

**Московский авиационный институт
(национальный исследовательский университет)**

**Институт информационные технологии и прикладной
математики**

Кафедра вычислительной математики и программирования

**Лабораторная работа №5 по курсу
«Информационный поиск»**

Студент:	Е.М. Стифеев
Преподаватель:	А.А. Кухтичев
Группа:	М8О-109М-21
Дата:	25.11.21
Оценка:	
Подпись:	

Москва, 2021

Лабораторная работа №5 «Поиск цитат, координатный индекс»

В этом задании необходимо расширить язык запросов булева поиска новым элементом – поиском цитат. Синтаксис этого элемента следующий:

- [«что где когда»] – кавычки, включают режим цитатного поиска для терминов внутри кавычек. Этому запросу удовлетворяют документы, содержащие в себе все термины *что*, *где* и *когда*, причём они должны встретиться внутри документа ровно в этой последовательности, без каких-либо вкраплений других терминов.
- [«что где когда» / 5] – аналогично предыдущему пункту, но допускаются вкрапления других терминов так, чтобы расстояние от первого термина цитаты до последнего не превышало бы 5.

Новый элемент может комбинироваться с другими стандартными средствами булева поиска, например:

- [«что где когда» && другъ]
- [«что где когда» || квн]
- [«что где когда» && !«хрустальная сова»]

Для реализации цитатного поиска нужно использовать координатный индекс, т.е. для каждого вхождения термина в документ построить и сохранить список позиций внутри документа, где этот термин встречался.

В отчёте нужно описать формат координатного индекса. Привести статистические данные:

Размер получившегося индекса.

- Время построения индекса.
- Общее количество позиций. Среднее количество позиций на термин и на пару термин-документ.
- Скорость индексации (кб входных данных в секунду)
- Время выполнения поисковых запросов.
- Примеры долго выполняющихся запросов.

Кроме того, нужно привести примеры запросов и результаты их выполнения. В выводах должны быть указаны недостатки работы, приведены примеры их решения. Что можно сделать, чтобы ускорить «долгие» запросы?

1. Описание

Корпус

Напомню, что корпус документов имеет следующую структуру, полученную по результатам ЛР1 (доступен по ссылке <https://cloud.mail.ru/public/ZfkX/gccM7hnDR>):

- Корпус документов
 - films1.txt (94 Мб, 15000 документов)
 - films2.txt (96 Мб, 15000 документов)
 - films3.txt (184 Мб, 15000 документов)
 - films4.txt (219 Мб, 15000 документов)
 - films5.txt (322 Мб, 15000 документов)
 - films6.txt (711 Мб, 15000 документов)
 - films7.txt (823 Мб, 15000 документов)
 - films8.txt (226 Мб, 15000 документов)
 - films9.txt (67 Мб, 15000 документов)
 - films10.txt (75 Мб, 15000 документов)
 - films11.txt (99 Мб, 15000 документов)
 - films12.txt (78 Мб, 15000 документов)
 - films13.txt (41 Мб, 6109 документов)

$$\Sigma_{Gb} = 2,899 \text{ Gb}, \Sigma_{docs} = 186109$$

Также, напомню, что получение одного документа могло включать проход по нескольким html-страницам и обработку динамически подгружаемых страниц, поэтому общее количество обработанных страниц было >800'000.

В каждом файле *.txt документы хранятся следующим образом:

- 1 строка 1 документ {...}
- 2 строка 2 документ {...}
- n строка n документ {...}

Каждый документ снабжён прямой ссылкой на источник, откуда был скачен, и хранит только выделенный из html-кода текст в кодировке UTF-8. Например, 234 строка файла films1.txt выглядит так:

```
{"page_url": "https://www.kinopoisk.ru/media/article/1773537/", "title": "Артур  
Смолянинов: «Я сомневался, что смогу сыграть ангела»", "body": "2 января в  
российский прокат вышла романтическая комедия Веры Сторожевой „Мой
```

парень — ангел“, главные роли в которой исполнили Артур Смольянинов и Анна Старшенбаум. Мы подготовили небольшой видеосюжет с участием создателей картины...Студентка Саша с большим трудом верит в чудеса. Ангелу Серафиму приходится приложить немало усилий, чтобы доказать ей, что ангелы существуют. Но он не учел одного: если девушка тебе поверит, она, скорее всего, тебя полюбит.\n\n\n\n\n\n\n\nАвтор: Дарико Цулая", "comments": ""}.

Индекс

Готовый индекс хранится в четырёх файлах:

- **docs_id.data** (42 Мб)

Файл служит для отображения индекса документа (doc_id) в его текстовое представление в файлах *.txt. Поддерживается переменная длина пути до файлов с документами.

- **terms.data** (54 Мб)

Файл служит для хранения словаря с терминами и ссылок (смещений) на файл с словопозициями и координатами. Поддерживается переменная длина термина. Термины упорядочены в лексикографическом порядке.

- **postings_list.data** (2.68 Гб)

Файл служит для хранения словопозиций и координат терминов в документе. Словопозиции упорядочены по возрастанию идентификаторов документов.

Структура

n_terms						
n_docs[0]	doc...doc	freq...freq	offset...offset	begin...begin	end...end	begin...
n_docs[1]	doc...doc	freq...freq	offset...offset	begin...begin	end...end	begin...
...						

Описание полей

Название	Тип	Назначение
n_terms	uint	Число терминов в корпусе/ число списков словопозиций и координат
n_docs[0],..., n_docs [n_terms -1]	uint	Число словопозиций для конкретного термина

doc[i][0],..., doc[i][n_docs[i]-1], i = 0...n_terms-1	*int	Вектор идентификаторов документов, в которых встречается термин (слопозиции)
freq[i][0],..., freq [i][n_docs[i]-1], i = 0...n_terms-1	*int	Вектор частот вхождений терма в документы
offset[i][0],..., offset[i][n_docs[i]-1], i = 0...n_terms-1	*uint	Относительные смещения до координат. Таким образом, если понадобятся координаты i-терма в j-м документе, то сначала выполнится смещение до нужной строки в таблице <i>posting_list.data</i> с помощью смещений в словаре. Затем, зная число документов n_docs[i], можно быстро считать freq и offset, не читая остальные данные. Далее, с помощью offset выполняется смещение до блока, в котором находятся координаты begin...begin, end...end терма в документе. Их количество равно значению freq.
begin... begin	*int	Координаты начал термина в документе. Координаты измеряются в символах от начала документа
end...end	*int	Координаты концов термина в документе. Координаты измеряются в символах от начала документа

- **tf.data** (939 Мб)

Файл служит для быстрого получения компонент документа, как вектора в пространстве терминов. Он нужен для быстрого ранжирования на основе косинуса между вектором запроса и вектором документа. См. подробности в ЛР по ранжированию.

Алгоритм цитатного поиска

Рассмотрим алгоритм на примере поиска “что где когда” \ 5. Сначала переведем расстояние в словах в расстояние в символах. Для этого сделаем предположение, что среднее число пробельных символов составляет 2

символа, а средняя длина слова – 7 символов. Тогда 5 требуемых слов переведётся в:

2 (пробел) + 3 (где) + 2 (пробел) + 5 (когда) + 2 + 7 + 2 + 7 + 2 + 7 = 32 символа.

Расстояние между словами считается как расстояние между их концами в символах.

Далее, получим словопозиции и координаты для терминов “что” и “где” из файлов индекса. Запустим обычный алгоритм пересечения списков словопозиций, но будем дополнительно отсекаать документы, если они не удовлетворяют требованиям близости терминов. Для этого создадим цикл по всем координатам терма “что” и, пользуясь отсортированностью координат терма “где”, с помощью *бинарного поиска* найдём ближайшие вхождения и проверим их на близость. Сохраним координат цитаты (“что где”). Далее, пересекаем получившийся список словопозиций со списком для терма “когда” аналогичным образом, сравнивая близость терма “когда” и цитаты “что где”.

Поиск по корпусу документов

Параметры

“что где когда” / 5

Поиск Загрузить индекс Построить индекс Выполнить слияние

Ссылка	Заголовок	Подробности
http://filmlplace.ru/film/morskaya-politsiya-s...	Морская полиция: Спецотдел	..и новые просто ни о ЧЕМ”, “почему тут уже 2...и!!!!”, “Я сериалы ГДЕ каждую серию событи...л замеч...
http://filmlplace.ru/film/tri-protsenta.html	Три процента	..риption”: “Будущее, ГДЕ мир разделён на две...еер, но когда нету ЧТО смотреть то можно г...ыгршл”,
http://filmlplace.ru/film/krepkaya-bronya.html	Крепкая броня	..ак Василий Русаков. ГДЕ то совсем еще дети, очно, сверковь))) ЧТО касается санитарок крылья
http://filmlplace.ru/film/intervyu-s-putinyim...	Интервью с Путиным	.. расклада ситуации.ГДЕ бы найти?))))”, “ЧТО за видео В.В. показа...ведут... Где найти КОГДА скачать по
http://filmlplace.ru/film/lozh-vo-imya-nauki...	Ложь во имя науки	.. голове. Слепил все ЧТО можно в одну кучу,вергнута), и все! ГДЕ опровергнуто? КОГДА, кем? А
http://filmlplace.ru/film/ischeznuvshij.html	Исчезнувший	.. на середине ленты. ЧТО? ГДЕ? КОГДА? Загадка...”, “Англ..
http://filmlplace.ru/film/nachalo-1.html	Начало	..ь в мир сновидений, ГДЕ значительное место ...я - то жалеть не за ЧТО? “Фильм не плохо.лемы ил
http://filmlplace.ru/film/kosmicheskaya-odis...	2001 год: Космическая одиссея	..ый район галактики, ГДЕ, как считается, скр...ми, а на счёт того, ЧЕГО кому смотреть, этопередаче
http://filmlplace.ru/film/vivarij.html	Вивариум	..из описания фразу: «ГДЕ им предстоит воспит...ейные ценности, или ЧТО то совершенно друго...о то
http://filmlplace.ru/film/ssd-strashliki-sovets...	С.С.Д.	.. пионерском лагере, ГДЕ предприимчивые теле...едущего программы “ЧТО? Где? КОГДА?””, “Ужас
http://filmlplace.ru/film/oprometchivij.html	Опрометчивый	..я от Гарварда, но в ЧЕМ же опрометчивость?о в школьные “ЧТО? Где? КОГДА?”, вряд ли кого-то...
http://filmlplace.ru/film/drugoj-mir-voinyi-kr...	Другой мир: Войны крови	..и: Skroty пишет: В ЧЕМ может быть дело? зн...ледовательности, что, ГДЕ, КОГДА, откуда...Но я, собст...
http://filmlplace.ru/film/zori-zdes-tihie.html	А зори здесь тихие...	..мяти и культуры. И, ЧТО характерно, за росс...сисек, сцена в бане ГДЕ девушки купались эт...ный: “А
http://filmlplace.ru/film/vnutrennyaya-imper...	Внутренняя империя	..на уже не понимает, ГДЕ заканчивается кинора в избытке, из-за ЧЕГО мозг начинает устав...гда так
http://filmlplace.ru/film/rasputin.html	Распутин	..уже Царское Село... ЧЕГО, почему, как... Под...л время!”, “Я может ГДЕ то не права, но Дел...н может
http://filmlplace.ru/film/ne-otstupat-i-ne-sd...	Не отступать и не сдаваться 2: Штурмовое предун...	..J. “comments”: [“Ну ЧТО сказать... Мда...Вс...ахватить Азию. ЧТО? ГДЕ? КОГДА? США - Захватила Фил...
http://filmlplace.ru/film/elitnyj-otryad.html	Элитный отряд	..ой власти в стране, ГДЕ царит бедность. В ф...и, но все же не то, ЧТО хотел увидеть - Ест...Вы что наро...
http://filmlplace.ru/film/ya-vizhu-ya-vizhu.ht...	Я вижу, я вижу	..ак и не получилось, ЧТО ж, сама виновата.ПС: Бесит КОГДА весь фильм тебя вод...ам в передаче “Чт
http://filmlplace.ru/film/goldnyie-igry-sojka...	Голодные игры: Сойка-пересмешница. Часть II	..недоумение, мол для ЧЕГО всё это делалось? Н...ончил на том месте, ГДЕ Китнисс охотится за...л вое
https://www.kinopoisk.ru/media/news/1815...	Премьеры США — 10 февраля	..оступа „Кейтэун“, ГДЕ герои Дензела Вашинг...реломлением цветов, ЧТО лишь добавляет прои...ла
https://www.kinopoisk.ru/media/article/2794...	Дизлайк, репост: Новые отечественные хорроры	..или в тюрьму, после ЧЕГО женщина сошла с ума...нает кровавое «ЧТО? Где? КОГДА?», и за каждый в
https://www.kinopoisk.ru/media/article/4001...	От Гая Фокса до Пугала: Самые знаменитые маск...	..адаёт в концлагерь, ГДЕ подвергается медици...енно похожую на ту, ЧТО была на Джокере воОу...?
https://www.kinopoisk.ru/media/article/3143...	Можем повторить: Как наши фильмы и сериалыа для АВС телеигру «ЧТО? ГДЕ? КОГДА?», а сейчас вместе ..
https://www.kinopoisk.ru/media/article/2250...	Джефф Дэниелс: «Люблю и разделяю точку зрен...	..НВО отвечает за то, ЧТО производит, и ты не...вных событий — что, ГДЕ, КОГДА и почему. Но осталь...
https://www.kinopoisk.ru/media/news/4002...	DeerFae дня: Заставка «Друзей», только вместоДрузя из передачи «ЧТО? ГДЕ? КОГДА?». В результате пол...
https://www.kinopoisk.ru/media/article/4004...	«Думаю, как все закончить» Чарли Кауфмана: Уду...	..ливый хоррор о том, ЧЕГО хотят мужчины”, “Бо...стического хоррора, ГДЕ законы энтропии неп...ик
http://filmlplace.ru/film/varg-veum-9-sputni...	«Золотой глобус» больше не золотой. Что о скан...	.. больше не золотой. ЧТО о скандале с премие...глобуса». В России, ГДЕ многие премии облад... «ЧО
http://filmlplace.ru/film/suhaya-kost.html	Варг Веум 9: Узы смерти	..одалеку от Бергена, ГДЕ произошло кровавое ...одолжение будет?!!! КОГДА?!!”, “Тогда смотри ЧТО poi
http://filmlplace.ru/film/ohota-na-monstra.h...	Сухая кость	..стречи в пустыне... ЧТО?ГДЕ?КОГДА? диалоги сумбурные, ..
http://filmlplace.ru/film/pomni.html	Охота на Монстра	..из Простоквашино - ЧТО ГДЕ происходит и КОГДА это все закончится?..
http://filmlplace.ru/film/tekst.html	Помни	..красно помнить все, ЧТО было до убийства, н...ыпается с вопросом «ГДЕ я?». Он идет по сле...ов пон
http://filmlplace.ru/film/idealnyj-muzhchina...	Текст	..камнем, утопил бы, ЧТО угодно. Но не так г... Люблю такой жанр, ГДЕ и актёру есть что с...рограмм
http://filmlplace.ru/film/ledenets.html	(НЕ)идеальный мужчина	..что то в нем есть о ЧЕМ нужно задуматься...в восторге от того, ГДЕ побывал твой ключ.тут, Чего
http://filmlplace.ru/film/robot-po-imeni-cha...	Леденец	..я... отсюда и имеем ЧТО имеем. От меня 7 из... “Фильм-перевертыш, ГДЕ меняются местами по... оби
http://filmlplace.ru/film/dvadcat-vosem-pan...	Робот по имени Чаппи	..ие, что как будь то ЧЕГО то не хватает. Как ...осищение. А фильмы ГДЕ показаны простые, и...напок
https://www.kinopoisk.ru/film/396906/	28 панфиловцев	..тация. Единственное ЧТО резануло взгляд это... понятие из фильма? ГДЕ шли сражения понятн... тани
https://www.kinopoisk.ru/film/43870/	Прорвёмся! (2006)	..щают внимания ни на ЧТО, кроме парашютного ...лом.\n\nЖанр. Экшн? ГДЕ?!\nПриключения, бое...
https://www.kinopoisk.ru/film/715299/	Служили два товарища (1968)	..го движения - самое ЧТО ни на есть декадент... Подписями, что, ГДЕ и КОГДА происходит, чтобы п...
https://www.kinopoisk.ru/film/1188952/	Пять лет и один день (2012)	..ржки, единственное, ЧТО есть у женщины — же...остая описывать что ГДЕ и КОГДА, потому что тот к
https://www.kinopoisk.ru/film/15061/	Прощание (2019)	..буде для рефлексии. ЧТО ж, кажется, “Прощан... взаимоотношениями, ГДЕ на фоне основного с...то
https://www.kinopoisk.ru/film/15061/	Смертельное оружие (1987)	..и детьми. Не знаю, КОГДА в последний раз люд...на работе. И знаешь ЧТО? На самом деле это ...ето

1/4

Параметры

Ссылка	Заголовок	Подробности
https://www.kinopoisk.ru/media/news/4002...	DeepFake дня: Заставка «Друзей», только вместо в...	...актеров — Александр ДРУЗЬ", "body": "Поместив..... Друзя из передачи «ЧТО? ГДЕ? КОГДА?». В результате
https://www.kinopoisk.ru/series/656730/	Что? Где? Когда?	..656730/", "title": "ЧТО? ГДЕ? КОГДА?", "alternative_tit..... Друзь", "Александр ДРУЗЬ", "Василий Уткин"]...
https://www.kinopoisk.ru/film/480702/	Параллельная дорога (1962)	.. Отвечает Александр ДРУЗЬ!»\n\nНет ничего стр..... скучной передачи «ЧТО? ГДЕ? КОГДА?», которую не

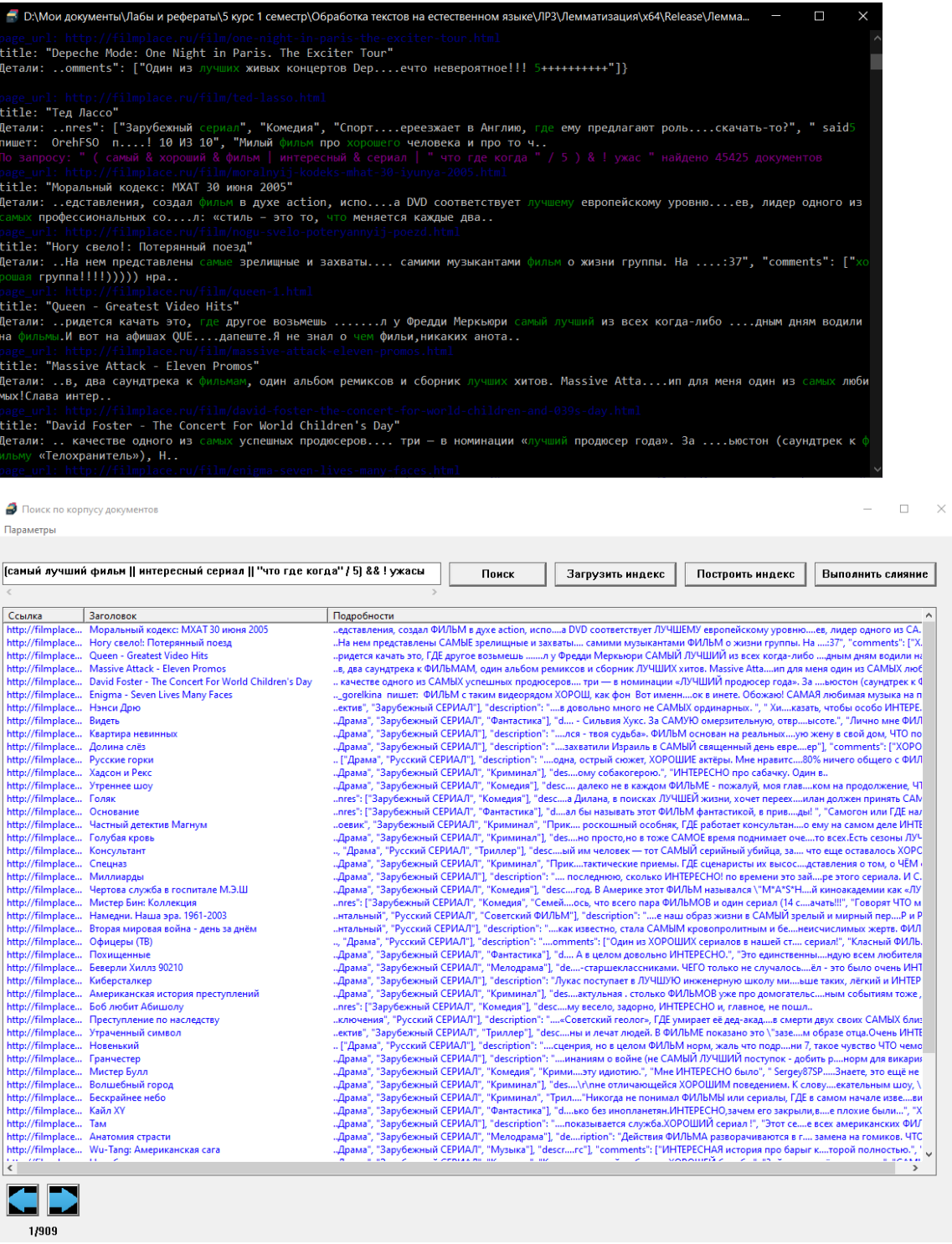


1/1

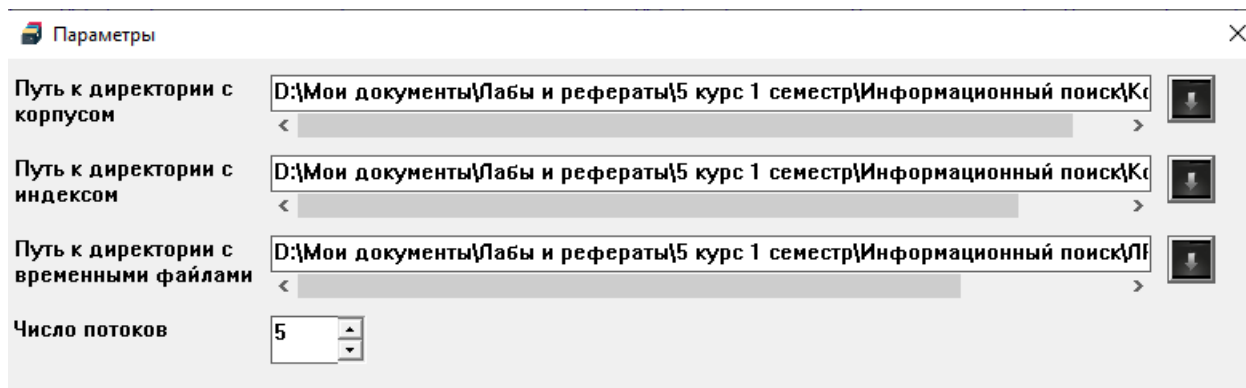
Интерфейс

Программа была реализована в двух интерфейсах: консольном и оконном.

Для последнего использовалась библиотека Windows.h (winapi). Примеры ниже:



Параметры поиска:



Параметры

Путь к директории с корпусом D:\Мои документы\Лабы и рефераты\5 курс 1 семестр\Информационный поиск\Кл

Путь к директории с индексом D:\Мои документы\Лабы и рефераты\5 курс 1 семестр\Информационный поиск\Кл

Путь к директории с временными файлами D:\Мои документы\Лабы и рефераты\5 курс 1 семестр\Информационный поиск\Кл

Число потоков 5

Примечание. Внимательный читатель мог заметить наличие лемматизации и сниппетов. Подробности реализации см. в соответствующих ЛР по курсу.

2. Исходный код

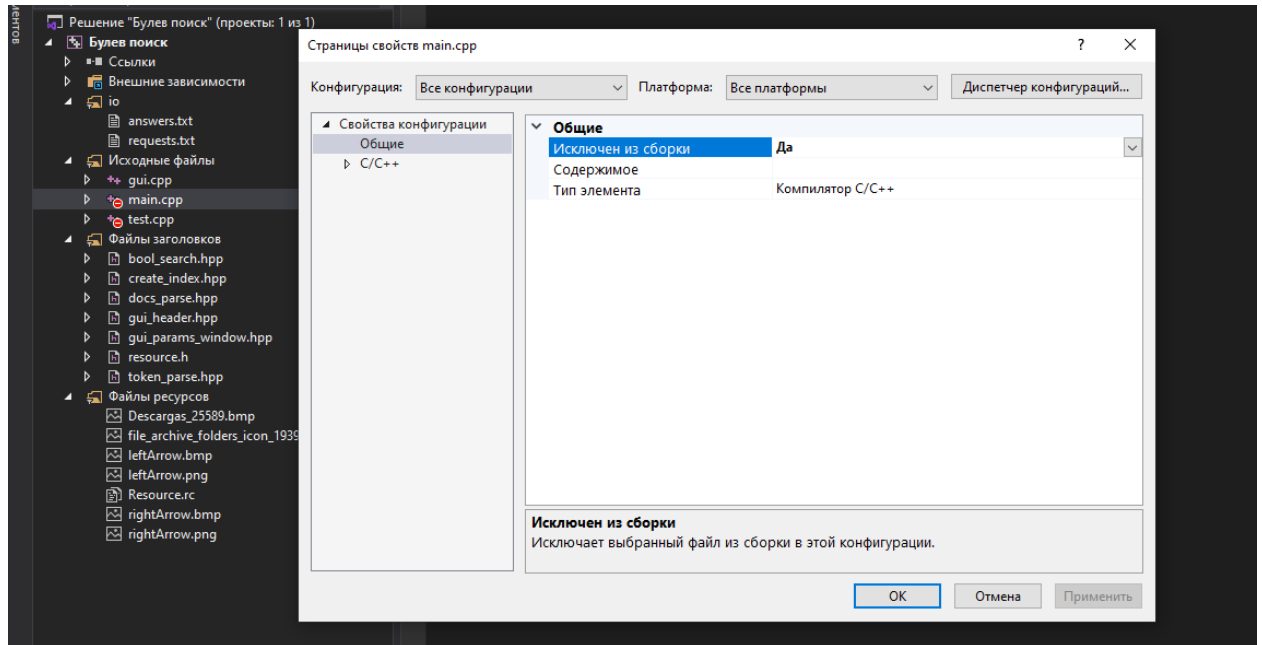
Структура проекта

- include
 - bool_search.hpp (булев поиск)
 - create_index.hpp (создание, чтение индекса)
 - defs.hpp (подключение внешних библиотек, макросы)
 - docs_parse.hpp (извлечение полей из корпуса)
 - gui_defs.hpp (подключение внешних библиотек, макросы, глобальные переменные)
 - gui_params_window.hpp (окно с выбором параметров)
 - quote_search.hpp (реализация цитатного поиска)
 - resource.h (подключение изображений, иконок и прочего)
 - token_parse.hpp (функции для преобразования токенов в термы)
- python
 - lemmatizator.py (лемматизация документа)
 - lemmatizator_setup.py (компиляция lemmatizator.py в exe-файл)
 - request_parse.py (лемматизация запроса)
 - request_parse_setup.py (компиляция request_parse.py в exe-файл)
- io
 - answers.txt
 - requests.txt
- src
 - gui.cpp (точка входа в оконный интерфейс)
 - main.cpp (точка входа в консольный интерфейс)
 - test.cpp (утилита для тестирования программы)
- resources (файлы ресурсов для оконного приложения)

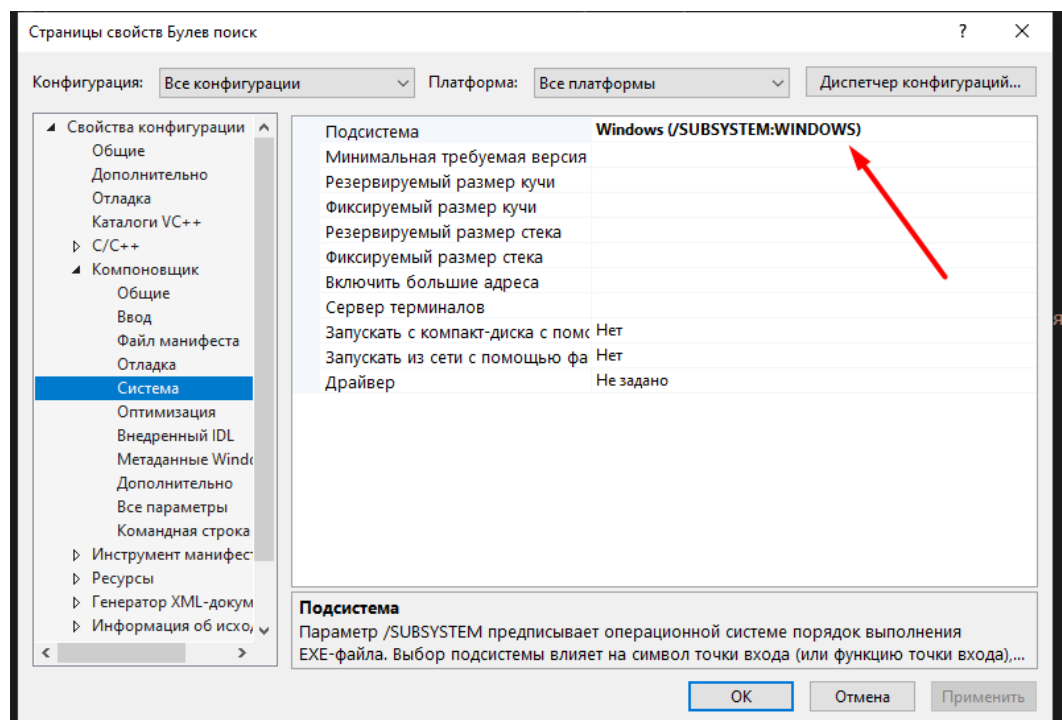
Проект был написан с помощью Microsoft Visual Studio 2019 эксклюзивно для ОС семейства Windows.

Запуск

Переключение между тремя точками входа осуществляется с помощью флага «исключить из сборки»:



Не забудь при переключении между консольными и оконными приложениями менять подсистему в настройке проекта:



Консольное приложение поддерживает флаги запуска:

- -i 'абс. путь к корпусу'
- -o 'абс. путь к индексу'
- -t 'абс. путь к директории с блочным индексом'
- -p кол-во_процессов_для_распараллеливания
- -create : создать блочный индекс
- -merge : выполнить слияние блочного индекса
- -clear : очистить папку с временными файлами после слияния
- -search : выполнить поиск

Утилита тестирования поддерживает следующие ключи:

- -i 'абс. путь к корпусу'
- -o 'абс. путь к индексу'
- -n1 число_запросов
- -n2 длина_запросы_в_термах

Пример создания индекса из корпуса:

```
$ ./Булев индекс.exe -merge -create -clear -i "D:\Мои документы\Лабы и рефераты\5 курс 1 семестр\Информационный поиск\Корпус" -o "D:\Мои документы\Лабы и рефераты\5 курс 1 семестр\Информационный поиск\Корпус_index" -t "D:\Мои документы\Лабы и рефераты\5 курс 1 семестр\Информационный поиск\ЛР3\Булев индекс\tmp"
```

```
[INFO] Создание индекса для блоков
```

```
[INFO] Thread 0 processing block 1/13 : D:\Мои документы\Лабы и рефераты\5 курс 1 семестр\Информационный поиск\Корпус\films13.txt
```

```
[INFO] Thread 1 processing block 2/13 : D:\Мои документы\Лабы и рефераты\5 курс 1 семестр\Информационный поиск\Корпус\films9.txt
```

```
[INFO] Thread 2 processing block 3/13 : D:\Мои документы\Лабы и рефераты\5 курс 1 семестр\Информационный поиск\Корпус\films10.txt
```

```
[INFO] Thread 3 processing block 4/13 : D:\Мои документы\Лабы и рефераты\5 курс 1 семестр\Информационный поиск\Корпус\films12.txt
```

```
[INFO] Block 1 has 232216 terms
```

```
<...>
```

```
[INFO] Block 10 has 597482 terms
```

[INFO] Block 11 has 669497 terms
 [INFO] Block 12 has 921883 terms
 [INFO] Block 13 has 1383148 terms
 [INFO] Создание очередей термов: 13 блок из 13
 [INFO] Слияние docs_id: 13 блок из 13
 [INFO] Слияние слопозиций термов
 [INFO] Осталось термов: 0
 [INFO] Очистка временных файлов
 [INFO] Общее число термов в словаре = 2809203

Время выполнения = 145,5 сек, размер корпуса = 2,899 Gb, документов = 186109

Средняя скорость на документ = 0,782 ms

Средняя скорость на килобайт = 0,048 ms

Пример работы тестировщика:

\$./Булев индекс.exe -n1 200 -n2 5 -search -i "..\..\Корпус" -o
 "..\..\Корпус_index" -t "tmp" <io/requests.txt >io/answers.txt

```
[INFO] Чтение термов
[INFO] Загружено 2809203 термов
Номер запроса|Документов|Время, ms|Запрос
1||0||47363||!завались-и || полубогов-наставников && !вкачал && unagi-sandжейпоп && омежному
2||97721||41163||трагичных || !draculiusваха || !фильм-эссе || !васантабалан || завертелись
3||0||49514||!пахнетыаха-kui && !севадоархеолога && отцом-священнослужителем || anime-9595-amanatsu && !акирыдальше
4||7||38243||конском && !ragamuffin || shackleton || галапаго || грира
5||142308||49725||!австрийско-американскогофильм || !lloydмато && !связуяща || разгильдяи-мужья || !лапирь
6||117798||55409||тщательно-тщательно || !заступом || отпечатавшиеся || экзотике && !шафтена
7||140308||34997||!perfectionist || !размытое || !rosseti && okayмлении || feuerbach
8||0||32456||отчаянней && takahashi && фыркают && теоретике && !uc-qsijjjsnjepvcowid3sw
9||0||35578||задутся || lasky && зятанутой-перетанутой || !marey && !изображению
10||0||55991||!невежля && 24-ому && саморазрушается && djunhorдекто || евро-1988
11||0||53903||temoins || timas || !diego3000представь && милый1920x1062 && !siliang
12||0||36817||!backpacking || !москвич-2141 || проводящих && костяными || меллерукуине
13||0||45504||!репортер-плэйбой && !200icq || керзон && натанутьтоксичный || !voda-i-ogon
14||0||43348||!таксандрию || минами-ловушками && братья-близнецы-они || yagakimi || !bo-ri
15||0||47784||!умничкаоднозначно && !момышулы && локдока && !predannaya-krasota && kitamuraэри
16||156072||42611||спецэффекты || !202605 || !баллы || !тщеславно-эгоцентричного || !наркодилеров-идиотов
17||0||56536||!незаземленным || !чаегиоликами && !артер850x1280850x1280850x1280850x1280850x1280850x1280797x1200
18||0||53416||попстрима || скалящая && !шер-хан && !окты && !дюбу
19||0||34299||суперзлодеях || расшалившиеся || !гинестра || veshi && !240088-obsuzhdenie-anime
20||0||40235||сторонюе && разспойлерувидел && !эротико-романтический && !кот-робинзон && тушилпирогами
21||0||49357||!пазрываюсь && !спойлеротрубывать || дора && существующую && !пслоу
22||130248||30793||!28223-death-parade && !милотыкогда && !самурая-террориста || !приветствовала && !стопроцентный
23||48145||51458||!tokotokos && !anime-6444-tegamibachi && !felicitas || детское || !прошманда
24||0||51957||!friendвпрочем && !социологами && первопроходцу && !бэрона || атмосферу-ретро
25||130000||40100||!сая || !изм-тукерто || !василист || !экикаваса || !скаваса
```

```

191|| 0|| 36098|| пейзажный || клочатые && полинялых && !плагиат-подделка && переколбасил
192|| 1152|| 49506|| !704920 && фармёжки || what || наклеен || вкушали
193|| 0|| 40179|| !франском || скромной && патмор && прог-рок-группы || трёхвалентный
194|| 134339|| 72172|| боала || !бееестранная && !ботан-ботан && momose || !янетти
195|| 0|| 42298|| !неозыданно || убивавшей && !нижерадзе || !k32 && хадисы
196|| 534|| 44258|| !мужик-мачо && !архетипчик && !zmopes && !срабатывание && !87489-toki-wo-kakeru-shoujo
197|| 8936|| 44787|| !злодея-мессии || !закончились && !вот1920x10801920x1080пешипгатакое && аналогayoutube && !клензендорфа
198|| 0|| 52473|| храбрец-красавчик && !дурачащих && фильм-интрига || !тупачка || !евнухом
199|| 31139|| 56695|| !amarelo && !лисапед || !markedoneороче || !ноября704x995 && !простынейтак
200|| 0|| 323291|| !сюжетно-воспитательные || минут06 && !sonambula || !паукоубиения && !ухxxxxxxxxxxдождался
Всего времени 11059715 мс, в среднем на запрос 55299 мс

```

3. Выводы

Размер индекса	3.69 Гб
Время построения индекса	4 часа (отечественный лемматизатор natasha работает очень медленно!)
Скорость индексации	268.7 Кб / сек
Время выполнения поисковых запросов	≈2 сек. Одну из них занимает вызов natash'и для лемматизации терминов из запроса.

Пример долго выполняющегося запроса (30 сек):

[фильм || сериал || мультфильм]

```
По запросу: " фильм | сериал | мультфильм " найдено 162370 документов
page_url: http://filmlplace.ru/film/jimmy-page-and-robert-plant-no-quarter-unledded.html
title: "Jimmy Page & Robert Plant - No Quarter - Unledded"
Детали: ..Некоторые фрагменты фильма были сняты в национ..
page_url: http://filmlplace.ru/film/slava-bobkov-i-gruppa-altaj-xxi-pereplavlennyiy-v-kolokol.html
title: "Слава Бобков и группа Алтай XXI: Переплавленный в колокол"
Детали: ..в Санкт-Петербурге. Фильм писали почти 7 часо..
page_url: http://filmlplace.ru/film/primal-fear.html
title: "Primal Fear - 16.6 All Over The World"
Детали: ..нным художественным фильмом, который незадолго ..
page_url: http://filmlplace.ru/film/madonna-the-re-invention-world-tour.html
title: "Madonna - The Re-Invention World tour"
Детали: ..ачный опыт. Музыка, фильмы, танцы – все, что я..
page_url: http://filmlplace.ru/film/sergej-lyubavin-nash-esenin.html
title: "Сергей Любавин - Наш Есенин"
Детали: .."], "description": "Фильм был снят 17 февраля..
page_url: http://filmlplace.ru/film/the-rolling-stones-the-stones-in-the-park.html
title: "The Rolling Stones: The Stones In The Park"
Детали: ..просмотра трудно. В фильме есть нечто гипнотич..
page_url: http://filmlplace.ru/film/the-rolling-stones-live-at-the-max.html
title: "The Rolling Stones: Live At The Max"
Детали: ..зный проект – сняли фильм, посвященный гастро..
page_url: http://filmlplace.ru/film/moralnyiy-kodeks-mhat-30-iyunya-2005.html
title: "Моральный кодекс: МХАТ 30 июня 2005"
Детали: ..едствления, создал фильм в духе action, испо..
page_url: http://filmlplace.ru/film/kontsert-eltona-dzhona-v-sssr-v-1979-godu.html
```

Причины: длинный список словопозиций-документов, ранжирование по методу косинусов (20 сек), создание сниппетов (8 сек).

В ходе выполнения лабораторной работы я научился выполнять координатный поиск для коллекции документов. Познакомился с winapi. Научился разрабатывать оконный интерфейс на языке C/C++. Научился выполнять python-скрипты внутри C++.

Литература

- [1] Кристофер Д.Маннинг, Прабхакар Рагхаван, Хайнрих Шютце. Введение в информационный поиск. 2020, изд. Вильямс.