пейзажист, плакатист XX века>>
4953. <<Христо Стаичев (болг. Христо Станчев / Christo Stanchev; 1 июля 1870, с. Адавр, Пловдивский округ, Болгария – 4 июня 1950, София, Османская Болгария), болгарский художник конца XIX – первой половины XX века>>
4954. <<Никола Димитров Танев (болг. Никола Димитров Танев / Nicolas Taneff / Nikola Tanev; 5 декабря 1890, Свиштов, Болгария (тогда – Османская империя) – 24 июля 1962 София, Народная Республика Болгария, болгарский художник>>
4955. <<Александр Телалим (18 апреля 1966, Владычье, СССР) – болгарский живописец, акварелист.
Биография
Родился в Бессарабии, Одесская область, Украина. Закончил Одесское художественное училище имени М. Б>>
4956. <<Цено Тодоров (болг. Цено Тодоров; 20 марта 1877, Враца, Османская Болгария – 20 ноября 1953, София) – болгарский художник-портретист, педагог, профессор Академии художеств Болгарии, один из зачинателей профессиональной>>
4957. <<Дечко Узунев (болг. Дечко Узунев / Dechko Uzunov; 22 февраля 1899, Казанлык, Болгария – 26 апреля 1986, София, Народная Республика Болгария), болгарский живописец-портретист, работавший над образами художников>>
4958. <<Иван Фулев (болг. Иван Фулев / Ivan Fulev; 24 июля 1900, Горна-Бешовица, Врачанский округ – 21 июля 1983, София, Болгария) – болгарский скульптор-коммунист, живописец, один из основателей «Товарищества новых>>
4959. <<Кирил Цонев (болг. Кирил Цонев / Kiril Tzonev / Kiril Tzonev; 1 января 1896, Юстендил, Болгария – 5 апреля 1961, София, Народная Республика Болгария) – болгарский живописец-портретист, мастер пейзажа, график>>

C:\WINDOWS\system32\cmd.exe
ершенные статьи о художниках\Ануфриев, Александр Сергеевич.txt"
number:175 pos:1432 file:"Живопись\Незавершенные статьи о живописи\Незавершенные статьи о художниках\Войнов, Константин Семенович.txt"
number:181 pos:1475 file:"Живопись\Незавершенные статьи о живописи\Незавершенные статьи о художниках\Войнов, Константин Семенович.txt"
number:155 pos:1189 file:"Живопись\Незавершенные статьи о живописи\Незавершенные статьи о художниках\Росаль-Воронов, Алексей Семенович.txt"
number:634 pos:4710 file:"Живопись\О художниках\Художники\Колосков, Аркадий Леонидович.txt"
number:600 pos:4427 file:"Живопись\Стили живописи\Лучизм.txt"
number:340 pos:2513 file:"Живопись\Стили живописи\Абстрактный экспрессионизм\Осенний ритм.txt"
number:600 pos:4427 file:"Живопись\Стили живописи\Абстрактный экспрессионизм\Осенний ритм.txt"
number:274 pos:2183 file:"Живопись\Стили живописи\Живопись маньеризма\Венера и Марс (Веронезе).txt"
number:532 pos:3868 file:"Живопись\Стили живописи\Импрессионизм\Импрессионизм (музыка).txt"
number:554 pos:4015 file:"Живопись\Стили живописи\Импрессионизм\Импрессионизм (музыка).txt"
number:1753 pos:12480 file:"Живопись\Хорошие статьи о живописи\Богоматерь с младенцем (Брубель).txt"
number:585 pos:4250 file:"Живопись\Хорошие статьи о живописи\Золотая осень (картина Левитана).txt"
number:491 pos:3717 file:"Живопись\Хорошие статьи о живописи\Золотая осень (картина Остроухова).txt"
number:922 pos:6636 file:"Живопись\Хорошие статьи о живописи\Новоселье (картина Петрова-Водкина).txt"
number:1592 pos:11542 file:"Живопись\Хорошие статьи о живописи\Обливание (картина Фешина).txt"
number:1767 pos:13049 file:"Живопись\Хорошие статьи о живописи\Плот «Медуза».txt"
number:2061 pos:15967 file:"Живопись\Хорошие статьи о живописи\Портрет муромской Ксении Петровой.txt"
number:1118 pos:8400 file:"Живопись\Хорошие статьи о живописи\Сатир в горах у крестьянина.txt"
number:81 pos:638 file:"Живопись\Хорошие статьи о живописи\Владимирка (картина).txt"
number:1278. провизител
number:45 pos:430 file:"Живопись\Красители\Пинакриптол.txt"
number:319 pos:3028 file:"Живопись\Красители\Антрахиноновые красители\Антрахинон-1,7-дисульфоксилота.txt"
number:2032 pos:15329 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Локшин, Давид Борисович.txt"
number:2322 pos:17832 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Локшин, Давид Борисович.txt"
number:467 pos:3101 file:"Живопись\Живопись по векам\Живопись XX века\Сurreалисты\Риго, Жак.txt"
number:12287 pos:97382 file:"Живопись\Живопись по странам\Художники по странам\Художники Великобритании\Маккартни, Пол.txt"
number:156 pos:1218 file:"Живопись\Живописные школы\Живописные школы по алфавиту\Киммерийская школа живописи.txt"
```

```
C:\WINDOWS\system32\cmd.exe
number:727 pos:5702 file:"Живопись\Живопись по векам\Живопись XX века\Баухаус\Иттен, Иоганнес.txt"
number:1913 pos:1519 file:"Живопись\Живопись по векам\Живопись XX века\Баухаус\Иттен, Иоганнес.txt"
number:93 pos:719 file:"Живопись\Живопись по векам\Живопись XX века\Геометрическая абстракция\Неогео.txt"
number:880 pos:6540 file:"Живопись\Живопись по векам\Живопись XX века\Сюрреализм\Арто, Антониетт.txt"
number:1141 pos:8323 file:"Живопись\Живопись по векам\Живопись XX века\Сюрреализм\Дали, Сальвадор.txt"
number:618 pos:4846 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Дели, Сальвадор.txt"
number:1141 pos:8323 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Дели, Сальвадор.txt"
number:186 pos:1398 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Делуш, Доминик.txt"
number:1068 pos:8044 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Делуш, Доминик.txt"
number:727 pos:5702 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Иттен, Иоганнес.txt"
number:1913 pos:1519 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Иттен, Иоганнес.txt"
number:364 pos:3587 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Качан, Владимир Владимирович.txt"
number:711 pos:5464 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Кендалл, Уильям.txt"
number:515 pos:3709 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Копланс, Джон.txt"
number:1542 pos:12609 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Миттов, Анатолий Иванович.txt"
number:221 pos:1741 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Ноланд, Кеннет.txt"
number:386 pos:3648 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Песин, Валерий.txt"
number:368 pos:2670 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Спасский, Евгений Дмитриевич.txt"
number:557 pos:4110 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Терешкину, Константин Андреевич.txt"
number:764 pos:5775 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Ханон, Юрий.txt"
number:345 pos:2561 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Шлосберг, Иза Мошавич.txt"
number:473 pos:3448 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Шлосберг, Иза Мошавич.txt"
number:540 pos:4095 file:"Живопись\Живопись по векам\Живопись XX века\Художники-абстракционисты\Ноланд, Кеннет.txt"
number:221 pos:1741 file:"Живопись\Живопись по векам\Живопись XX века\Художники-абстракционисты\Парсонс, Бетти.txt"
number:221 pos:1741 file:"Живопись\Живопись по векам\Живопись XX века\Художники-минималисты\Ноланд, Кеннет.txt"
number:220 pos:1766 file:"Живопись\Живопись по векам\Живопись XX века\Художники-супрематисты\Хадид, Заха.txt"
number:93 pos:719 file:"Живопись\Живопись по векам\Живопись XXI века\Неогео.txt"
number:1011 pos:8329 file:"Живопись\Живопись по векам\Живопись XXI века\Абстракционизм\Абстракционизм.txt"

C:\WINDOWS\system32\cmd.exe
number:198 pos:1431 file:"Живопись\Живопись по странам\Художники по странам\Художники Ирландии\О'Келли, Алонизус.txt"
number:176 pos:1242 file:"Живопись\Живопись по странам\Художники по странам\Художники Италии\Тасси, Агостино.txt"
238770. целлулоид
number:28 pos:260 file:"Живопись\Незавершенные статьи о живописи\Незавершенные статьи о компьютерной графике\Сел-шейдинг.txt"
238771. красящая
number:273 pos:2100 file:"Живопись\Живопись по странам\Художники по странам\Художники Италии\Микелино да Безоццо.txt"
238772. исакова
number:350 pos:2868 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Бруни, Лев Александрович.txt"
number:580 pos:4813 file:"Живопись\Живопись по странам\Художники по странам\Художники Армении\Шагинян, Аршам Арташевич.txt"
number:798 pos:6735 file:"Живопись\Живопись по странам\Художники по странам\Художники Армении\Шагинян, Аршам Арташевич.txt"
number:445 pos:376 file:"Живопись\Незавершенные статьи о живописи\Незавершенные статьи о художниках\Гельцер, Анатолий Федорович.txt"
238773. летающего
number:1268 pos:9080 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Шнуров, Сергей Владимирович.txt"
number:1268 pos:9080 file:"Живопись\Живопись по векам\Живопись XXI века\Художники XXI века\Шнуров, Сергей Владимирович.txt"
number:827 pos:6162 file:"Живопись\Живопись по странам\Живопись Китая\Даосский триптих У Даоцзы.txt"
238774. папанастасиу
number:293 pos:2256 file:"Живопись\Живопись по странам\Художники по странам\Художники Греции\Когевинас, Ликургос.txt"
238775. синерополя
number:139 pos:1112 file:"Живопись\Живопись по векам\Живопись XIX века\Художники XIX века\Макко, Георг.txt"
number:236 pos:1881 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Барачини, Николай Андреевич.txt"
number:191 pos:723 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Горлов, Николай Николаевич.txt"
number:139 pos:1112 file:"Живопись\Живопись по векам\Живопись XX века\Художники XX века\Макко, Георг.txt"
number:139 pos:1112 file:"Живопись\Живопись по странам\Художники по странам\Художники Германии\Макко, Георг.txt"
number:191 pos:723 file:"Живопись\Незавершенные статьи о живописи\Незавершенные статьи о художниках\Горлов, Николай Николаевич.txt"
number:397 pos:2842 file:"Живопись\Произведения живописи\Панорамы\Штурм Перекла (панорама).txt"
number:458 pos:3372 file:"Живопись\Произведения живописи\Панорамы\Штурм Перекла (панорама).txt"
238776. понтий
number:731 pos:5645 file:"Живопись\Живопись по странам\Художники по странам\Художники Армении\Хачатурян, Гаянэ Левоновна.txt"
number:731 pos:5645 file:"Живопись\Живопись по странам\Художники по странам\Художники Грузии\Хачатурян, Гаянэ Левоновна.txt"
238777. патриархия
number:304 pos:2376 file:"Живопись\Иконопись\Православная иконография\Сбор новомучеников и исповедников Церкви Русской.txt"
238778. газетире
number:238 pos:1774 file:"Живопись\Живопись по странам\Художники по странам\Художники Китая\Чжу Хаогю.txt"
```

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.

Список файлов		
Имя файла		
Длина строки, 4 байта	Один символ Юникод в строке, 2 байта	...

Список описаний файлов		
Текст описания		
Длина строки, 4 байта	Один символ Юникод в строке, 2 байта	...

Список токенов		
Токен в виде строки		
Длина строки, 4 байта	Один символ Юникод в строке, 2 байта	...
Структура с информацией о токенах		
Индекс файла, 4 байта	Номер токена в файле, 4 байта	Позиция в файле, 4 байта

Вывод

В ходе выполнения лабораторной работы был рассмотрен вариант индексирования корпуса документов. В качестве хеш-функции была выбрана функция murmurhash2 как наиболее подходящая к задаче при заданных условиях. Также были самостоятельно написаны все необходимые STL-контейнеры.