МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

Институт №8 «Компьютерные науки и прикладная математика»

Лабораторные работы
по курсу «Информационный поиск»

Выполнил: Жиденко Александр Сергеевич
Группа: М8О-401Б-22
Преподаватель: Кухтичев Антон Алексеевич

Москва, 2025

Содержание

1 Введение

В рамках курса «Информационный поиск» реализована полноценная учебная поисковая система на базе корпуса статей Lenta.ru. Все этапы — от добычи данных до индексации и булева поиска — выполнены на языке Python (Flask для веб-интерфейсов), без привлечения сторонних библиотек для основных структур данных индексации и поиска. Дополнительно подготовлены CLI-утилиты и веб-прототипы для демонстрации работы каждой лабораторной.

2 Лабораторная №1: Добыча корпуса документов

2.1 Цель

Собрать корпус не менее 30 000 текстовых документов единой тематики для дальнейшей индексации.

2.2 Источник и структура

Использован архив новостей Lenta.ru. Каждый документ содержит заголовок, источник, URL, дату и основной текст в UTF-8.

2.3 Реализация

Написан скрипт-коллектор (Python, requests, BeautifulSoup) с повторными попытками, логированием и сохранением состояния. Документы сохраняются в пары `doc_XXXXXX.txt` / `doc_XXXXXX.meta.json`.

2.4 Результаты

Собрано 30000 документов для разработки.

3 Лабораторная №2: Поисковый робот

3.1 Цель

Создать веб-краулер с вежливостью, поддержкой `robots.txt`, очередью URL и сохранением статуса.

3.2 Реализация

Класс `WebCrawler` реализует:

- разбор `robots.txt` и учёт `crawl-delay`;
- нормализацию URL и дедупликацию;
- извлечение ссылок и текста (несколько CSS-селекторов + fallback по `body`);
- сохранение документов в формат корпуса;
- CLI и веб-интерфейс (Flask) для запуска и мониторинга.

3.3 Результаты

Получен воспроизводимый краулер с логами, резюмированием и статистикой посещённых URL.

4 Лабораторная №3: Токенизация

4.1 Цель

Разбить текст на токены, нормализовать регистр, удалить пунктуацию и (опционально) стоп-слова.

4.2 Реализация

Класс `Tokenizer`:

- regex для русских/английских слов и чисел;
- опции: `lowercase`, `remove_punctuation`, `min_length`, `remove_stopwords`;
- подсчёт частот и словаря;
- пакетная обработка корпуса с сохранением статистики и токенов.

4.3 Результаты

Подготовлены токены и частоты для дальнейших лабораторных; есть CLI и веб-демо.

5 Лабораторная №4: Стемминг

5.1 Цель

Привести токены к базовой форме с помощью стемминга.

5.2 Реализация

Класс `Stemmer` (русский — правилоудаление суффиксов; английский — упрощённый Porter-like):

- обработка рефлексивных, прилагательных, глагольных и именных суффиксов;
- частоты основ, словарь основ, отображение токен $\rightarrow$ основа;
- пакетная обработка корпуса, CLI и веб-интерфейс.

5.3 Результаты

Уменьшено количество уникальных форм, подготовлены данные для индексации.

6 Лабораторная №5: Закон Ципфа

6.1 Цель

Проверить распределение частот токенов и соответствие закону Ципфа.

6.2 Реализация

Класс `ZipfAnalyzer`:

- расчёт частот, рангов, константы C и корреляции;
- графики log–log и rank–frequency (matplotlib, numpy);
- экспорт статистики и графиков, CLI и веб-интерфейс с встраиваемыми изображениями.

6.3 Результаты

Получено распределение частот, подтверждающее ожидаемую гиперболическую зависимость (корреляция для корпуса порядка 30 000 документов ожидается $> 0{,}8$).

7 Лабораторная №6: Булев индекс

7.1 Цель

Построить булев индекс «термин $\rightarrow$ множество документов».

7.2 Реализация

Класс `BooleanIndex`:

- построение из корпуса с токенизацией и (опц.) стеммингом;
- сериализация/загрузка в JSON, текстовый экспорт;
- статистика: объём, топ-термины, среднее число термов на документ.

7.3 Результаты

Индекс готов для булева поиска; предусмотрены CLI и веб-обвязка.

8 Лабораторная №7: Булев поиск

8.1 Цель

Реализовать обработку булевых запросов (AND, OR, NOT) на основе индекса.

8.2 Реализация

Класс `BooleanSearch`:

- разбор запроса в постфиксную нотацию с приоритетом NOT > AND > OR;
- оценка через операции над множествами документов;
- выдача с метаданными, лимитирование результатов;
- CLI (одиночный, пакетный, интерактивный) и веб-интерфейс (Flask).

8.3 Результаты

Полностью функциональный булев поиск по корпусу. Запросы с несколькими операторами и скобками поддерживаются.

9 Заключение

Реализован полный цикл учебной поисковой системы: сбор данных, подготовка текста, стемминг, статистический анализ, построение индекса и булев поиск. Подготовлены CLI-утилиты, веб-интерфейсы и автотесты для каждого этапа. Код и структура данных позволяют масштабировать корпус до 30 000+ документов, сохраняя совместимость с отчётными требованиями курса.