

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

**Институт информационные технологии и прикладной
математики**

Кафедра вычислительной математики и программирования

**Лабораторная работа №4 по курсу
«Обработка текстов на естественном языке»**

Студент:	Е.М. Стифеев
Преподаватель:	А.А. Кухтичев
Группа:	М8О-109М-21
Дата:	04.12.21
Оценка:	
Подпись:	

Москва, 2021

Лабораторная работа №4 «Построение сниппетов»

Необходимо добавить в поисковую систему построение цитат (сниппетов), реферирование документов, найденных по запросу.

Сниппеты должны содержать слова запроса и давать пользователю представление о том, насколько документ отвечает поисковому запросу. Длина сниппета должна быть ограничена двумя-тремя строчками.

В отчёте нужно привести описание алгоритма построения сниппетов, примеры.

1. Описание

Корпус

Поисковая система обрабатывает запросы для корпуса документов, хранящегося на диске.

По итогам лабораторных работ по курсу с помощью веб-скрапинга по нескольким сайтам был получен корпус документов (доступен по ссылке <https://cloud.mail.ru/public/ZfkX/gccM7hnDR>), который имеет следующую структуру:

- films1.txt (94 Мб, 15000 документов, UTF-8)
- films2.txt (96 Мб, 15000 документов, UTF-8)
- films3.txt (184 Мб, 15000 документов, UTF-8)
- films4.txt (219 Мб, 15000 документов, UTF-8)
- films5.txt (322 Мб, 15000 документов, UTF-8)
- films6.txt (711 Мб, 15000 документов, UTF-8)
- films7.txt (823 Мб, 15000 документов, UTF-8)
- films8.txt (226 Мб, 15000 документов, UTF-8)
- films9.txt (67 Мб, 15000 документов, UTF-8)
- films10.txt (75 Мб, 15000 документов, UTF-8)
- films11.txt (99 Мб, 15000 документов, UTF-8)
- films12.txt (78 Мб, 15000 документов, UTF-8)
- films13.txt (41 Мб, 6109 документов, UTF-8)

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

Получение одного документа зачастую включало проход по нескольким html-страницам и обработку динамически подгружаемых страниц, поэтому общее количество обработанных страниц было >800'000.

Всего была обкачено три сайта:

- <https://www.kinopoisk.ru/>
- <https://shikimori.one/>
- <http://filmplace.ru/>

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 Смольянинов: «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\n\nАвтор: Дарико Цулая", "comments": ""}.
```

Индекс

Готовый индекс хранится в четырёх файлах (доступен по ссылке <https://cloud.mail.ru/public/wynT/adagiBjh9>):

- **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.

- **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.

Построение индекса для корпуса с учётом лемматизации терминов с помощью отечественной NLP-системы Natasha [1] занимает 4 часа при распараллеливании на 4 ОМР-потока (больше не позволяет размер оперативной памяти) процессора Intel Core i7 9700K (3.6 GHz). Блочный индекс (до слияния) доступен по ссылке <https://cloud.mail.ru/public/F3Fe/eTEiUPHt6>.

Построение сниппетов

Был реализован простейший алгоритм построения сниппетов. Пусть дан запрос в любой форме, состоящий из терминов $T = \{T_0, \dots, T_{m-1}\}$. Для документа, для которого требуется построить сниппет, с помощью координатного индекса возвратим все координатные вхождения в него терминов из запроса и отберём из них по *одному* для каждого термина запроса. В простейшем случае можно выбрать только первое или случайное вхождение. В более сложном нужно учитывать зону вхождения: заголовок, тело, метаданные, вступление, заключение; тепловую карту документа (с помощью eye-трекинга), машинное обучение и т.д. В лабораторной работе рассматривается только первое по порядку вхождение относительно начала документа. Далее, для каждого термина в определённом радиусе (в реализации – 20 символов) выводится текст находящийся слева и справа (с учётом границ документа, других терминов, которые могут попасть в радиус). Сам термин выделяется цветом или капитализацией.

Илка	Заголовок	Подробности
rs://www.kinopoisk....	Самые яркие и безумные новости года	..528553/", "title": "САМЫЕ яркие и безумные но...гове материалы:\n\nЛУЧШИЕ ФИЛЬМЫ 2014 года по версии..
rs://www.kinopoisk....	С Новым Годом!	..кучных праздников и САМЫХ ЛУЧШИХ подарков, позитивно.....личество интересных ФИЛЬМОВ и кинособытий, о ко..
rs://www.kinopoisk....	Кейси Аффлек убьет Брэда Питта	..Оппонентом Кейси по ФИЛЬМУ станет Брэд Питт, з.... считается одним из ЛУЧШИХ стрелков на Западе.....меется, хочет стать САМ...
k://filmlace.ru/film...	BBC: Живая природа. Ребятam о зверятах	..ательный сериал для САМЫХ маленьких зрителей. В этом ФИЛЬМЕ дети смогут не толь....но.Качайте", "Очень ХОРОШИЙ док сер...
rs://www.kinopoisk....	Читатели КиноПоиска и «ВКонтакте» подве...	..осование за главные ФИЛЬМЫ, САМЫХ достойных актеров и.....лено 15 номинаций: «ЛУЧШИЙ фильм», «Лучший реж...
rs://www.kinopoisk....	Подожди, пожалуй (2002)	..3", "description": "ФИЛЬМ о любви и смерти.",.....ными невзгодами. Не САМУЮ ЛУЧШУЮ анимацию (хотя отме...
rs://www.kinopoisk....	Читатели КиноПоиска и «ВКонтакте» выбер...	..«ВКонтакте» выберут ЛУЧШИЕ ФИЛЬМЫ за 15 лет", "body":":...режиссер, сделавший САМЫЙ большой вклад в рос...
k://filmlace.ru/film...	Лучшие из райских уголков... Австралия, Ег...	..is.html", "title": "ЛУЧШИЕ из райских уголков.....", "description": "В ФИЛЬМЕ рассказывается о САМЫХ красивых и интересн...
rs://www.kinopoisk....	КиноПоиск поздравляет с Новым годом!	..третье Новый год с САМЫМИ близкими и интересн.....каникулы, набраться ХОРОШЕГО настроения и, конеч.....любимый праздни...
k://filmlace.ru/film...	Великие художники	..т из документальных ФИЛЬМОВ в которых рассказы..... жизни и творчестве САМЫХ знаменитых художник.....ожниках, 9/10 Очень...
rs://www.kinopoisk....	The Red Stuff (2000)	..ты\лЛетчики, Отбор, САМЫЕ ЛУЧШИЕ, Первый в космосе,окой документальный ФИЛЬМ про Советских космо...
k://filmlace.ru/film...	BBC: Живая природа: Собаки	..ion": "Собаки - это САМЫЕ преданные и ЛУЧШИЕ друзья человека, вс.....ts": ["Великолепный ФИЛЬМ! С юмором, иронией ..
rs://www.kinopoisk....	«Фаворитку» признали лучшим британски...	..«Фаворитку» признали ЛУЧШИМ британским независимым ФИЛЬМОМ", "body":": "В выходн.....их категориях:\n\n\nСАМЫМ много...
rs://www.kinopoisk....	Австралийская киноакадемия назвала «Пар...	..назвала «Паразитов» ФИЛЬМОМ года", "body":": "Авс.....антино отметили как ЛУЧШЕГО режиссера за фильмбыть в курсе всех ...
rs://www.kinopoisk....	Пользователи «ВКонтакте» выбрали лучши...	..«ВКонтакте» выбрали ЛУЧШИЕ ФИЛЬМЫ 2017-го вместе с Ки.....ck & White, ставшую САМЫМ кассовым отечествен...
rs://shikimori.one/a...	«Оскар» за лучший монтаж: История номи...	..title": "«Оскар» за ЛУЧШИЙ монтаж: История ном.....аполучить статуэтку ФИЛЬМАМ «Вестсайдская истор..... использует один из С...
rs://shikimori.one/a...	Проект Кей: Семь историй. Воспоминания о...	..вал посмотреть данный ФИЛЬМ перед просмотром пе.....ленное произведение ЛУЧШИМ из цикла «семь историй», а также САМЫМ ...
rs://www.kinopoisk....	Киноакадемия не будет вручать «Оскар» за вручать «Оскар» за ЛУЧШИЙ популярный ФИЛЬМ", "body":": "Американ.....валось номинировать САМЫЕ кассовые ленты, час...
k://filmlace.ru/film...	Discovery: Машины Зла	..шины зла - выставка САМЫХ страшных изобретени..... человечества. Этот ФИЛЬМ - для смелых зрителе.....садо-мазо", "Всё бы ХОР...
rs://www.kinopoisk....	Пол Томас Андерсон напишет новый филь...	..ерсон напишет новый ФИЛЬМ в соавторстве с доч.....нимает истории о не САМЫХ приятных людях в не.....аботать идею. Очень ХО...
rs://www.kinopoisk....	«Трансформеры 3»: Видеоинтервью с Патри...	..извлечь из нее все САМОЕ ЛУЧШЕЕ29 июня в российском.....ате стартовал новый ФИЛЬМ Майкла Бэя «Трансфо...
rs://www.kinopoisk....	Том Круз снова встанет в ряды лучших стре...	..нова встанет в ряды ЛУЧШИХ стрелков", "body":":ия сиквела одной из САМЫХ успешных своих лент.....щения о продолжении Ф...

Рис. 1 – Построение сниппетов для выборки документов по запросу «самый лучший фильм» в графическом интерфейсе

```
По запросу: "!! самый хороший фильм" найдено 54498 документов
[INFO] Обработка результатов поиска
page_url: https://www.kinopoisk.ru/media/news/2528553/
title: "Самые яркие и безумные новости года"
Детали: ..528553/", "title": "САМЫЕ яркие и безумные но....гове материалы:\n\nЛУЧШИЕ фильмы 2014 года по версии..
page_url: https://www.kinopoisk.ru/media/news/1127648/
title: "С Новым Годом!"
Детали: ..кучных праздников и самых лучших подарков, позитивно....личество интересных фильмов и кинособытий, о ко..
page_url: https://www.kinopoisk.ru/media/news/139973/
title: "Кейси Аффлек убьет Брэда Питта"
Детали: ..Оппонентом Кейси по фильму станет Брэд Питт, з.... считается одним из лучших стрелков на Западе.....меется, хочет стать самым лучшим, сместив с т..
page_url: http://filmlace.ru/film/bbc-zhivaya-priroda-rebyatam-o-zveriyatah.html
title: "BBC: Живая природа. Ребятam о зверятах"
Детали: ..ательный сериал для самых маленьких зрителей. В этом фильме дети смогут не толь....но.Качайте", "Очень хороший док сериал! Детям, ..
page_url: https://www.kinopoisk.ru/media/news/3898389/
title: "Читатели КиноПоиска и «ВКонтакте» подведут итоги 2017 года"
Детали: ..осование за главные фильмы, самых достойных актеров и.....лено 15 номинаций: «Лучший фильм», «Лучший реж..
page_url: https://www.kinopoisk.ru/film/465042/
title: "Подожди, пожалуй (2002)"
Детали: ..3", "description": "Фильм о любви и смерти.",.....ными невзгодами. Не самую лучшую анимацию (хотя отме..
page_url: https://www.kinopoisk.ru/media/news/3280840/
title: "Читатели КиноПоиска и «ВКонтакте» выберут лучшие фильмы за 15 лет"
Детали: ..«ВКонтакте» выберут лучшие фильмы за 15 лет", "body":":...режиссер, сделавший самый большой вклад в рос..
page_url: http://filmlace.ru/film/luchshie-iz-rajskih-ugolkov-avstraliya-egipet-kuba-tunis.html
title: "Лучшие из райских уголков... Австралия, Египет, Куба, Тунис"
Детали: ..is.html", "title": "ЛУЧШИЕ из райских уголков.....", "description": "В фильме рассказывается о самых красивых и интересн..
```

Рис. 2 – Построение сниппетов для выборки документов по запросу «самый лучший фильм» в консольном интерфейсе

Поисковая система. Интерфейс

Реализовано десктоп-приложение для ОС семейства Windows.

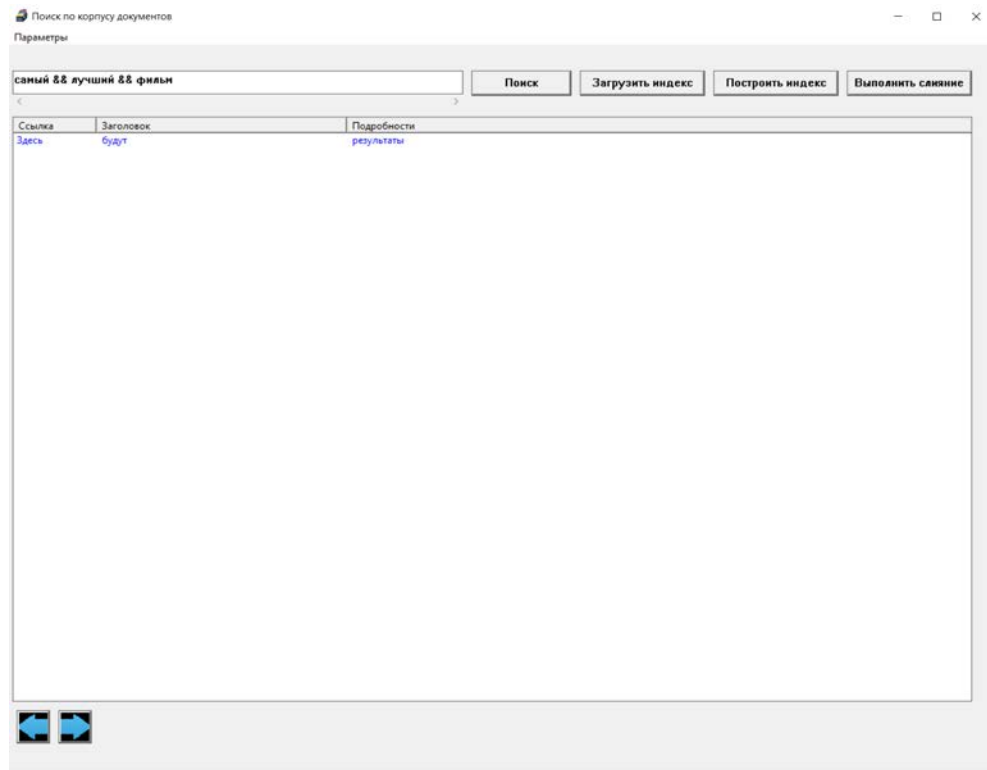
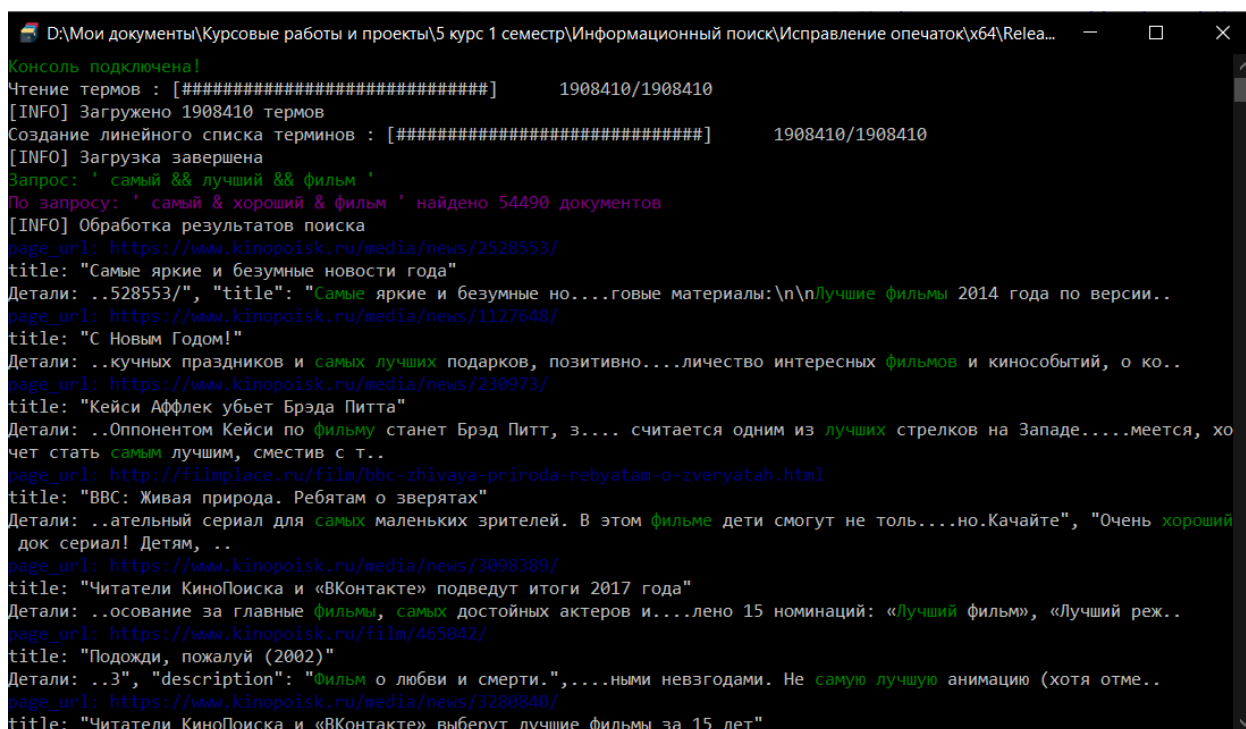


Рис. 3 – Стартовое графическое окно приложения

При запуске приложения пользователь видит два окна: графическое (с элементами управления) и консольное – для логирования.



```
Консоль подключена!
Чтение термов : [#####] 1908410/1908410
[INFO] Загружено 1908410 термов
Создание линейного списка терминов : [#####] 1908410/1908410
[INFO] Загрузка завершена
Запрос: ' самый && лучший && фильм '
По запросу: ' самый & хороший & фильм ' найдено 54490 документов
[INFO] Обработка результатов поиска
page_url: https://www.kinopoisk.ru/media/news/2528553/
title: "Самые яркие и безумные новости года"
Детали: ..528553/", "title": "Самые яркие и безумные но....говые материалы:\n\nлучшие фильмы 2014 года по версии..
page_url: https://www.kinopoisk.ru/media/news/1127648/
title: "С Новым Годом!"
Детали: ..кучных праздников и самых лучших подарков, позитивно....личество интересных фильмов и кинособытий, о ко..
page_url: https://www.kinopoisk.ru/media/news/230973/
title: "Кейси Аффлек убьет Брэда Питта"
Детали: ..Оппонентом Кейси по фильму станет Брэд Питт, э.... считается одним из лучших стрелков на Западе....меется, хо
чет стать самым лучшим, сместив с т..
page_url: http://filmplace.ru/film/bbc-zhivaya-priroda-rebyatam-o-zveryatah.html
title: "BBC: Живая природа. Ребятам о зверятах"
Детали: ..ательный сериал для самых маленьких зрителей. В этом фильме дети смогут не толь....но.Качайте", "Очень хороший
док сериал! Детям, ..
page_url: https://www.kinopoisk.ru/media/news/3098389/
title: "Читатели КиноПоиска и «ВКонтакте» подведут итоги 2017 года"
Детали: ..осование за главные фильмы, самых достойных актеров и....лено 15 номинаций: «Лучший фильм», «Лучший реж..
page_url: https://www.kinopoisk.ru/film/465042/
title: "Подожди, пожалуй (2002)"
Детали: ..3", "description": "Фильм о любви и смерти.",....ными невзгодами. Не самую лучшую анимацию (хотя отме..
page_url: https://www.kinopoisk.ru/media/news/3280840/
title: "Читатели КиноПоиска и «ВКонтакте» выберут лучшие фильмы за 15 лет"
```

Рис. 4 – Консольное окно приложения

При взаимодействии пользователя с системой последняя через консольное окно ведёт оповещение, чем она «занята» в данный момент. Прежде делать запрос к корпусу необходимо загрузить индекс с помощью кнопки «Загрузить индекс» или, если индекс не создан, то необходимо создать блочный индекс и затем выполнить слияние с помощью соответствующих кнопок. Корпус должен храниться в директории на диске в формате, описанном в информации о корпусе (нумерация файлов необязательна – названия - произвольные). При этом один файл с документами считается системой одним блоком, которые обрабатываются параллельно, так что к выбору размера файла нужно подходить разумно.

В меню «параметры» настраивается количество потоков и пути до нужных директорий. При закрытии и запуске приложения система запоминает все пути, количество потоков, а также последний сделанный пользователем запрос.

?

Параметры

×

Путь к директории с корпусом

и документы\Лабы и рефераты\5 курс 1 семестр\Информационный поиск\Корпус

←

→

↓

Путь к директории с индексом

менты\Лабы и рефераты\5 курс 1 семестр\Информационный поиск\Корпус_index

←

→

↓

Путь к директории с временными файлами

я и проекты\5 курс 1 семестр\Информационный поиск\Исправление опечаток\tp

←

→

↓

Число потоков

5

↑

↓

Рис. 5 – Выбор параметров

После загрузки индекса, можно делать запросы в формах, описанных в предыдущем разделе. Результаты сортируются по косинусному правилу TF-IDF.

Поиск по корпусу документов

Параметры

— □ ×

[самый && лучший && фильм] || "что где когда" / 5

Поиск

Загрузить индекс

Построить индекс

Выполнить слайзинг

Ссылка	Заголовок	Подробности
https://www.ki...	«Оскар-2017»: Шорт-лист фильмов с лучшими ви...	...ар-2017»: Шорт-лист ФИЛЬМОВ с ЛУЧШИМИ визуальными эффекта...етендентов значатся САМЫЙ кассовый фильм 2016...тасти...
https://filmplice...	Непростые вопросы: Путешествие в страну шама...	...нтальный", "Русский ФИЛЬМ", "description": "...яться и радоваться САМЫМ простым вещам... В ...ло. Я могу отличить ХОРОШЕИ...
https://www.ki...	9 лучших трейлеров недели: Человек-невидимка,2525/", "title": "9 ЛУЧШИХ трейлеров недели: Ч...анипат в 1761 году, ГДЕ сразились силы импе... сражение считается САМЫМ кру...
https://www.ki...	«Август. Восьмого»: Репортаж со съемочной пл...	...596/", "title": "10 ЛУЧШИХ трейлеров недели: Т...льно главные роли в ФИЛЬМЕ о противостоянии дв...мы предупредим вас, КОГД...
https://www.ki...	10 лучших трейлеров недели: Темные воды, альп...	...5337/", "title": "9 ЛУЧШИХ трейлеров недели: Х...лился на «Оскар». В ФИЛЬМЕ «Вне игры» есть все, ЧТО может привлечь внима...а...
https://www.ki...	9 лучших трейлеров недели: Харли Квинн, новень...	...1759/", "title": "8 ЛУЧШИХ трейлеров недели: А...емало драматических ФИЛЬМОВ, но это первая карт...е генераторы, из-за ЧЕГО...
https://www.ki...	8 лучших трейлеров недели: Атомные самураи, д...	...ериале «Теин Пикс», ГДЕ он сыграл франкокана...ерифа Эрла МакГро в ФИЛЬМАХ «От заката до рассв... навсегда останется ЛУЧШ...
https://www.ki...	Ушел из жизни актер Майкл Паркс	...Драма», "Зарубежный ФИЛЬМ", "description": "...тонаселенный район, ГДЕ проживала девушка, ... на свои злодеяния. САМОЕ п...
https://filmplice...	37	...s", "title": "Высы: ФИЛЬМ / Air Movie", "comm...ни... Единственное ЧТО не понравилось в фи...у — что это одна из ЛУЧШИХ полн...
https://shikimo...	Высы: Фильм / Air Movie	...297/", "title": "10 ЛУЧШИХ трейлеров недели: Д...я весной 1917 года, КОГДА от действий двух со...дупредим вас, когда ФИЛЬМ стан...
https://www.ki...	10 лучших трейлеров недели: Джентльмены, хищ...	...но и на Западе. В ФИЛЬМЕ повествуется об обы...й можно делать все, ЧТО угодно, а потом вык... меня ностальгию))) ХОРОШИЙ
https://filmplice...	Приключения пингвиненка Лоло	...е Rotten Tomatoes у ФИЛЬМА лишь 17% «свежести»...скусства, созданное ЛУЧШИМИ людьми в мире. Я пр...ожет стать одним из...
https://www.ki...	Звезда «Кошек» Джейсон Дрүроул ответил критика...	...первой ленты серии, ГДЕ за главными героями...нстр. Кто сыграет в ФИЛЬМЕ и КОГДА ждать его выхода, п...ужой» стал одним из...
https://www.ki...	«Чужого» подвергнут перезагрузке	...пяти номинациям на ЛУЧШИЙ ФИЛЬМ", "body": "Десять фи...миллиона зрителей, ЧТО на 16 % меньше, чем...церемонии оказал...
https://filmplice...	Киноакадемия может вернуться к пяти номинаци...	...ючения", "Советский ФИЛЬМ", "description": "...ебыванием в местах, ГДЕ живут люди, создают...а это делает... Ни ЧЕГО сложного...
https://www.ki...	Новые приключения Дони и Микки	...0403/", "title": "7 ЛУЧШИХ трейлеров недели: ...ый бульдозер, после ЧЕГО поехал на нем верши...гинцева на создание ФИЛЬМА...
https://www.ki...	7 лучших трейлеров недели: «Правосудие Спенсе...	...6120/", "title": "9 ЛУЧШИХ трейлеров недели: ...нсы в первых восьми ФИЛЬМАХ режиссера. Создатель...ом и стал одним из САМ...
https://www.ki...	9 лучших трейлеров недели: «Звездные войны», К...	...5596/", "title": "9 ЛУЧШИХ трейлеров недели: Х...снимают очень много ФИЛЬМОВ ужасов. Новый хорро...мы предупредим вас, К...
https://www.ki...	9 лучших трейлеров недели: Холодное сердце, Ир...	...ы 3 сезон сняли.", "ХОРОШИЙ и интересный сериал...не сериал зашел, не ЧЕГО особенного, дешево...е всеми остальными, СА...
https://filmplice...	Рagnarøk	...могих стал одним из САМЫХ ожидаемых ФИЛЬМОВ года, он может заяв...е один фильм, после ЧЕГО хочет отойти от биз...смляе...
https://www.ki...	Режиссер «Новолуния» хочет уйти из кино	...0-х Ривер с помощью ФИЛЬМА «Бешеный бык» заста...ачнешь играть — вот ЧТО ты начнешь делать", ...с считался одним из СА...
https://www.ki...	Хоакин Феникс поблагодарил брата Ривера за сво...	...ую ветвь», одну из САМЫХ престижных кинопрем...ера Пака. Произведа ХОРОШЕЕ впечатление. Ки-у...мы предупредим вас, Н...
https://www.ki...	Премьера на КиноПоиске: Трейлер «Паразитов»	...ть и ждать", "что ЛУЧШЕ ждать пока те сериа...ностью или смотреть ФИЛЬМЫ?@KiritoAsuna, @Roma...йдет, чем дождешься КОГ...
https://shikimo...	Chain Chronicle: Haecceitas no Hikari Part 2	...Драма», "Зарубежный ФИЛЬМ", "Комедия", "Мелод...т, что в комплексе, ГДЕ он живет, вселилась...(мормоны - одна из САМЫХ с...
https://filmplice...	Порыв ветра	...етраженный", "Русский ФИЛЬМ", "description": "...ого человека. День, КОГДА все против него. Де...талантливо. 6/10", "ХОРОШАЯ д...
https://www.ki...	Что? Где? Когда?	...656730/", "title": "ЧТО? ГДЕ? КОГДА?", "alternative_tit...
https://filmplice...	Лиса и Заяц я этот мультфильм!", "САМОЕ ЛУЧШЕЕ для меня в этом мульт...зверушек почти весь ФИЛЬМ показывается в анфа... думал, даже что ег...
https://www.ki...	5 лучших трейлеров недели: «Харли Квинн», «Хок...	...0958/", "title": "5 ЛУЧШИХ трейлеров недели: «...винн, «Хокусай» и «ЧЕМ мы заняты в тени»...мы предупредим вас, КОГДА прои...
https://www.ki...	Кирилл Серебренников напишет и поставит мини...	...арковский — один из САМЫХ прославленных совет...ких режиссеров, чьи ФИЛЬМЫ хорошо известны дал...включаются в список...
https://www.ki...	Пять лет и один день (2012)	...аре, ее муж живет с ЛУЧШЕЙ подружкой, которую б...рхжи, единственное, ЧТО есть у женщины — же... этот замечательный ФИЛЬ...
https://filmplice...	Помадные джунгли	... "После долгого дня САМОЕ тоТакой лёгкий)))", ...лицу. Но, по моему, ЛУЧШЕ жить плодотворной, ...му за 30!", "Искала ФИЛЬМ г...
https://www.ki...	День, когда я нравился девушкой (2006)	.../", "title": "День, КОГДА я нравился девушкой...мною, считаются его ЛУЧШЕЙ работой. А конкретн... больше понравился СА...
https://filmplice...	Генезис 2.0	...льный", "Зарубежный ФИЛЬМ", "description": "...ые можно продать за ХОРОШИЕ деньги на «чёрном» ...аторию Южной Кореи, Г...
https://www.ki...	«Оскар-2018»: Лонг-лист фильмов с лучшими виз...	...ар-2018»: Лонг-лист ФИЛЬМОВ с ЛУЧШИМИ визуальными эффекта...листе оказались как САМЫЕ кассовые фильмы 201...извест...
https://filmplice...	Легенда о царе Соломоне	...ми решениями даже в САМЫХ сложных ситуациях... сином. Анимационный ФИЛЬМ «Легенда о царе Сол...менный зритель см...
https://www.ki...	Проклятые фотографии (2006)	...ews": ["Фотографии\ФИЛЬМ мне попался совершенно... случайно, я искала ХОРОШИЙ фильм ужасов и наткнулась на этот КОГДА...
https://www.ki...	Министр культуры не исключил раздвижек кино...	...а релиза зарубежных ФИЛЬМОВ, выход которых запл...ики — время премьер ЛУЧШИХ российских фильмов...й декабристов. Ж...
https://www.ki...	Нарушение общественного порядка	...lres": ["Зарубежный ФИЛЬМ", "Комедия", "desc...прожил в Австралии, ГДЕ работал обычным про... встречался с тремя ЛУЧШИ...
https://www.ki...	Интерстеллар: Премьера финального дублירו...	...ь ленту. В итоге за ФИЛЬМ взялся Кристофер Но...нтерстеллар» станет САМЫМ длинным фильмом Кри...акрытом показе, это Л...
https://www.ki...	Проклятие куклы Роберт	...lres": ["Зарубежный ФИЛЬМ", "Ужасы", "descp...на работу в музей, ГДЕ много странных эксп... именем Роберт. Все САМОЕ не...

←

→

1/1091

Рис. 6 – Поисковая выдача по комбинированному запросу (графическое окно)

```
D:\Мои документы\Курсовые работы и проекты\5 курс 1 семестр\Информационный поиск\Исправление опечаток\64\Release...
все всех самых важных новостей из ..
Запрос: ' (самый && лучший && фильм) | "что где когда" / 5 '
По запросу: ' (самый & хороший & фильм) | "что где когда" / 5 ' найдено 54525 документов
[INFO] Обработка результатов поиска
page_url: https://www.kinopoisk.ru/media/news/2866427/
title: "«Оскар-2017»: Шорт-лист фильмов с лучшими визуальными эффектами"
Детали: ..ар-2017»: Шорт-лист фильмов с лучшими визуальными эффектами...етендентов значатся самый кассовый фильм 2016....
статистические твари и где они обитают).\n\nЧл....известен 24 января, когда будут обнародованы ..
page_url: http://filmpolice.ru/film/neprostye-voprosyi-puteshestvie-v-stranu-shamanov-altaj.html
title: "Непростые вопросы: Путешествие в страну шаманов - Алтай"
Детали: ..нтальный", "Русский фильм"], "description": "...яться и радоваться самым простым вещам... В ....ло. Я могу о
тличить хорошее от плохого, я вижу....я ее только тогда, когда мне это действительно...плохой человек. А что же не та
к? Почему я....ьше? ..... Кто я? Где мой дом? Со всеми э..
page_url: https://www.kinopoisk.ru/media/news/3432525/
title: "9 лучших трейлеров недели: Человек-невидимка, Ип Ман и душа Pixar"
Детали: ..2525/", "title": "9 лучших трейлеров недели: Ч...анипат в 1761 году, где сразились силы импе.... сражение счи
тается самым крупным в XVIII век....мы предупредим вас, когда проект станет досту....риях. Одна из них – фильм о мире ду
ш, где зар....о-то инопланетного, что не поддается объясн..
page_url: https://www.kinopoisk.ru/media/article/1592949/
title: "«Август. Восьмого»: Репортаж со съемочной площадки"
Детали: .. Она сама не знает, чего ей хочется, поэтому....вильон «Мосфильма», где режиссер Джаник Фай.... создает свой н
овый фильм «Август. Восьмого»....ой, и стал одним из самых кассовых российских....).\n\nВ тот день, когда КиноПоиск п
риехал в....ость.\n\nВпрочем, лучше посмотрите все сами..
page_url: https://www.kinopoisk.ru/media/news/3413596/
title: "10 лучших трейлеров недели: Темные воды, альпинисты и хранители"
Детали: ..596/", "title": "10 лучших трейлеров недели: Т...льно главные роли в фильме о противостоянии дв....мы предупр
едем вас, когда фильм выйдет в кино....аются заработать на самых известных произведе....развивается в поле, где происход
ит что-то с..
page_url: https://www.kinopoisk.ru/media/news/3435337/
title: "9 лучших трейлеров недели: Харли Квинн, новенький Соник и губка в бегах"
```

Рис. 7 – Поисковая выдача по комбинированному запросу (консольное окно)

По результатам запроса пользователь открывает ссылку в первом столбце в любом, привычном ему, браузере.

Также существует полностью консольная версия приложения, существующая в т.ч. для тестирования и измерения метрик поисковой системы. Ей на вход в качестве аргументов программы подаются на вход все вышеуказанные параметры, а также путь к данным для подсчета метрик (см. подробности в соответствующей ЛР).

```
D:\Мои документы\Курсовые работы и проекты\5 курс 1 семестр\Информационный поиск\Исправление опечаток\64\Release\Исправление о...
По запросу: ' " тихий место 2 " ' найдено 50 документов
Запрос: ' аватар скачать '
По запросу: ' !! аватар скачать ' найдено 474 документов
Запрос: ' "re zero" / 4 '
По запросу: ' " re zero " / 4 ' найдено 266 документов
Запрос: ' режиссер назад в будущее '
По запросу: ' !! режиссер назад в будущее ' найдено 6233 документов
Запрос: ' фильм для интеллектуалов '
По запросу: ' !! фильм для интеллектуал ' найдено 1286 документов
Запрос: ' фильм сериал с самым большим рейтингом '
По запросу: ' !! фильм сериал с самым большим рейтингом ' найдено 4820 документов
Запрос: ' фильмы Макото Синкай '
По запросу: ' !! фильм макото синкай ' найдено 271 документов
Запрос: ' лучшие фильмы квентина тарантино '
По запросу: ' !! хороший фильм квентин тарантино ' найдено 1266 документов
Запрос: ' как звали главного героя коносуэ '
По запросу: ' !! как звать главный герой коносуэ ' найдено 85 документов
Запрос: ' самый лучший фильм '
По запросу: ' !! самый хороший фильм ' найдено 54490 документов
Запрос: ' сериалы с рейтингом 18+ '
По запросу: ' !! сериал с рейтингом 18 ' найдено 2061 документов
Запрос: ' высшая школа демонов '
По запросу: ' !! высший школа демон ' найдено 1478 документов
Запрос: ' джокер '
По запросу: ' джокер ' найдено 1353 документов
Точность на уровне 5 = 0.400000
DCG на уровне 5 = 1.289830
nDCG на уровне 5 = 0.437459
ERR на уровне 5 = 0.717949
```

Рис. 8 Утилита тестирования системы

2. Исходный код

Структура проекта

- include
 - algebra.hpp (простейшие операции с векторами)
 - 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)
 - resource.h (подключение изображений, иконок и прочего)
 - search.hpp (реализация всех видов поиска)
 - token_parse.hpp (функции для преобразования токенов в термы)
 - typos_correction.hpp (реализация исправления опечаток)
- 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 тестирования программы)
- resources (файлы ресурсов для оконного приложения)

Проект был написан с помощью Microsoft Visual Studio 2019 эксклюзивно для ОС семейства Windows. Исходный код доступен по <https://github.com/Stifeev/Information-retrieval/tree/main/Курсовой%20проект>.

Запуск и сборка

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

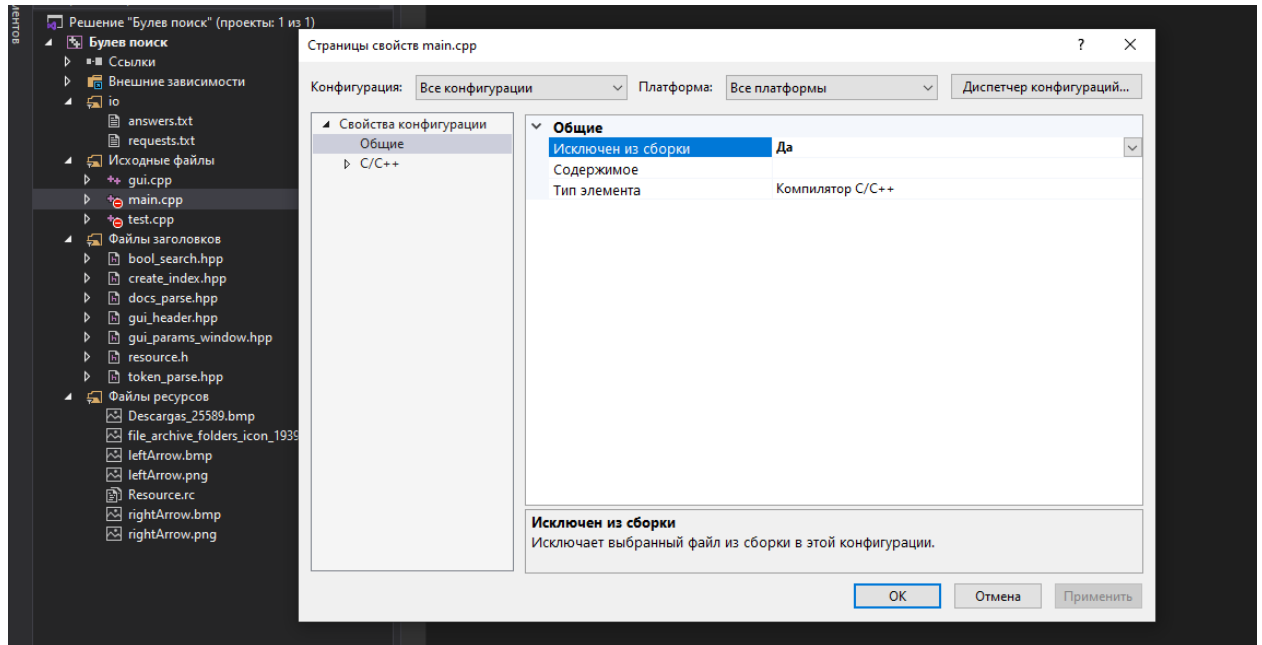


Рис. 9 – Настройка переключения между точками входа

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

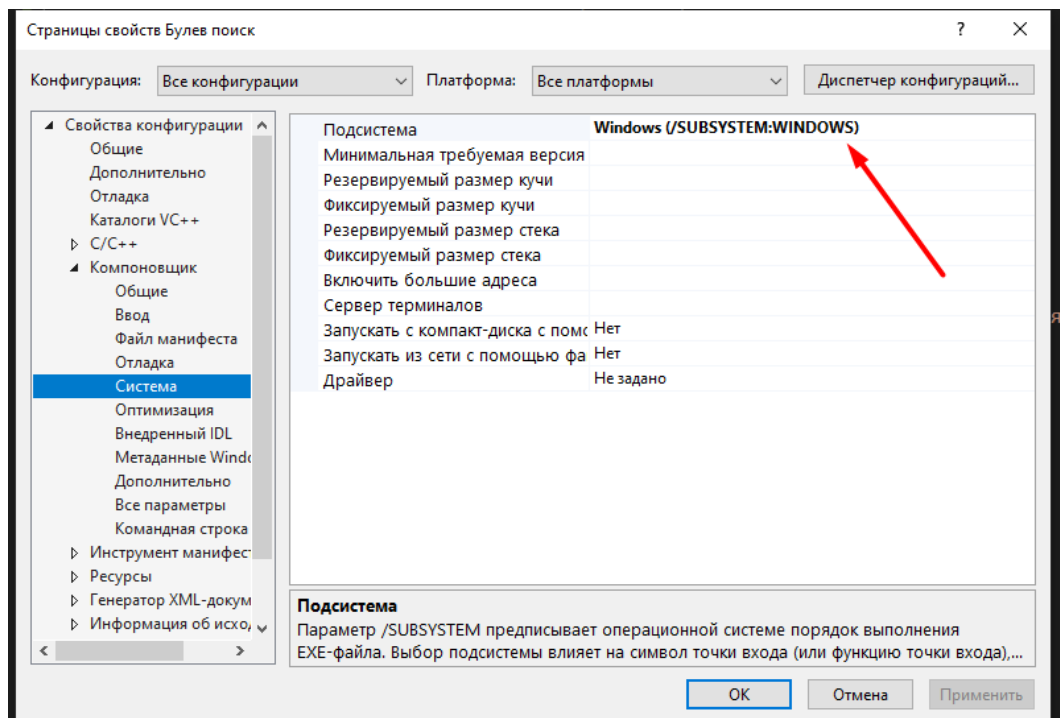


Рис. 10 – Настройка подсистемы: консоль или окно

Консольное приложение поддерживает флаги запуска:

- -i 'путь к корпусу'
- -o 'путь к индексу'
- -t 'путь к директории с блочным индексом'
- -m 'путь к директории с эталонами для метрик'
- -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
- -metric : высчитать метрики

Пример создания блочного индекса из корпуса (время указано до внедрения лемматизации):

```
$ ./prog.exe -p 4 -create -i "..\..\Корпус" -o -t "tmp"
```

Вывод

```
[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 слияние блочного индекса:

```
$ ./prog.exe -p 4 -merge -clear -i "..\..\Корпус_index" -t "tmp"
```

Вывод

```
[INFO] Слияние блочного индекса
[INFO] Создание очередей термов:   13 блок из   13
[INFO] Слияние docs_id:   13 блок из   13
[INFO] Слияние словозиций термов
[INFO] Осталось термов:           0
[INFO] Общее число термов в словаре = 1908410
Документов = 186109
[INFO] Время на слияние блочного индекса: 35 sec
[INFO] Вычисление статистики
Первый проход. Термов осталось:           0
Второй проход. Документов осталось:           0
[INFO] Вычисление статистики закончено
```

Пример просчёта метрик:

```
$ ./prog.exe -p 4 -metric -i "..\..\Корпус" -o "..\..\Корпус_index"
-m "..\..\Корпус_metric"
```

Вывод

```
Чтение термов : [#####] 1908410/1908410
[INFO] Загружено 1908410 термов
Запрос: ' "тихое место 2" '
По запросу: ' " тихий место 2 " ' найдено 50 документов
Запрос: ' аватар скачать '
По запросу: ' !! аватар скачать ' найдено 474 документов
Запрос: ' "re zero" / 4 '
По запросу: ' " re zero " / 4 ' найдено 266 документов
```

Запрос: ' режиссёр назад в будущее '

По запросу: ' !! режиссер назад в будущее ' найдено 6233 документов

Запрос: ' фильм для интеллектуалов '

По запросу: ' !! фильм для интеллектуал ' найдено 1286 документов

Запрос: ' фильм сериал с самым большим рейтингом '

По запросу: ' !! фильм сериал с самый больший рейтинг ' найдено 4820 документов

Запрос: ' фильмы Макото Синкая '

По запросу: ' !! фильм макото синкай ' найдено 271 документов

Запрос: ' лучшие фильмы квентина тарантино '

По запросу: ' !! хороший фильм квентин тарантино ' найдено 1266 документов

Запрос: ' как звали главного героя коносубы '

По запросу: ' !! как звать главный герой коносуб ' найдено 85 документов

Запрос: ' самый лучший фильм '

По запросу: ' !! самый хороший фильм ' найдено 54490 документов

Запрос: ' сериалы с рейтингом 18+ '

По запросу: ' !! сериал с рейтинг 18 ' найдено 2061 документов

Запрос: ' высшая школа демонов '

По запросу: ' !! высокий школа демон ' найдено 1478 документов

Запрос: ' джокер '

По запросу: ' !! джокер ' найдено 1353 документов

Точность на уровне 30 = 0.241026

DCG на уровне 30 = 2.655877

nDCG на уровне 30 = 0.289893

ERR на уровне 30 = 0.732372

3. Выводы

Построение сниппетов – важная часть работы поисковой системы. С помощью них пользователь решает, какие документы открывать, а какие – нет. Качественные сниппеты экономят время пользователя, а не качественные – наоборот.

В лабораторной работе реализована простейшая их реализация, однако даже она работает и даёт представление о чём документ, стоящий в поисковой выдаче.

В ходе лабораторной работы я научился выполнять строить сниппеты для коллекции документов.

Литература

- [1] <https://natasha.github.io/>
- [2] Кристофер Д.Маннинг, Прабхакар Рагхаван, Хайнрих Шютце. Введение в информационный поиск. 2020, изд. Вильямс.