

**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
https://www.ki...	Что? Где? Когда?	..656730/, "title": "ЧТО? ГДЕ? КОГДА?", "alternative_tit...
https://shikimo...	Shinkansen Henkei Robo Shinkalion: Mirai kara Kita S...	..tube", "youtu.be", "ЧТО то ни ГДЕ не нашёл. Блин КОГДА будет. Если кто зна...
https://www.ki...	Весной (1929)	..нологов и творцов - КОГДА технология становится...адавание загадок' - ГДЕ, ЧТО, как и когда снято ...
https://www.ki...	Большой вопрос	..нственная телеигра, ГДЕ выигрыш зависит нет качества юмора. А ЧТО принесет победу — и...авить на 'Что, Где, КОГДА':\n\nИт...
http://filmplace...	Л.О.Р.Д. Легенда о разорении династий	..ё ляжет. Только при ЧЕМ тут Один и алхимики...ял сюжет кто кого и ГДЕ куда.", "Смотрел см...то где происходит и КОГДА это за...
https://www.ki...	Пять лет и один день (2012)	..ржи, единственное, ЧТО есть у женщины — же...остях описывать что ГДЕ и КОГДА, потому что тот кто...
http://filmplace...	Исчезнувший город	..т направится туда, ГДЕ теперь лежат руины...ема не раскрыта ", "ЧТО вообще происходит и...лей загадок ЧТО ГДЕ КОГДА и я пр...
https://www.ki...	DeepFake дня: Заставка «Друзей», только вместоДрузя из передачи «ЧТО? ГДЕ? КОГДА?». В результате пол...
http://filmplace...	Варг Бейм 9: Узы смерти	..одалеку от Бергена, ГДЕ произошло кровавое ...одолжение будет?! КОГДА?!", "Тогда смотри ЧТО попроще, где думать...
https://shikimo...	Kemono to Chat	.. бывает XDТолько... ГДЕ они взяли эти станн... тут больше и не на ЧТО, либо будет интерес...Ragapoid, что? где? КОГДА? в любом...
https://www.ki...	Сексуальный вор (1973)	..суального вора» все ЧТО мог и поэтому фильм...ный программы «Что? ГДЕ? КОГДА?»: какая часть женс...
http://filmplace...	Сухая кость	..стречи в пустыне... ЧТО?ГДЕ?КОГДА? диалоги сумбурные...
https://www.ki...	Приключения охотника на драконов (2010)	..лАктёры.\n\nПервое, ЧТО очень бесит в этом ...силумом или Тромой, ГДЕ в ролях зачастую иг...ым только один раз, КОГДА мал...
https://www.ki...	Случайные связи	..есь, пока они живы. ЧЕМУ? Валить дурака хотя...на на подобии «Что? ГДЕ? КОГДА?», но по приколу. И...
http://filmplace...	Исчезнувший	.. на середине ленты. ЧТО? ГДЕ? КОГДА? Загадка...", "Англ...
http://filmplace...	Интервью с Путиным	.. расклада ситуации.ГДЕ бы найти?!!!!", "ЧТО за видео В.В. показа...ведут... Где найти КОГДА сказать помогите...
https://www.ki...	Рождённые в СССР. Четырнадцатилетние (ТВ, 1998)	..к интересно узнать, ЧТО же происходит с реб...ывают бомбоубежище, ГДЕ будут прятаться. Бо...? - "Свобода - это КОГДА что хо...
https://www.ki...	Морская полиция: Спецотдел	..и новые просто ни о ЧЕМ", "почему тут уже 2...!!!!!!", "Я сериалы ГДЕ каждую серию событи...л замечательный, ну КОГДА же буд...
http://filmplace...	Не отступай и не сдавайся 2: Штормовое предуп...	..), "comments": ["Ну ЧТО сказать... Мда...Вс...ахватить Азиро. ЧТО? ГДЕ? КОГДА?США - Захватила Фил...
http://filmplace...	Опрометчивый	..я от Гарварда, но в ЧЕМ же опрометчивость? ...о в школьные \"Что? ГДЕ? КОГДА?\", вряд ли кого-то...
https://www.ki...	Z/Rex: The Jurassic Dead (2017)	..тсутствие динамики, ЧТО надо понимать, для ...ть на телеигру 'Что ГДЕ КОГДА', либо Анатолию Вас...
http://filmplace...	Я вижу, я вижу	..ак и не получилось, ЧТО ж, сама виновата.ПС: Бесит КОГДА весь фильм тебе вод...ам в передаче \"Что?ГДЕ?Когда?\"Ибо на прот...
https://www.ki...	Охота на Монстра	..из Простоквашино - ЧТО ГДЕ происходит и КОГДА это все закончится?...
https://www.ki...	Параллельная дорога (1962)	.. скучной передачи «ЧТО? ГДЕ? КОГДА?», которую не мешал...
https://www.ki...	Джефф Дэниелс: «Люблю и разделяю точку зрен...	..НВО отвечает за то, ЧТО производит, и ты не...вных событий — что, ГДЕ, КОГДА и почему. Но остаьл...
https://www.ki...	Дикей хмель (1985)	..на обуюную фабрику, ГДЕ работала ее мать. З...ние на вопросы, кто, ЧТО, где, КОГДА, с кем, чем и не ин...
https://www.ki...	Королева варваров 2: Сражение за скипетр Аркар...	..тельно талантливая, ЧТО не скажешь о других...ор. Непонятно, что, ГДЕ, КОГДА должно произойти. Ф...
https://shikimo...	Грязная парочка / Dirty Pair	..и в главной роли. В ЧЕМ то даже похожа на ... есть тип сложной, ГДЕ чтоб понять шутку и ...м:Поясни: что, где, КОГДА?Две ум...
https://www.ki...	Прорвёмся! (2006)	..щают внимания ни на ЧТО, кроме парашютного ...лом.\n\nЖанр. Экшн? ГДЕ?\n\nПриключения, бое...ье, кто кого, где, КОГДА и...
http://filmplace...	Внутренняя империя	..на уже не понимает, ГДЕ заканчивается кино ...ра в избытке, из-за ЧЕГО мозг начинает устав...гда так и не узнал, КОГДА начался...
https://www.ki...	Дельфин: История мечтателя (2009)	..ешество по океану, ГДЕ его подстерегают ра...гадаться. Может, в 'ЧТО? ГДЕ? КОГДА? поиграем? И еще м...
https://www.ki...	Выкрутасы (1987)	..я уютное гнездышко, ГДЕ ему будет комфортно...мы можем заниматься ЧЕМ угодно и КОГДА угодно, где нам ник...
https://shikimo...	«Думаю, как все закончить» Чарли Кауфмана: Уду...	..ливый хоррор о том, ЧЕГО хотят мужчины", "bo...стического хоррора, ГДЕ законы энтропии неп...икторине «Что? ГДЕ? КОГДА?»)
https://shikimo...	Pumpkin Scissors	..ня, просто убивает ЧТО ГДЕ как КОГДА почему зачем кто ко...
https://shikimo...	Suisei no Gargantia: Meguru Kouro, Haruka	..ею вздыхая) "Начало ЧЕГО? Если овалек то нет... возможно подскажете, где уже есть перевод ..." @probox_fly, что? ГДЕ? Когда?",
https://www.ki...	Премьеры США — 10 февраля	..оступа 'Кейптаун', ГДЕ герои Дензела Вашинг...реломлением цветов, ЧТО лишь добавляет прои...ла: кто, что, где и КОГДА дела...
http://filmplace...	Леденец	..я... отсюда и имеем ЧТО имеем. От меня 7 из... "Фильм-перевертыш, ГДЕ меняются местами по... общем мне нравится КОГДА в...
https://www.ki...	Дизлайк, репост: Новые отечественные хорроры	..или в тюрьму, после ЧЕГО женщина сошла с ума...нает кровавое «Что? ГДЕ? КОГДА?», и за каждый непр...
https://www.ki...	Курить/Не курить (1993)	..тельной условности, ГДЕ возможно все ЧТО угодно, как угодно и КОГДА угодно, вслед за др...
https://www.ki...	Автоответчик: Удаленные сообщения (2010)	..иноПоиска в поисках ЧЕГО бы такого глянуть ...е недорогие студии, ГДЕ можно с минимальным... тут кого, где чем, КОГДА и за...
https://www.ki...	Призрак Красной реки (2005)	..совы совсем не те, ЧЕМ кажутся. Ребят, бро...Вуд. В тех местах, ГДЕ она не совсем дотяг...осы, а именно: Где? КОГДА? Куда? Отк...

1/4

Поиск по корпусу документов

Параметры

"что где когда" / 5 && другъ

Поиск

Загрузить индекс

Построить индекс

Выполнить слияние

Ссылка	Заголовок	Подробности
https://www.ki...	Что? Где? Когда?	..656730/", "title": "ЧТО? ГДЕ? КОГДА?", "alternative_tit.... Другъ", "Александр ДРУЗЬ", "Василий Уткин"],...
https://www.ki...	DeerFake дня: Заставка «Друзей», только вместо в...	..актеров — Александр ДРУЗЬ", "body": "Поместив.... Другъ из передачи «ЧТО? ГДЕ? КОГДА?». В результате пол..
https://www.ki...	Параллельная дорога (1962)	.. Отвечает Александр ДРУЗЬ>\n\nНет ничего стр.... скучной передачи «ЧТО? ГДЕ? КОГДА?», которую не мешал..

1/1

Интерфейс

Программа была реализована в двух интерфейсах: консольном и оконном.

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

```
D:\Мои документы\Лабы и рефераты\5 курс 1 семестр\Обработка текстов на естественном языке\ЛР3\Лемматизация\k64\Re...
title: "Параллельная дорога (1962)"
Детали: .. Отвечает Александр Друзья!.. Нет ничего стр... скучной передачи «Что? Где? Когда?», которую не мешал..
По запросу: " ( самый & хороший & фильм | интересный & сериал | " что где когда " / 5 ) & ! ужас " найдено 45425 докумен
тов
[INFO] Обработка результатов поиска
page_url: http://filmlplace.ru/film/tajnyi-istorii-robin-gud.html
title: "National Geographic: Тайны истории"
Детали: ..: ["Документальные сериалы", "Загадки истории"... может быть, она ни когда не будет разгадана,...ра, пытайс
ь узнать, что стоит за легендами,...к", "comments": ["В фильме говорится о двух по...быть Робин Гудом. Интересный фи
льм"]}]
page_url: https://shikimori.one/animes/1471-city-hunter-2/summaries
title: "Городской охотник 2 / City Hunter 2"
Детали: ..кольно видеть серию где нам дают интригу...ение очень неплохой сериал! Качественная сериа...ет), неплохие
фоны, интересные сюжет (пусть и одно..., атмосфера 80-х. И самое главное главные гер.... дорогой, 30 лет не чего новог
о не придумали...вом сезоне: яркая и хорошая рисовка, прекрасная....ть в ужасный вечер, когда просто хотелось отв..
page_url: http://filmlplace.ru/film/bbc-afrika.html
title: "BBC: Африка"
Детали: ..: ["Документальные сериалы", "Природа и животн....дивительно красивый фильм.Съёмки просто потря..... - но эт
от сериал что то особенное !!!! Р.... Lulu пишет: styorКогда появится, мы обновим....кого есть ", "Очень интересный фил
м!!! А какие пе....тка инфы. Для детей самое то, да и взрослым б
page_url: http://filmlplace.ru/film/tri-vozhdy.html
title: "Три вождя"
Детали: ..я", "Документальные сериалы", "Исторический", "....всеобъемлющего, чем когда бы то ни было ранее....ments": ["
Смотрится интересно, совершенно не нуд..
page_url: https://www.kinopoisk.ru/media/news/3133075/
title: "Андрей Звягинцев поставит остросюжетный сериал для Paramount"
Детали: ..тavit остросюжетный сериал для Paramount", "bo...Оскар" в категории «Лучший фильм на иностранном язык....ория
, которая будет интересна международному зрит....nt — первый случай, когда крупная международн....ы быть в курсе всех са
чих важных новостей из ..
```

Поиск по корпусу документов

Параметры

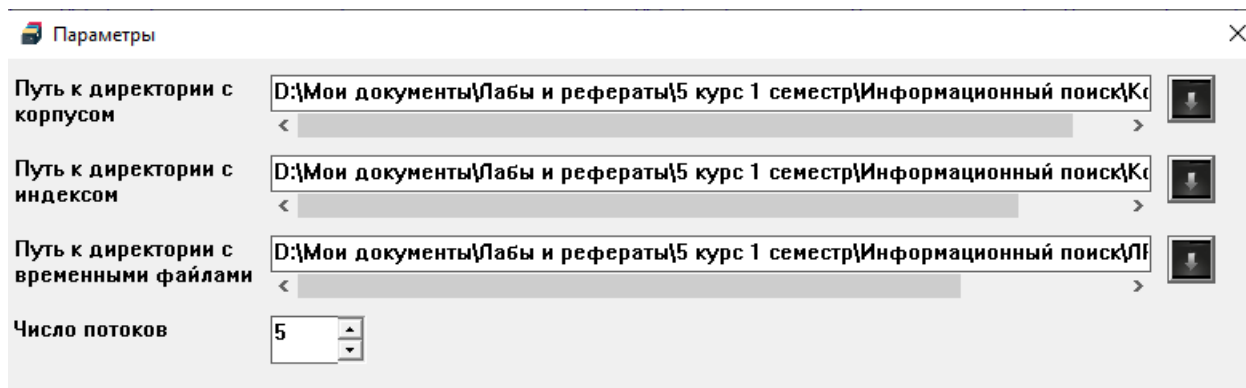
[самый лучший фильм || интересный сериал || "что где когда" / 5] && ! ужасы

Поиск Загрузить индекс Построить индекс Выполнить слайне

Ссылка	Заголовок	Подробности
http://filmlplace...	National Geographic: Тайны истории	..: ["Документальные СЕРИАЛЫ", "Загадки истории"... может быть, она ни КОГДА не будет разгадана,...ра, пытаюсь узнать, что ст...
https://shikimo...	Городской охотник 2 / City Hunter 2	..:кольно видеть серию где нам дают интригу...ение очень неплохой сериал! Качественная сериа...ет), неплохие
http://filmlplace...	BBC: Африка	фоны, интересные сюжет (пусть и одно..., атмосфера 80-х. И самое главное главные гер.... дорогой, 30 лет не чего новог
https://www.ki...	Три вождя	о не придумали...вом сезоне: яркая и хорошая рисовка, прекрасная....ть в ужасный вечер, когда просто хотелось отв..
https://www.ki...	Андрей Звягинцев поставит остросюжетный сери...	page_url: http://filmlplace.ru/film/bbc-afrika.html
http://filmlplace...	Держись, Чарли!	title: "BBC: Африка"
https://www.ki...	Самые аркие и безумные новости года	Детали: ..: ["Документальные сериалы", "Природа и животн....дивительно красивый фильм.Съёмки просто потря..... - но эт
https://shikimo...	Выск: Фильм / Air Movie	от сериал что то особенное !!!! Р.... Lulu пишет: styorКогда появится, мы обновим....кого есть ", "Очень интересный фил
http://filmlplace...	Меллесса и Дюкуи	м!!! А какие пе....тка инфы. Для детей самое то, да и взрослым б
http://filmlplace...	Рождения убивать	page_url: http://filmlplace.ru/film/tri-vozhdy.html
https://shikimo...	Ратнарэк	title: "Три вождя"
https://shikimo...	One Room Special	Детали: ..я", "Документальные сериалы", "Исторический", "....всеобъемлющего, чем когда бы то ни было ранее....ments": ["
http://filmlplace...	Помадные джунгли	Смотрится интересно, совершенно не нуд..
https://www.ki...	Ушел из жизни актер Майкл Паркс	page_url: https://www.kinopoisk.ru/media/news/3133075/
http://filmlplace...	Медвежий упол	title: "Андрей Звягинцев поставит остросюжетный сериал для Paramount"
http://filmlplace...	Новые уполки	Детали: ..тavit остросюжетный сериал для Paramount", "bo...Оскар" в категории «Лучший фильм на иностранном язык....ория
https://www.ki...	10 лучших трейлеров недели: Темные воды, альп...	, которая будет интересна международному зрит....nt — первый случай, когда крупная международн....ы быть в курсе всех са
https://shikimo...	Yoshinaga-sanchi no Gargoyles	чих важных новостей из ..
http://filmlplace...	Лучшие из райских уголков... Австралия, Египет, ...	
http://filmlplace...	Выкуп	
http://filmlplace...	Ангелы Чарли	
https://shikimo...	Гаргульи дома Ёсинэга / Yoshinaga-sanchi no Garg...	
http://filmlplace...	Виллетта	
http://filmlplace...	Секс по дружбе	
http://filmlplace...	Космос: возможные миры	
http://filmlplace...	Утиси	
https://www.ki...	6 лучших трейлеров недели: «Кэндимен», «Расска...	
https://shikimo...	Как и ожидалось, моя школьная романтическая ...	
https://shikimo...	Суббу-Ду! Корпорация заеда	
https://www.ki...	Атака титанов: Оборона Skytree (2017)	
http://filmlplace...	Сын моего отца	
https://shikimo...	Ajin 2nd Season OVA	
https://www.ki...	Раздолл: Фильм (ТВ, 2000)	
http://filmlplace...	BBC: Прогулки по ЮАР	
http://filmlplace...	Деять жизней Хлои Кинг	
https://www.ki...	Нерпосты: поспрос: Путешествие в страну шаман...	
https://www.ki...	Ребека Холл станет террористкой ИРА	
http://filmlplace...	Дело Дойлов	
http://filmlplace...	Благоприятные стеры	

1/909

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

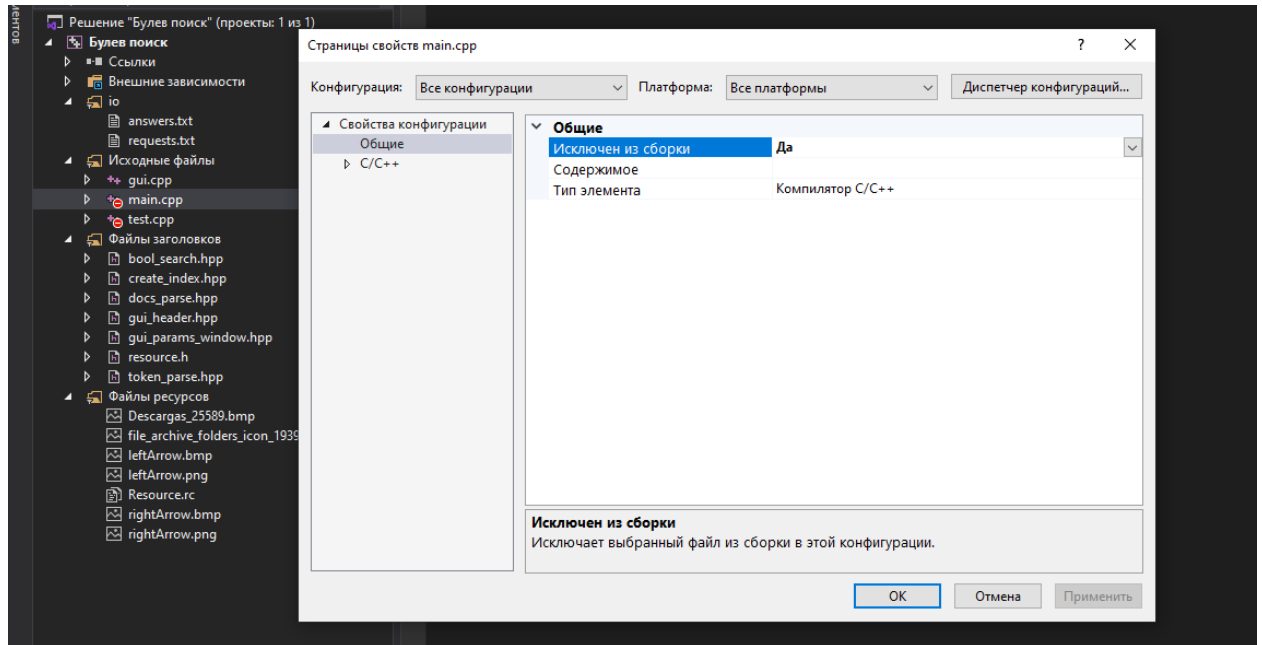
Структура проекта

- 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)

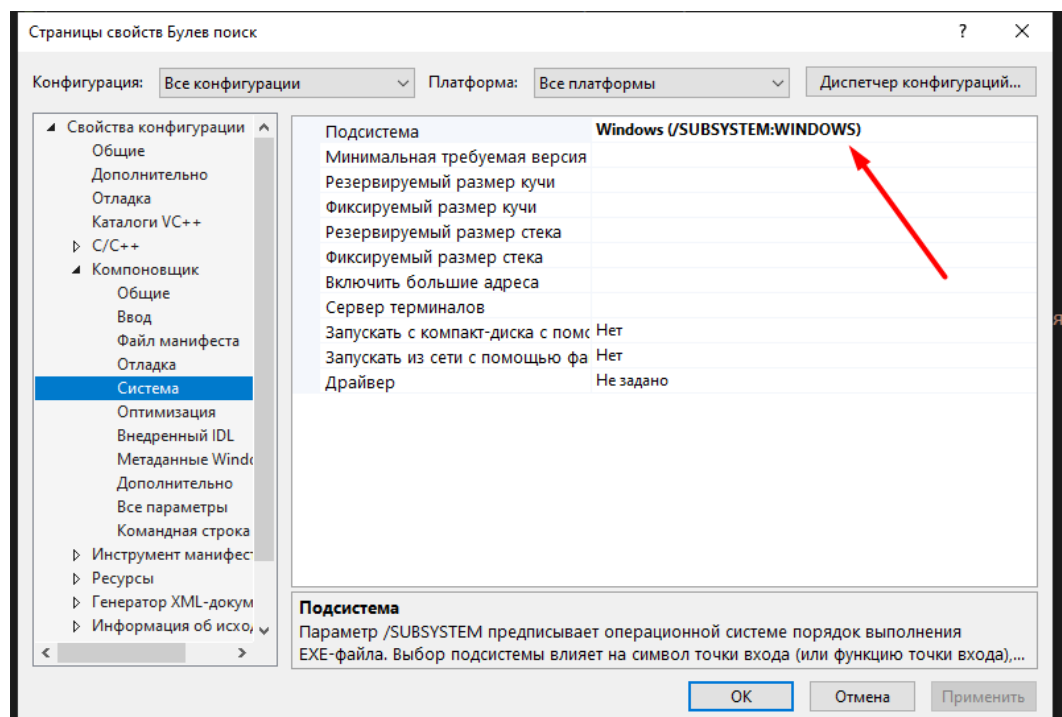
Проект был написан с помощью Microsoft Visual Studio 2019 эксклюзивно для ОС семейства Windows.

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-saпджейпоп && омежному
2|  97721|  41163|  трагичных || !draculiusваша || !фильм-эссе || !васантабалан || завертелись
3|  0|  49514|  !пахнетыаha-kui && !сеvдоархеолога && отцом-священнослужителем || anime-9595-amanatsu && !акирыдальше
4|  7|  38243|  конском && !ragamuffin || shackleton || галапаго || грира
5|  142308|  49725|  !австрийско-американскогофильм || !lloydмато && !связующа || разгильдяи-мужья || !лапирь
6|  117798|  55409|  тцательно-тцательно || !заступом || отпечатавшиеся || экзотике && !шафтена
7|  140308|  34997|  !perfectionist || !размытое || !rosseti && okayмленин || feuerbach
8|  0|  32456|  отчаянней && takahashi && фыркают && теоретике && !uc-qsijj1snjepvcow1d3sw
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 || меллеруukine
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|  !умничкаодназначно && !момышулы && локдока && !predannaya-krasota && kitamuraэри
16|  156072|  42611|  спецэффакты || !202605 || !баллы || !тщеславно-эгоцетричного || !наркодилеров-идиотов
17|  0|  56536|  !незаземлённым || !чаегиоликами && !артер850x1280850x1280850x1280850x1280850x1280850x1280797x1200
18|  0|  53416|  попстрима || скалящая && !шер-хан && !окты && !дюбу
19|  0|  34299|  суперзлодеях || расшалившиеся || !гинестра || veshi && !240088-obsuzhdenie-anime
20|  0|  40235|  сторонуюе && разспойлерувидел && !эротико-романтический && !кот-робинзон && тушилпирогамн
21|  0|  49357|  !пазрываюсь && !спойлеротрубывать || дора && существующую && !пслоу
22|  130248|  30793|  !28223-death-parade && !милотыкогда && !самурая-террориста || !приветствовала && !стопроцентный
23|  48145|  51458|  !tokotokoc && !anime-6444-tegamibachi && !felicitas || детское || !прошманда
24|  0|  51957|  friendвпрочем && !социологами && первопроходцу && !бэрона || атмосферу-ретро
25|  130000|  40100|  !саян || !тань-туканто || !васантабалан || !заканчиваю || !скалящая
```

```

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 && !вот1920x10801920x1080нешипгатакое && аналогayoutube && !клензендорфа
198||      0||      52473|| храбрец-красавчик && !дурачащих && фильм-интрига || !тупачка || !евнухом
199||     31139||      56695|| !amarello && !лисaped || !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
[INFO] Обработка результатов поиска
page_url: http://filmlplace.ru/film/zabyityij-den-rozhdeniya-sbornik-multfilmov.html
title: "Забытый день рождения. Сборник мультфильмов"
Детали: ..ь рождения. Сборник мультфильмов", "alternative_titl..
page_url: http://filmlplace.ru/film/skazki-russkih-pisatelej-vyipusk-3.html
title: "Сказки русских писателей. Выпуск 3"
Детали: ..enres": ["Советские мультфильмы"], "description": "..
page_url: http://filmlplace.ru/film/v-mire-skazok-sbornik-multfilmov-vyipusk-4.html
title: "В мире сказок. Сборник мультфильмов. Выпуск 4"
Детали: ..ире сказок. Сборник мультфильмов. Выпуск 4", "altern..
page_url: http://filmlplace.ru/film/po-doroge-s-oblakami-ot-silvercinema.html
title: "По дороге с облаками"
Детали: .."Детский", "Русские мультфильмы"], "description": "..
page_url: http://filmlplace.ru/film/v-gostyah-u-skazki-sbornik-multfilmov.html
title: "В гостях у сказки. Сборник мультфильмов"
Детали: ..х у сказки. Сборник мультфильмов", "alternative_titl..
page_url: http://filmlplace.ru/film/pro-bolshih-i-malenkih-sbornik-multfilmov.html
title: "Про больших и маленьких. Сборник мультфильмов"
Детали: ..маленьких. Сборник мультфильмов", "alternative_titl..
page_url: http://filmlplace.ru/film/lesnyie-skazki-sbornik-multfilmov.html
title: "Лесные сказки. Сборник мультфильмов"
Детали: ..ные сказки. Сборник мультфильмов", "alternative_titl..
page_url: http://filmlplace.ru/film/ladushki-ladushki-sbornik-multfilmov.html
title: "Ладушки, ладушки. Сборник мультфильмов"
Детали: ..и, ладушки. Сборник мультфильмов", "alternative titl..
```

Из них:

- 2 сек на поиск документов.
- 27 сек на ранжирование результатов по 162370 документам.
- 1 сек на создание сниппетов на первой странице выдачи (50 документов)

Причины: длинный список словопозиций-документов, неоптимизированное ранжирование по методу косинусов (несмотря на то, что нормализованные компоненты векторов-документов загружаются с диска).

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.