

МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

Институт №8 «Компьютерные науки и прикладная математика»

**Отчет по лабораторным работам
по курсу «Технологии параллельного программирования»**

Выполнил: Полиха Александр
Владимирович
Группа: М8О-403Б-22
Преподаватель: Кухтичев Антон
Алексеевич

Москва, 2025

Цель работы

Разработать поисковой движок на C++ для выбранного корпуса документов. Этот корпус документов должен собираться поисковым роботом, хранится корпус в MongoDB.

Привести несколько запросов к существующим поисковикам, указать недостатки.

Описание данных

В качестве корпуса данных был выбран

В качестве первого источника был выбран набор статей из Wikipedia, а именно – Биографии музыкальных деятелей. Этот ресурс известен своим большим количеством статей. В качестве второго источника был выбран belcanto, в нем насчитывается порядка 5 000 биографий, а в Wikipedia около больше 30 000. Для парсинга был написан поисковой робот на Python, который собирает html Документы, выполняя правила robots.txt.

Текст у этих документов состоит из сырого html, есть мета информация по кодировке, описанию, ключевые слова, автор.

Документ содержит:

1. Основной контент статьи
2. Заголовок
3. Дата/время публикации
4. Рубрика/тема
5. Текст статьи
6. Изображения/видео (иногда) с подписями/источниками
7. Ссылки внутри текста и блоки по теме
8. Навигация и сервисные блоки
9. Элементы авторизации/регистрации
10. Канонический URL

Для выбранных источников существует поисковик на сайте, также можно воспользоваться поиском Google, Yandex. Поэтому их можно использовать для выполнения лабораторной работы.

Примеры существующих поисковиков

Поиск на сайте:

Результаты поиска

Q: музыкант Расширенный поиск: Поиск по: (Основное)

Просмотреть (предыдущие 20 | следующие 20) (20 | 50 | 100 | 250 | 500)

Музыкант
себя: музыкант-исполнитель, артист инструменталист и т. д. Распространённые словосочетания: поп-музыкант, камер-музыкант, фолк-музыкант. музыкант // Толковый...
9 KB (523 слова) - 19:43, 26 октября 2025

Дельфин (музыкант)
Дельфин (англ. Dolphin) — российский rap-исполнитель, мелодекламатор, певец, музыкант, композитор, музыкальный продюсер, автор песен и поэт. Бывший участник...
211 KB (13 218 слов) - 12:31, 25 октября 2025

Воробьев, Алексей Владимирович (артист) (перенаправление с Алексей Владимирович Воробьев (музыкант))
кинорежиссёр, сценарист, кинопродюсер, кинокомпозитор, монтажёр, певец, музыкант и телеведущий. Автор таких хитов, как «Сумасшедшая», «Я тебя люблю», «Всегда...
64 KB (3234 слова) - 12:33, 15 ноября 2025

Джин Симmons (музыкант)
род. 25 августа 1949 года, Тират-Кармель, Израиль) — американский рок-музыкант, бас-гитарист, вокалист, актёр и предприниматель. Один из основателей группы...
26 KB (1542 слова) - 10:23, 18 декабря 2025

Рисунок 1 – поиск на сайте

site:wikipedia.org "Музыканты"

All Images Videos Short videos Shopping News More Tools

Википедия https://ru.wikipedia.org › wiki › К... · Translate this page · Категория:Музыканты по алфавиту
В данную категорию автоматически добавляются страницы, содержащие шаблоны: {{Музыкант}},; из категории «Шаблоны, встраиваемые в шаблоны-карточки:Музыканты», ... [Read more](#)

Википедия https://ru.wikipedia.org › wiki › К... · Translate this page · Категория:Музыканты
Страницы в категории «Музыканты». Показаны 2 страницы из 2, находящихся в данной категории. Список ниже может не отражать последних изменений. Музыкант ... [Read more](#)

Википедия https://ru.wikipedia.org › wiki › К... · Translate this page · Музыкант
... «Музыканты», где речь идет о певцах. В толковании С. И. Ожегова музыкант — это артист, играющий на музыкальном инструменте, а также вообще человек ... [Read more](#)

Википедия https://ru.wikipedia.org › wiki › С... · Translate this page · Список самых продаваемых музыкальных исполнителей
Музыканты перечислены как с заявленными продажами, так и с общим количеством ... Список самых продаваемых музыкантов США. Заметки. править. ↑ Чтобы быть ... [Read more](#)

Рисунок 2 – поиск через Google

Недостатки существующих поисковиков:

- Для обычного пользователя любого браузера может быть трудно найти нужное содержимое без дополнительных знаний по составлению грамотного запроса.
- Необъективность, при одинаковом запросе, без использования параметров поисковика, результаты могут быть разными, например, Яндекс будет продвигать их маркетплейс на самую первую позицию при запросе «купить наушники», в то время как google подвинет на первую позицию в выпадающем списке ozon.

- Часто после выполнения запроса, в результатах отображаются только популярные источники.

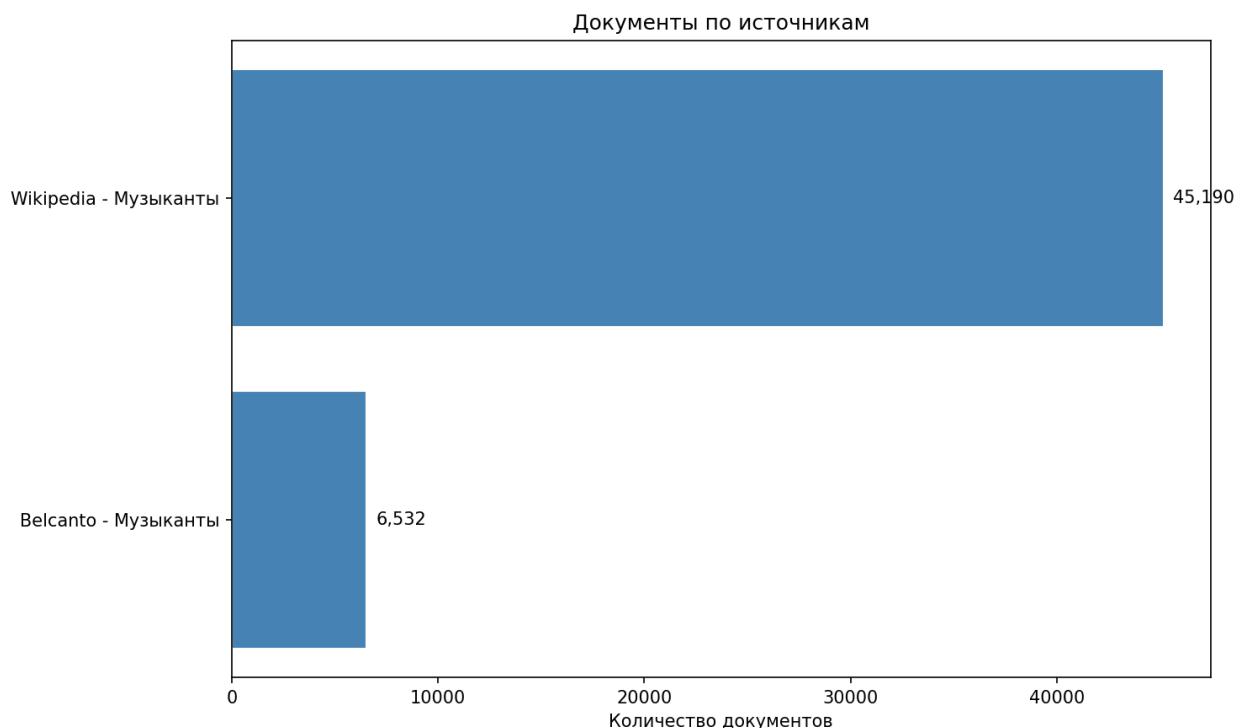


Рисунок 3 – Распределение документов по источникам

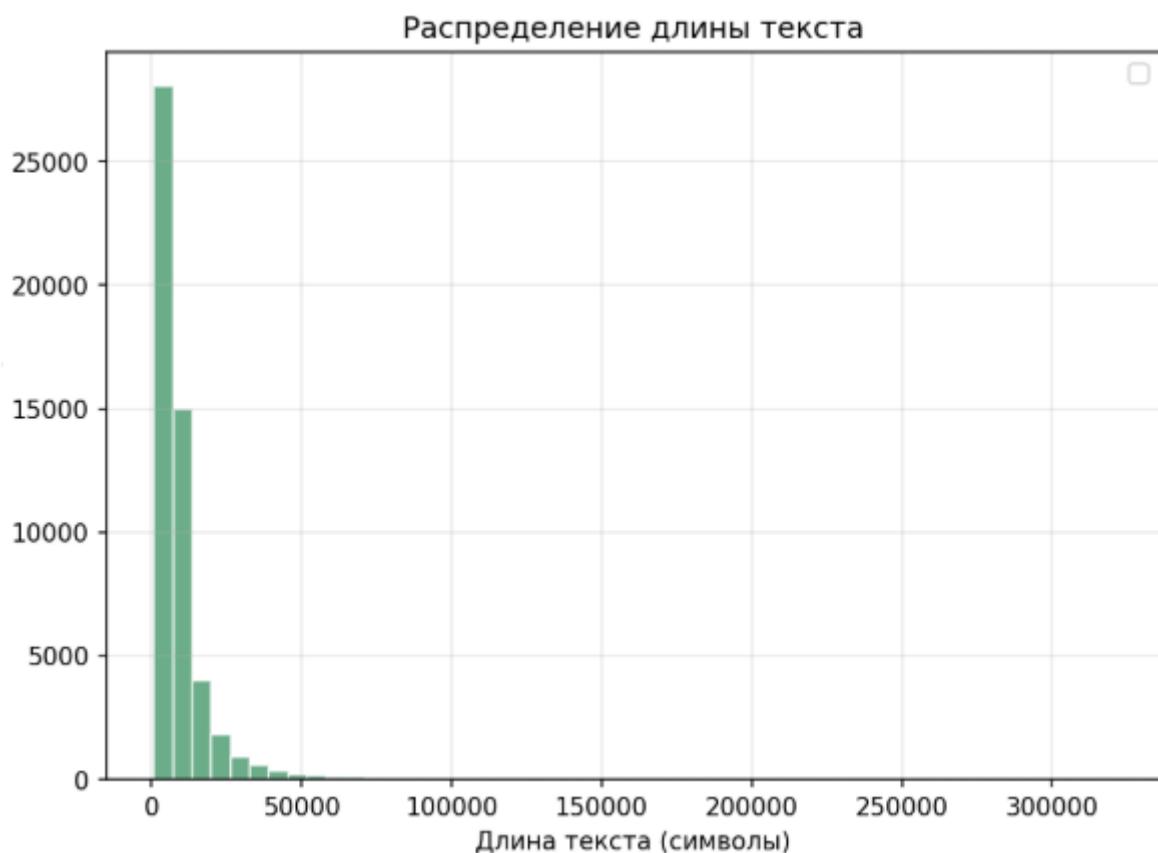


Рисунок 4 – Распределение количества символов в документах

После работы поискового робота мы получаем:

- 51 722 – количество документов.
- 10 287 символов – средний размер текста в документе.
- Объем корпуса – 1.47 GB

Токенизация

С токенизацией есть проблемы, например, токен «абдаль-мумин» - содержит дефис, для чего нужны отдельные правила для токенизации.

После работы индексатора были получены следующие результаты:

Документов: 51,722

Всего токенов: 53,295,593

Уникальных токенов: 1,321,300

Уникальных стемов: 835,856

Средняя длина токена: 7.15 символов

Размер индекса: 155.86 MB

Скорость токенизации: 2812.43 KB/сек

Постройка индекса заняла 1682.31 секунды

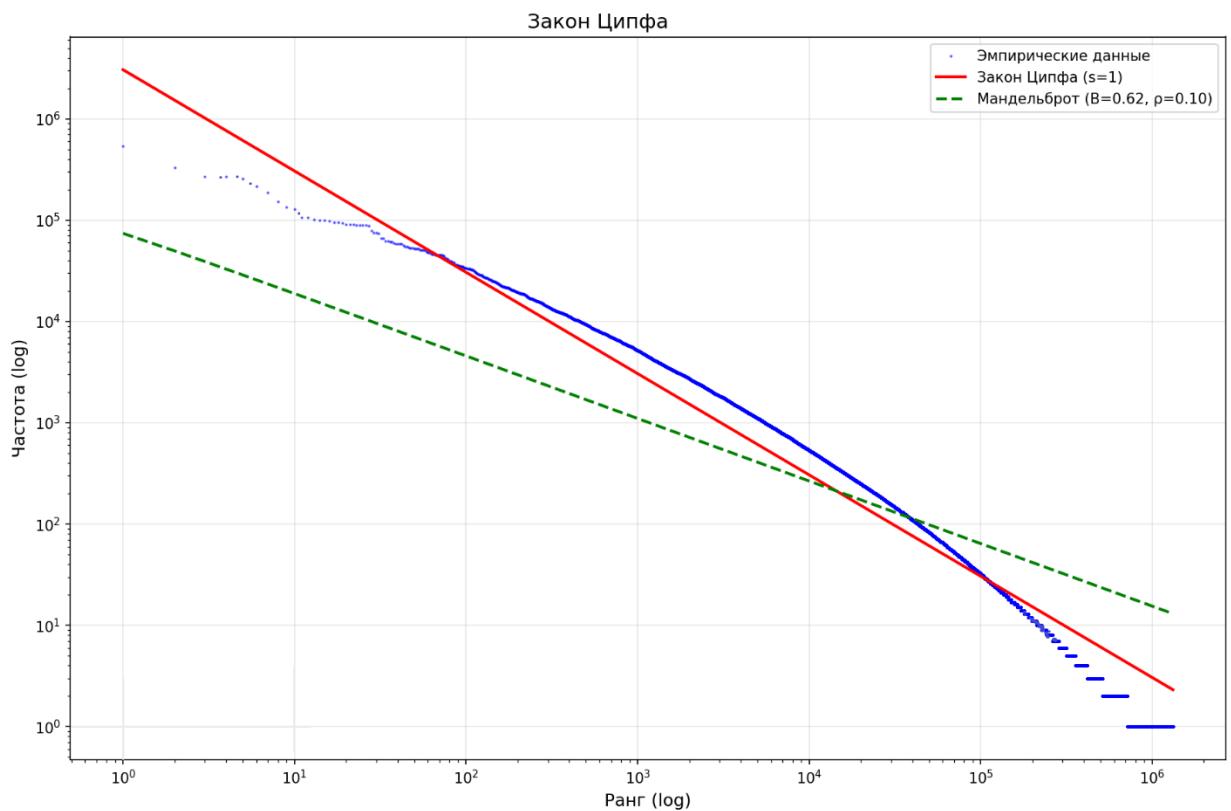


Рисунок 5 – закон Ципфа

Стемминг

Плюсы:

- Быстрый поиск
- Меньший размер индекса (меньше уникальных термов)

Минусы:

- Нельзя искать по точной форме слова

Запрос: пианист

Без стемминга: будет найдено только «пианист»

Со стеммингом: «пиан»: Будет найдено пианист, пианино, фортепиано.

Булев индекс

ПРЯМОЙ ИНДЕКС (Forward Index):

Для каждого документа:

4 байта - doc_id

2 байта - длина title

N байт - title (UTF-8)

2 байта - длина url

M байт - url (UTF-8)

ОБРАТНЫЙ ИНДЕКС (Inverted Index):

4 байта - количество термов

Для каждого терма:

1 байт - длина терма

N байт - терм (UTF-8)

4 байта - количество документов (DF)

4*DF байт - массив doc_id (отсортирован)

Булев поиск

В наличии есть следующие операции: AND, OR, NOT, группировка(круглые скобки).

Результатами поиска являются ссылками на источники

Примеры поиска

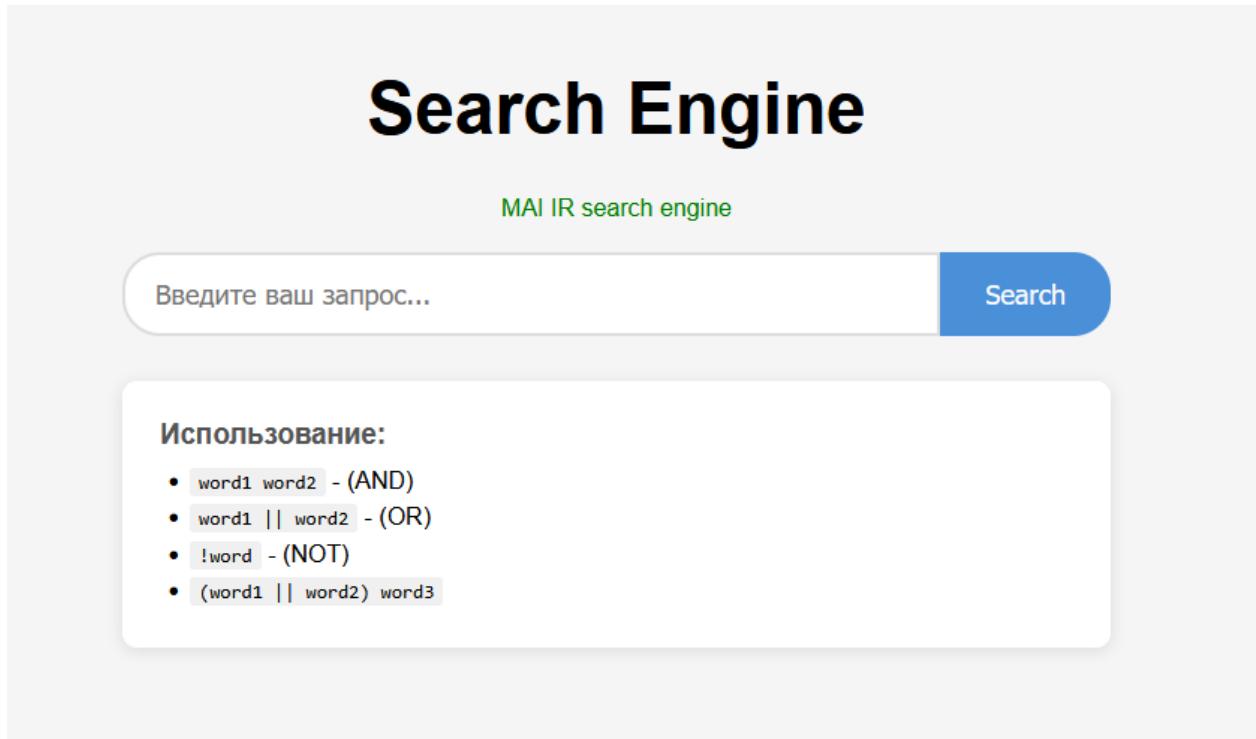


Рисунок 6 – web-интерфейс

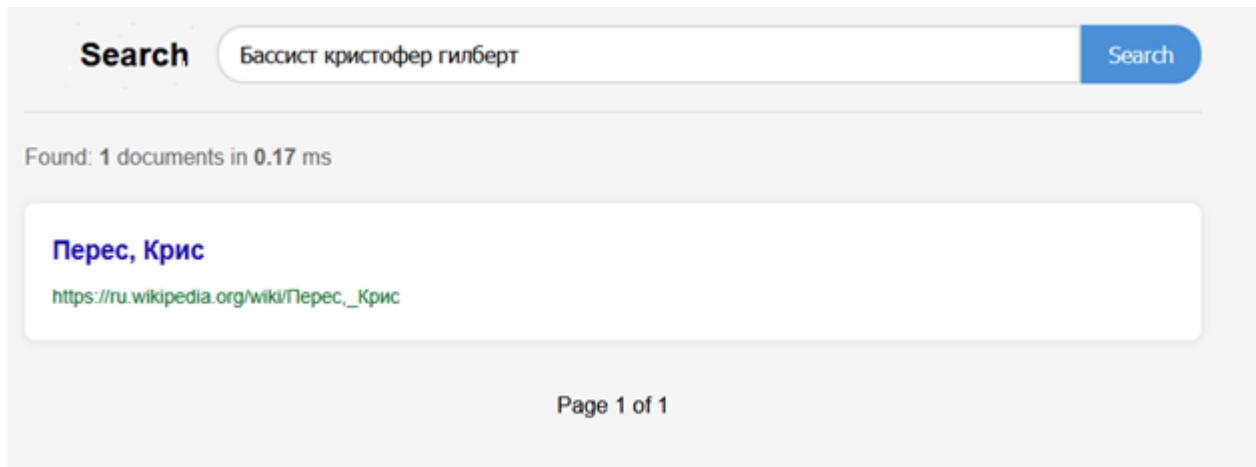


Рисунок 7 – результаты поиска

Search

певец артист

Search

Found: 3185 documents in 1.57 ms

Павел Захарович Андреев | Belcanto.ru

<https://www.belcanto.ru/andreev.html>

Зураб Иванович Анджапаридзе (Zurab Andzshaparidze) | Belcanto.ru

<https://www.belcanto.ru/andzhaparidze.html>

Марио Дель Монако | Belcanto.ru

<https://www.belcanto.ru/delmonaco.html>

Пермские «Риголетто» и «Паяцы» в Московской филармонии | Belcanto.ru

<https://www.belcanto.ru/25120401.html>

Джоаккино Россини | Belcanto.ru

<https://www.belcanto.ru/rossini.html>

Опера Сергея Рахманинова «Алеко» | Belcanto.ru

<https://www.belcanto.ru/aleko.html>

Нижегородский театр оперы и балета (Nizhny Novgorod Opera and Ballet Theatre) | Belcanto.ru

<https://www.belcanto.ru/nizhegorodsky.html>

Саратовский академический театр оперы и балета | Belcanto.ru

<https://www.belcanto.ru/saratov.html>

Опера Антона Рубинштейна «Демон» | Belcanto.ru

<https://www.belcanto.ru/demon.html>

Рисунок 8 - результаты поиска «певец артист»

Search

певец !артист

Search

Found: 9431 documents in 3.04 ms

[Луи Армстронг | Belcanto.ru](#)

https://www.belcanto.ru/armstrong_louis.html

[Жорж Брассенс | Belcanto.ru](#)

<https://www.belcanto.ru;brassens.html>

[Алессандро Бончи | Belcanto.ru](#)

<https://www.belcanto.ru/bonci.html>

[Уго Бенелли \(Ugo Benelli\) | Belcanto.ru](#)

<https://www.belcanto.ru/benelli.html>

[О «Кармен» в Королевском театре Мадрида | Belcanto.ru](#)

<https://www.belcanto.ru/25121902.html>

[Театр «Ла Скала» в Милане | Belcanto.ru](#)

<https://www.belcanto.ru/scala.html>

[В эмпиреях бельканто с Альбиной Шагимуратовой | Belcanto.ru](#)

<https://www.belcanto.ru/25120901.html>

[«Ночь перед Рождеством» в Баварской опере | Belcanto.ru](#)

<https://www.belcanto.ru/25120702.html>

[Турок в жёлтом и синем | Belcanto.ru](#)

<https://www.belcanto.ru/25120202.html>

[Опера Флотова «Алессандро Страделла» | Belcanto.ru](#)

<https://www.belcanto.ru/alessandro.html>

Рисунок 9 – результаты поиска «певец !артист»



Search

певец

Search

Found: 12616 documents in 1.35 ms

Павел Захарович Андреев | Belcanto.ru

<https://www.belcanto.ru/andreev.html>

Луи Армстронг | Belcanto.ru

https://www.belcanto.ru/armstrong_louis.html

Зураб Иванович Анджапаридзе (Zurab Andzshaparidze) | Belcanto.ru

<https://www.belcanto.ru/andzhaparidze.html>

Рисунок 10 – результаты поиска певец

Выводы

В процессе выполнения лабораторных работ были изучены процесс работы современных поисковых систем, процесс индексации. Была реализована своя поисковая система для конкретного корпуса данных.