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 документова
- 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 проходился по всем ссылкам на сайте и складывал html-код страницу 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.

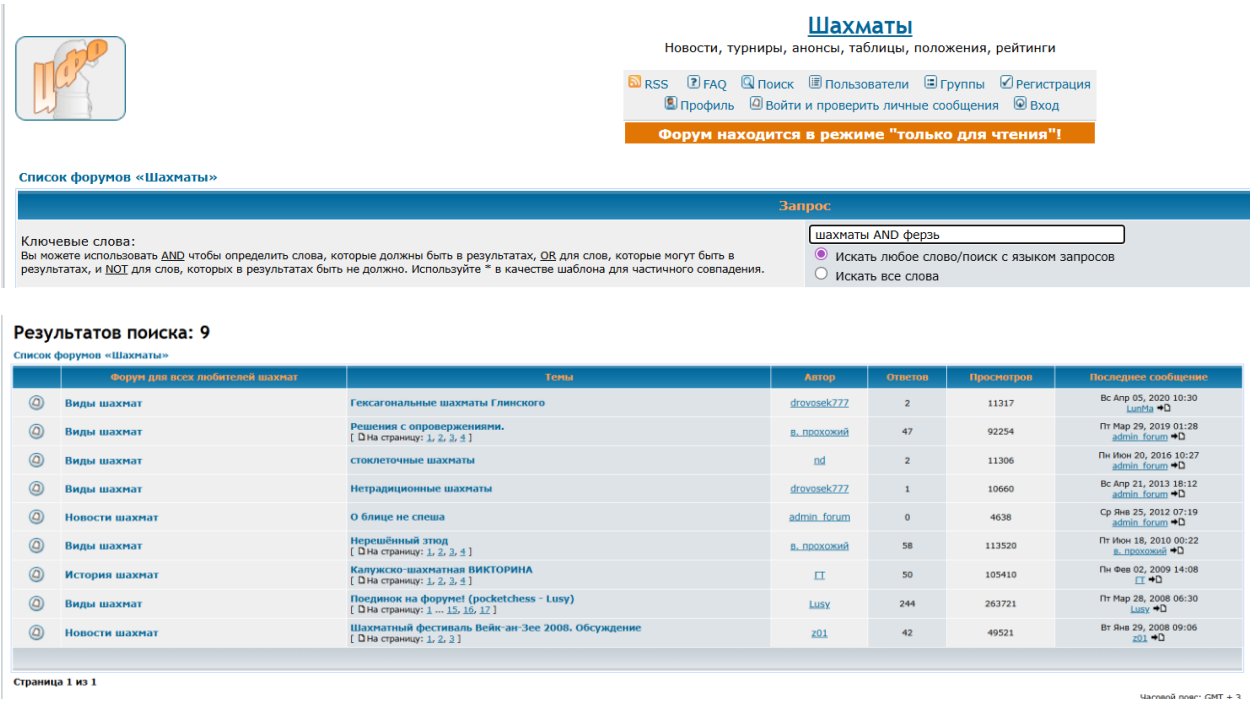
После обкачки корпуса документов и парсинга текста из полученных данных я собрал соответствующую статистику:

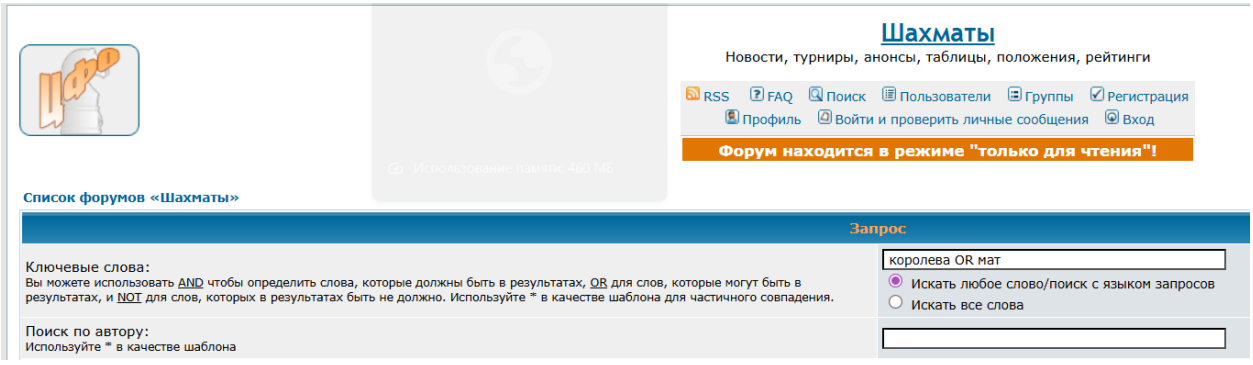
Всего документов	33 258
Общий объем «сырых данных»	3.5 Gb
Общий объем чистого текста	1.32 Gb
Средний размер сырого документа	110.54 KB
Средний объем текста	41.7 KB
Среднее количество символов	23 922
Среднее количество слов	2 215

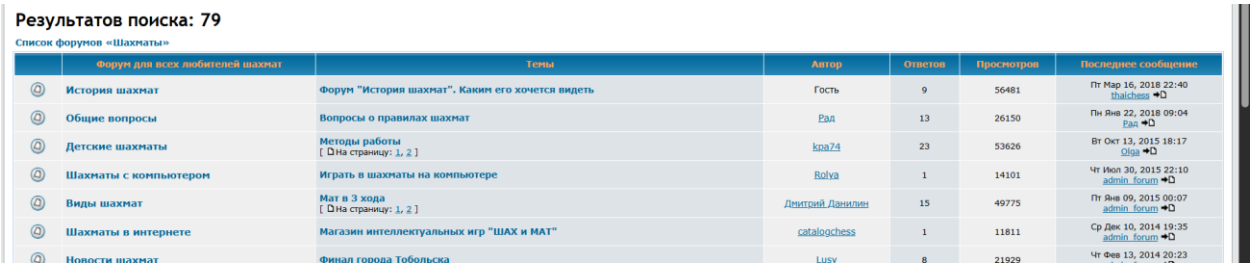
ПРИМЕР ПОИСКА НА ИСПОЛЬЗУЕМОМ РЕСУРСЕ

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

- шахматы AND ферзь
- королева OR мат







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).

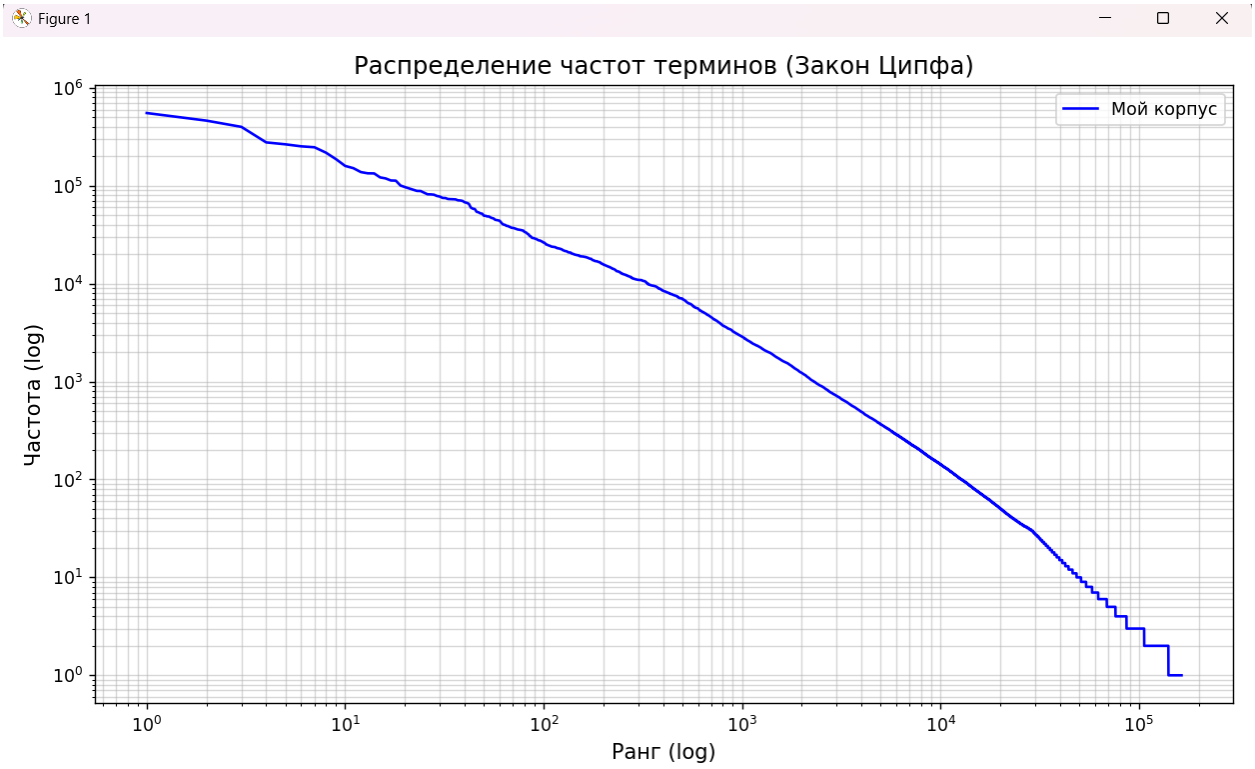
Общее количество токенов: 44 424 762

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 терма 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.

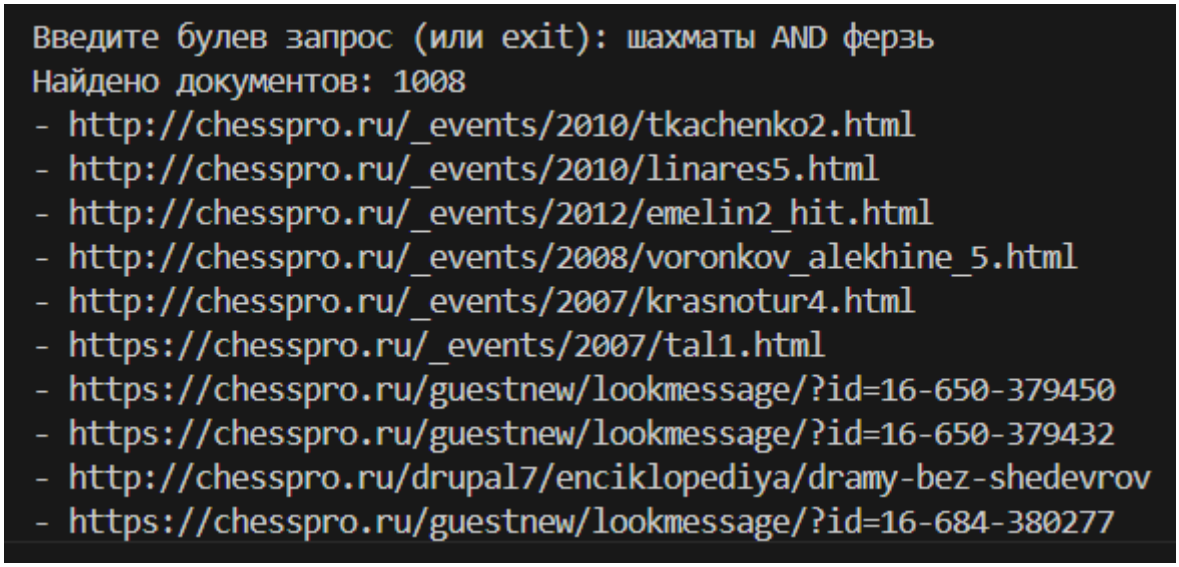
Булев поиск позволяет находить документы, удовлетворяющие логическим условиям (AND, OR, NOT). Алгоритм реализован как последовательный обработчик токенов поискового запроса.

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): 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.