МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ

Федеральное государственное бюджетное образовательное учреждение высшего образования

«Московский Авиационный Институт»

(Национальный Исследовательский Университет)

Институт: №8 «Информационные технологии и прикладная математика»
Кафедра: 806 «Вычислительная математика и программирование»

Отчет по лабораторной работе №1

по предмету «Информационный поиск»

«Добыча корпуса документов»

Группа: М8О-412Б-22

Студент: Муратов А.А.

Оценка:

Дата сдачи:

Москва, 2025

Задание

Необходимо проанализировать корпус документов, который будет использован при выполнении остальных лабораторных работ:

- Скачать примеры документов к себе на компьютер. В отчёте нужно указать источник данных. Источников в итоговом индексе должно быть не менее двух;
- Ознакомиться с ним, изучить его характеристики. Из чего состоит текст? Есть ли дополнительная мета-информация? Если разметка текста, какая она?
- Выделить текст.
- Найти существующие поисковики, которые уже можно использовать для поиска по выбранному набору документов (встроенный поиск Википедии, поиск Google с использованием ограничений на URL или на сайт). Если такого поиска найти невозможно, то использовать корпус для выполнения лабораторных работ нельзя!
- Привести несколько примеров запросов к существующим поисковикам, указать недостатки в полученной поисковой выдаче.

В результатах работы должна быть указаны статистическая информация о корпусе:

- Размер примеров «сырых» документов.
- Количество документов.
- Размер текста, выделенного из «сырых» данных.
- Средний размер документа, средний объём текста в документе.

Описание задания

Тема:

Статьи из российских СМИ. Выбранные мной источники - популярные СМИ, имеющие хорошо поддерживаемую структуру HTML и структуру сайтов.

Источники данных: www.interfax.ru,tass.ru.

Что внутри документов:

- Interfax: стандартный HTML с og:* мета-тегами, article:published_time, article:tag, canonical/amp ссылки, favicon и CSS/JS. Основной текст в блоке статьи, есть дата/время, рубрика, теги.

- TASS: HTML, собранный Next.js, содержит og:* и JSON-LD (application/ld+json), прелоады шрифтов/изображений, data-v атрибуты. Основной текст и заголовок в блоках с классами article-page, присутствуют дата, автор/редакция, рубрика. Исходный HTML был

минимизирован в одну строку. Большой размер документов по сравнению со вторым источником.

Доступные поисковики:

- Встроенный поиск Interfax: https://www.interfax.ru/search/?sw=запрос (ищет по всему сайту, без сложных фильтров).

- Встроенный поиск TASS: https://tass.ru/search?text=запрос (есть сортировка по дате/релевантности).

- Внешний: Google/Yandex с ограничением `site:interfax.ru` или `site:tass.ru`.

Примеры запросов и ограничения:

- `site:interfax.ru аэропорт Нижний Новгород`; выдача смешивает новости и обзоры, часто есть дубли по AMP/print.

- `site:tass.ru экономика нефть`; попадают лонгриды и дайджесты, нет контроля по дате, встречаются мультимедийные страницы без текста.

- `interfax поиск "Москва"` во встроенном поиске: нет фильтрации по рубрике, много старых материалов.

- `tass поиск "военная операция"` во встроенном поиске: релевантность слабая, выдаёт фотоленты и видео, которые плохо подходят для текстового корпуса.

Статистика корпуса:

Источник	Кол-во	Сырые байты (сум/сред)	Текст, символы (сум/сред)
InterFax	5	303 721 / 60 744	27 146 / 5 429
TASS	5	2 832 338 / 566 468	42 454 / 8 491

Выводы

Корпус документов используемый в работе пригоден для выполнения дальнейших работ, существует потенциал для улучшения поиска по данным сайтам.