Московский авиационный институт (национальный исследовательский университет)

Факультет информационных технологий и прикладной математики

Кафедра вычислительной математики и программирования

Лабораторная работа №1 по курсу «Информационный поиск»

Студент: Н. С. Федоров
Преподаватель: А. А. Кухтичев
Группа: М8О-410Б

Дата:
Оценка:
Подпись:

Москва, 2025

Лабораторная работа №1 «Добыча корпуса документов»

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.

1 Описание

В рамках лабораторной работы был подготовлен корпус документов для дальнейшего использования в задачах информационного поиска. В качестве источников данных были выбраны сайты *Billboard.com* и *Pitchfork.com*, представляющие собой крупные англоязычные онлайн-издания, публикующие новости, обзоры и аналитические материалы, посвящённые музыкальной индустрии.

Данные сайты были выбраны по следующим причинам: регулярное обновление контента, наличие большого количества текстовых материалов, чёткая структура статей, а также существование встроенных и внешних средств поиска, что делает корпус пригодным для выполнения последующих лабораторных работ.

Для формирования корпуса HTML-страницы были разобраны и очищены от служебной информации. Из документов были удалены элементы навигации, рекламные блоки, скрипты и иные части страницы, не относящиеся к содержанию статьи. В результате из каждого документа был выделен чистый текст, включающий заголовок, авторов и основной текст статьи.

Для выбранных источников существуют готовые поисковые системы. Оба сайта обладают встроенным поиском по опубликованным материалам. Кроме того, поиск по данным корпусам возможен с использованием внешних поисковых систем, таких как Google и Яндекс, с применением ограничений на домен (например, `site:billboard.com` или `site:pitchfork.com`). Таким образом, корпус удовлетворяет требованиям лабораторной работы и может быть использован в дальнейших заданиях.

Для выбранного корпуса существуют готовые поисковые системы, которые могут быть использованы для поиска по документам. Оба сайта-источника, *Billboard.com* и *Pitchfork.com*, обладают встроенными средствами поиска по опубликованным материалам. Кроме того, поиск по данным ресурсам возможен с использованием внешних поисковых систем, таких как Google и Яндекс, с применением ограничений на домен.

В качестве примера был рассмотрен поисковый запрос «Justin Bieber». На рисунках 1 и 2 представлена поисковая выдача Google и Яндекса соответственно, полученная с использованием ограничений на сайт музыкальных изданий. На рисунках 3 и 4 показаны результаты встроенного поиска сайтов Billboard и Pitchfork.

В результате выполнения работы была получена следующая статистическая информация о корпусе:

- Размер сырых данных: примеры HTML-документов имеют размер около 1051 КБ для Billboard и около 715 КБ для Pitchfork.
- Общее количество документов: около 30 000.
- Размер выделенного текста: для примеров документов Billboard объём очищенного текста составляет 5 КБ, Pitchfork - 3 КБ.

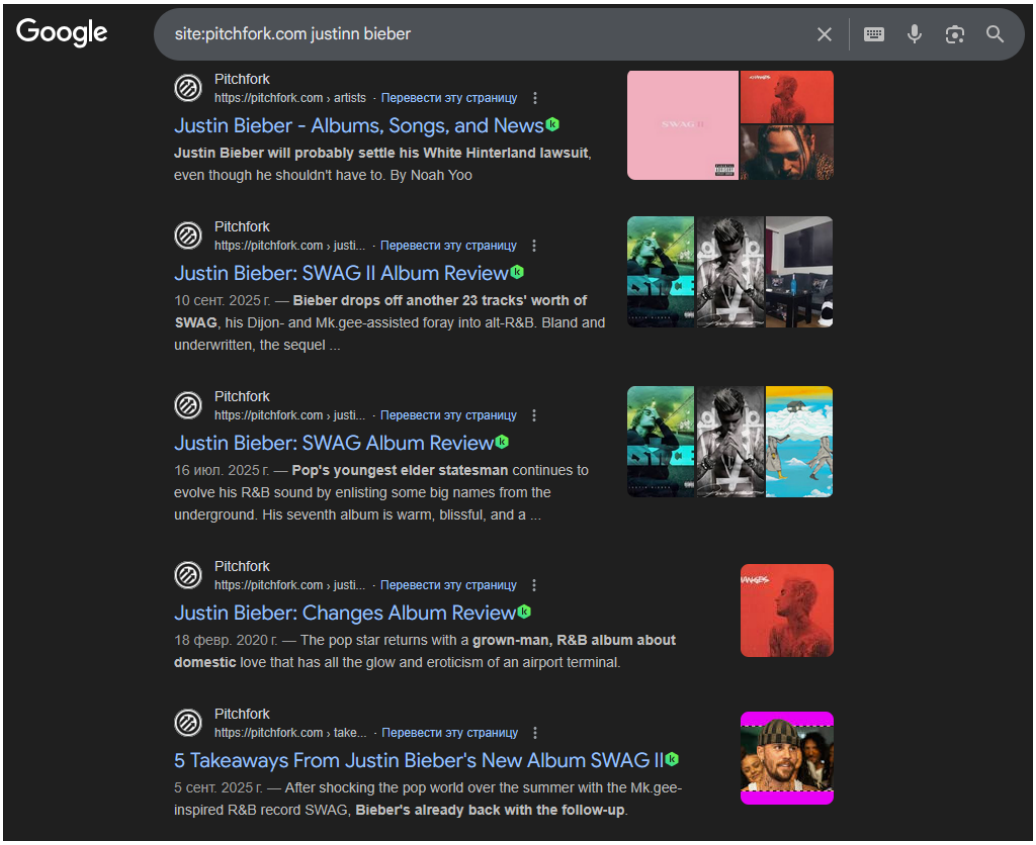

Рис. 1: Результаты поиска запроса «Justin Bieber» в поисковой системе Google.

- Средние значения размера документа и объёма текста в документе будут рассчитаны на последующих этапах работы.

Рис. 2: Результаты поиска запроса «Justin Bieber» в поисковой системе Яндекс.

2 Исходный код

Для подготовки корпуса была разработана программа на языке Python, предназначенная для обработки HTML-документов и извлечения текстового содержимого ста-

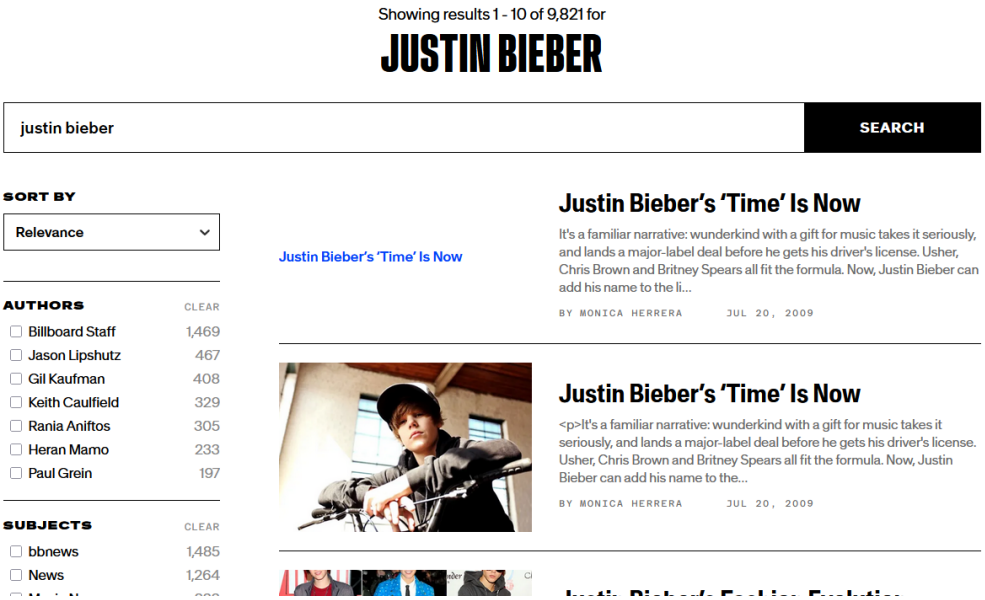

Рис. 3: Результаты встроенного поиска сайта Billboard.com по запросу «Justin Bieber».

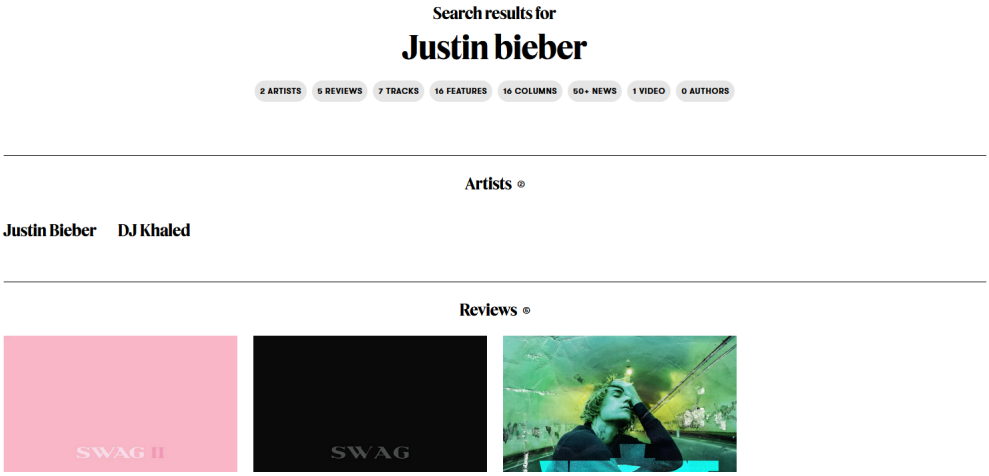

Рис. 4: Результаты встроенного поиска сайта Pitchfork.com по запросу «Justin Bieber».

тей. Работа программы организована в виде последовательного конвейера обработки данных.

На первом этапе осуществляется загрузка и разбор HTML-страниц с использованием библиотеки для парсинга HTML-документов. Каждая страница обрабатывается как отдельный документ корпуса.

На следующем этапе выполняется извлечение структурированных элементов статьи: заголовка, информации об авторах и основного текстового содержимого. Для этого используются характерные HTML-теги и атрибуты, специфичные для структуры сайтов Billboard и Pitchfork. Такой подход позволяет отделить полезный текст от вспомогательных элементов страницы.

После извлечения текст очищается от лишних пробелов и объединяется в единое текстовое представление документа. Полученные документы сохраняются для дальнейшего использования в задачах информационного поиска, а также используются для подсчёта статистических характеристик корпуса.

Результатом работы программы является структурированный корпус текстовых документов, готовый для индексации, анализа и применения различных методов поиска и ранжирования.

3 Выводы

В ходе выполнения первой лабораторной работы по курсу «Информационный поиск» я познакомился с процессом формирования текстового корпуса из реальных веб-источников. На практике были изучены особенности работы с «сырыми» HTML-данными и сложности, возникающие при извлечении полезного текстового содержимого.

В процессе выполнения работы стало понятно, что даже структурированные сайты содержат большое количество служебной информации, не относящейся к основному тексту, и требуют аккуратной очистки и обработки. Также был получен опыт выделения мета-информации документов, такой как авторы и заголовки, которая может быть полезна при дальнейшем анализе и поиске.

Дополнительно была рассмотрена работа существующих поисковых систем и выявлено, что, несмотря на их удобство для конечных пользователей, они имеют ограничения с точки зрения исследовательских задач информационного поиска. Это подчёркивает необходимость создания собственных поисковых решений для проведения экспериментов и анализа.

Полученные навыки будут полезны при выполнении последующих лабораторных работ, а также в задачах анализа текстов, построения поисковых систем и обработки больших объёмов неструктурированных данных.

Список литературы

[1] Маннинг, Рагхаван, Шютце *Введение в информационный поиск* — Издательский дом «Вильямс», 2011. Перевод с английского: доктор физ.-мат. наук Д. А. Клю-

шина — 528 с. (ISBN 978-5-8459-1623-4 (рус.))