МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

Институт №8 «Компьютерные науки и прикладная математика»

Лабораторная работа №*1*
по курсу «Информационный поиск»

«Добыча корпуса документов»

Выполнил: *Ермаков Ярослав Валерьевич*
Группа: *М8О-407Б-22*
Преподаватели: *А.А. Кухтичев*

Москва, *2025*

Условия

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

- Скачать его к себе на компьютер. В отчёте нужно указать источник данных.
- Ознакомиться с ним, изучить его характеристики. Из чего состоит текст? Есть ли дополнительная мета-информация? Если разметка текста, какая она?
- Разбить на документы.
- 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 «сырых» данных.
- Количество документов.
- Размер текста, выделенного из «сырых» данных.
- Средний размер документа, средний объём текста в документе.

Описание

Целью работы является подготовка корпуса документов для последующего использования в лабораторных работах по информационному поиску. В ходе выполнения работы требуется выбрать не менее двух источников документов, загрузить примеры документов, изучить структуру данных, выделить текст из “сырого” формата и собрать статистические характеристики корпуса.

В качестве источников данных были выбраны два независимых открытых научных источника: **ACL Anthology** и **arXiv**. Эти источники подходят для задач поиска, так как содержат большое количество документов с устойчивыми URL, имеют доступ к HTML-страницам публикаций

- **ACL Anthology** - открытая библиотека публикаций по компьютерной лингвистике и смежным областям. Документы доступны по стабильным URL вида https://aclanthology.org/... и представлены в HTML с заголовком публикации и текстовым содержимым страницы.
- **arXiv** - открытый архив научных статей. Документы доступны по URL вида https://arxiv.org/abs/<id> и представлены HTML-страницами с метаинформацией и содержательными фрагментами.

По результатам выгрузки корпуса получены следующие статистические характеристики:

1. **Размер «сырых» данных:** 529 704 878 байт
2. **Количество документов:** 58 866.
3. **Размер текста, выделенного из «сырых» данных:** 529 704 878 байт. Выделение текста из HTML-разметки выполнялось на этапе подготовки корпуса, поэтому дальнейшие измерения проводились по уже очищенному текстовому представлению документов, используемому в последующих лабораторных работах.
4. **Средний размер документа:** 8998,49 байт.
5. **Средний объём текста в документе:** около 9 КБ очищенного текстового содержимого

Исходный код

Для первой ЛР использовалась утилита export_corpus.py, предназначенная для подготовки корпуса документов к дальнейшим лабораторным работам. Программа подключается к MongoDB, выбирает последнюю сохранённую версию HTML-документа для каждого URL, извлекает из HTML видимый текст и заголовок страницы, нормализует пробелы и сохраняет очищенный текст в файлы docs/*.txt. Дополнительно формируется таблица meta.tsv с метаданными (URL, источник, время загрузки, заголовок, длина текста), а при необходимости сохраняется и “сырой” HTML в сжатом виде.

Выводы

В ходе первой лабораторной работы был выбран и проанализирован корпус документов из двух открытых источников - ACL Anthology и arXiv. Были изучены структура исходных документов, наличие метаинформации и существующие средства поиска по данным источникам. Документы сохранены в исходном формате и подготовлены для последующей обработки и использования в дальнейших лабораторных работах по информационному поиску.