

МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)
Институт №8 «Компьютерные науки и прикладная математика»

Лабораторная работа
по курсу «Информационный поиск»

Выполнил: Козырев П.А

Группа: М8О-412Б-22

Преподаватели: Кухтичев А.А.

Москва, 2025

ЦЕЛЬ РАБОТЫ

- Найти источники, которые будут использоваться в работе
- Реализовать поискового робота для получения корпуса документа
- Подготовить документы к дальнейшему использованию:
 - выделить текст
 - убрать ненужную информацию
- Привести несколько примеров запросов к существующим поисковикам, указать недостатки в полученной поисковой выдаче.
- Реализовать токенизацию
- Реализовать стемминг
- Реализовать булев индекс и булев поиск

ОПИСАНИЕ ДАННЫХ

В качестве источника данных я выбрал 3 интернет-ресурса на тему шахмат:

- ChessPro Main - <https://chesspro.ru/>
- Chess World - <https://chess-world.net/>
- Obninsk Chess Forum - <https://www.obninskchess.ru/forum2/>

На данные сайты посвящены шахматам, на них есть много различной информации: новости, статьи, форумы и т.д.

Для получения данных я написал скрипт на Python, который с помощью сторонних библиотек, таких как BeautifulSoup, requests, urllib.parse проходил по всем ссылкам на сайте и складывал html-код страницы в MongoDB.

Для очистки данных и получения чистого текста я написал еще один скрипт, который удаляет всю техническую информацию (html-теги) и складывает полученный текст в БД.


После обкатки корпуса документов и парсинга текста из полученных данных я собрал соответствующую статистику:

Всего документов	33 258
Общий объем «сырых данных»	3.5 Gb
Общий объем чистого текста	1.32 Gb
Средний размер сырого документа	110.54 KB
Средний объем текста	41.7 KB
Среднее количество символов	23 922
Среднее количество слов	2 215

ПРИМЕР ПОИСКА НА ИСПОЛЗУЕМОМ РЕСУРСЕ

На сайте реализован булев поиск, вот такие результаты он выдал для двух запросов:

- шахматы AND ферзь
- королева OR мат



Шахматы

Новости, турниры, анонсы, таблицы, положения, рейтинги

[RSS](#) [FAQ](#) [Поиск](#) [Пользователи](#) [Группы](#) [Регистрация](#)
[Профиль](#) [Войти и проверить личные сообщения](#) [Вход](#)

Форум находится в режиме "только для чтения"!

Список форумов «Шахматы»

Запрос

шахматы AND ферзь

☒ Искать любое слово/поиск с языком запросов
☐ Искать все слова

Ключевые слова:
Вы можете использовать **AND** чтобы определить слова, которые должны быть в результатах, **OR** для слов, которые могут быть в результатах, и **NOT** для слов, которых в результатах быть не должно. Используйте * в качестве шаблона для частичного совпадения.


Результатов поиска: 9

Список форумов «Шахматы»

	Форум для всех любителей шахмат	Темы	Автор	Ответов	Просмотров	Последнее сообщение
1	Виды шахмат	Гексагональные шахматы Глинского	drovosek777	2	11317	Бс Apr 05, 2020 10:30 Lunika ➡
2	Виды шахмат	Решения с опровержениями. [На страницу: 1, 2, 3, 4]	в.прокопий	47	92254	Пт Mar 29, 2019 01:28 admin_forum ➡
3	Виды шахмат	стоклеточные шахматы	nd	2	11306	Пн Июн 20, 2016 10:27 admin_forum ➡
4	Виды шахмат	Нетрадиционные шахматы	drovosek777	1	10660	Бс Apr 21, 2013 18:12 admin_forum ➡
5	Новости шахмат	О близе не сша	admin_forum	0	4638	Ср Яне 25, 2012 07:19 admin_forum ➡
6	Виды шахмат	Нерешённый этюд [На страницу: 1, 2, 3, 4]	в.прокопий	58	113520	Пт Июн 18, 2010 00:22 в.прокопий ➡
7	История шахмат	Калужско-шахматная ВИКТОРИНА [На страницу: 1, 2, 3, 4]	CT	50	105410	Пн Фев 02, 2009 14:08 CT ➡
8	Виды шахмат	Поединок на форушет (pocketchess - Lusy) [На страницу: 1 -- 15, 16, 17]	Lusy	244	263721	Пт Mar 28, 2008 06:30 Lusy ➡
9	Новости шахмат	Шахматный Фестиваль Вейк-ан-Зее 2008. Обсуждение [На страницу: 1, 2, 3]	z01	42	49521	Вт Яне 29, 2008 09:06 z01 ➡

Страница 1 из 1

Часовой пояс: GMT + 3



Шахматы

Новости, турниры, анонсы, таблицы, положения, рейтинги

[RSS](#) [FAQ](#) [Поиск](#) [Пользователи](#) [Группы](#) [Регистрация](#)
[Профиль](#) [Войти и проверить личные сообщения](#) [Вход](#)

Форум находится в режиме "только для чтения"!

Список форумов «Шахматы»

Запрос

королева OR мат

☒ Искать любое слово/поиск с языком запросов
☐ Искать все слова

Ключевые слова:
Вы можете использовать **AND** чтобы определить слова, которые должны быть в результатах, **OR** для слов, которые могут быть в результатах, и **NOT** для слов, которых в результатах быть не должно. Используйте * в качестве шаблона для частичного совпадения.

Поиск по автору:
Используйте * в качестве шаблона

Результатов поиска: 79

Список форумов «Шахматы»

	Форум для всех любителей шахмат	Темы	Автор	Ответа	Просмотров	Последнее сообщение
1	История шахмат	Форум "История шахмат". Каким его хочется видеть	Гость	9	56481	Пт Mar 16, 2018 22:40 thatchess ➡
2	Общие вопросы	Вопросы о правилах шахмат	Рал	13	26150	Пн Яне 22, 2018 09:04 Рал ➡
3	Детские шахматы	Методы работы [На страницу: 1, 2]	kora74	23	53626	Вт Окт 13, 2015 18:17 Olga ➡
4	Шахматы с компьютером	Играть в шахматы на компьютере	Bolya	1	14101	Чт Июл 30, 2015 22:10 admin_forum ➡
5	Виды шахмат	Мат в 3 хода [На страницу: 1, 2]	Дмитрий Данилен	15	49775	Пт Яне 09, 2015 00:07 admin_forum ➡
6	Шахматы в интернете	Магазин интеллектуальных игр "ШАХ и МАТ"	catalogchess	1	11811	Ср дек 10, 2014 19:35 admin_forum ➡
7	Новости шахмат	Финал годовда Тобольска	Lusy	8	21929	Чт Фев 13, 2014 20:23 Lusy ➡

ТОКЕНИЗАЦИЯ И СТЕММИНГ

Для обработки текста был использован метод токенизации на основе фильтрации символов кириллицы и суффиксальный стемминг.

Линейный проход по тексту позволяет обрабатывать большие объемы данных за минимальное время. Также стемминг значительно сокращает размер инвертированного индекса, объединяя различные словоформы в один терм.

Однако надо указать и на недостатки: суффиксальный метод может ошибочно отрезать часть корня (оверстемминг) или не до конца нормализовать слово (андерстемминг).

Общее количество токенов: 44 424 762

Средняя длина токена: 5.49 символов

Скорость токенизации: 15317 КБ/сек

Время выполнения: 90.54 сек

График распределения частот терминов (Закон Ципфа):



РЕАЛИЗАЦИЯ ИНДЕКСА И БУЛЕВОГО ПОИСКА

В основе поисковой системы лежит структура инвертированного индекса. В отличие от прямого индекса, инвертированный индекс сопоставляет каждому уникальному слову список идентификаторов документов, в которых оно встречается.

В моей работе индекс реализован с использованием собственных структур данных: Внешний контейнер: хеш-таблица, где ключом является нормализованный терм (после стемминга), а значением - множество документов. Использование хеш-таблицы обеспечивает поиск термина в среднем за время $O(1)$. Внутренний контейнер: множество уникальных идентификаторов документов. Использование хеш-множества позволяет выполнять проверку наличия документа в списке за $O(1)$, что важно для операций пересечения множеств.

Булев поиск позволяет находить документы, удовлетворяющие логическим условиям (AND, OR, NOT). Алгоритм реализован как последовательный обработчик токенов поискового запроса.

Основные этапы поиска:

- Парсинг запроса: Строка запроса разбивается на токены. Выделяются ключевые слова и логические операторы.
- Нормализация: Каждое слово запроса проходит через стемминг (приведение к основе), чтобы соответствовать ключам в инвертированном индексе.
- Итеративная обработка: Программа хранит промежуточный результат и текущий активный оператор (по умолчанию AND)

РЕЗУЛЬТАТ

Пример работы программы:

```
Введите булев запрос (или exit): шахматы AND ферзь
Найдено документов: 1008
- http://chesspro.ru/_events/2010/tkachenko2.html
- http://chesspro.ru/_events/2010/linares5.html
- http://chesspro.ru/_events/2012/emelin2_hit.html
- http://chesspro.ru/_events/2008/voronkov_alekhine_5.html
- http://chesspro.ru/_events/2007/krasnotur4.html
- https://chesspro.ru/_events/2007/tal1.html
- https://chesspro.ru/guestnew/lookmessage/?id=16-650-379450
- https://chesspro.ru/guestnew/lookmessage/?id=16-650-379432
- http://chesspro.ru/drupal7/enciklopediya/dramy-bez-shedevrov
- https://chesspro.ru/guestnew/lookmessage/?id=16-684-380277
```

```
Введите булев запрос (или exit): королева OR мат
Найдено документов: 79
- http://chesspro.ru/drupal7/hitparad/prusikin_doctor
- http://chesspro.ru/details/tkachenko_studies305
- https://chesspro.ru/_events/2011/bologan1_hit.html
- https://chesspro.ru/_events/2008/mainz3.html
- https://chesspro.ru/_events/2008/weik29.html
- http://chesspro.ru/_events/2008/zinar.html
- http://chesspro.ru/drupal7/details/tkachenko_studies172
- http://www.chesspro.ru/details/tkachenko_studies115
- https://chesspro.ru/_events/2010/grabuzova_obzor3_tur.html
```

ЗАКЛЮЧЕНИЕ

В ходе выполнения лабораторной работы был разработан поисковый робот, осуществляющий сбор и обработку корпуса документов. В процессе работы был реализован парсинг документов, включающий извлечение текстовой информации и её предварительную обработку. Для нормализации текстовых данных были применены методы токенизации и стемминга, что позволило привести слова к базовой форме и сократить размер словаря.

На основе обработанных данных был построен булев индекс, обеспечивающий эффективное хранение информации о вхождении терминов в документы. Также был реализован механизм булевого поиска, позволяющий выполнять поисковые запросы с использованием логических операций. В результате выполненной работы была изучена и практически освоена базовая архитектура информационно-поисковых систем, а также получены навыки разработки и анализа алгоритмов индексирования и поиска текстовой информации.