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: М. М. Сисенов
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.

Описание

Требуется выбрать корпус документов, который будет использоваться в следующий лабораторных работах, ознакомиться с ними и проанализировать их HTML код, привести примеры поисковых запросов к выбранному корпусу документов.

Источник данных

Были выбраны 2 сайта, главной тематикой которых являются статьи связанные с психологией:

- **b17.ru** `https://www.b17.ru/` — сайт с огромным количеством статей и возможностью общения с профессиональными психологами.
- **psychologies.ru** `https://psychologies.ru/` — сайт имеет более популярный и менее научный формат, акцентирующий внимание на актуальных новостях и трендах.

Описание корпуса документов

Причины выбора *b17.ru* и *psychologies.ru* :

- *Много текста*: На этих сайтах публикуются полноценные длинные статьи, а не короткие заметки. Это дает хороший объем данных, который необходим для качественной проверки закона Ципфа и работы стемминга.
- *Встроенный поиск*: Сайты имеют внутреннюю поисковую систему, что может облегчить сравнение с внешними поисковиками.
- *Простая верстка*: Структура сайтов понятна и логична (обычный HTML). Заголовки и тексты статей легко вытащить программно, не прибегая к сложным инструментам для обхода защиты или обработки скриптов.

Предварительный анализ структуры документов

Каждая статья на сайтах представляет собой отдельный HTML-документ. По предварительному анализу можно выделить общие структурные элементы:

- *Заголовок*: Обычно размещается в теге `<h1>`.
- *Основной текст*: Содержимое разбито на абзацы (`<p>`) и смысловые блоки. Часто используется микроразметка (например, атрибут `itemprop="articleBody"` или `class="article__block article__block_type-text"`).
- *Разметка*: Страницы используют современные семантические теги HTML5 (например, `<article>`, `<section>`), но также содержат большое количество служебных элементов (меню, реклама, ссылки), которые требуют фильтрации при парсинге.

Примеры документов

Пример документа с **b17.ru**:

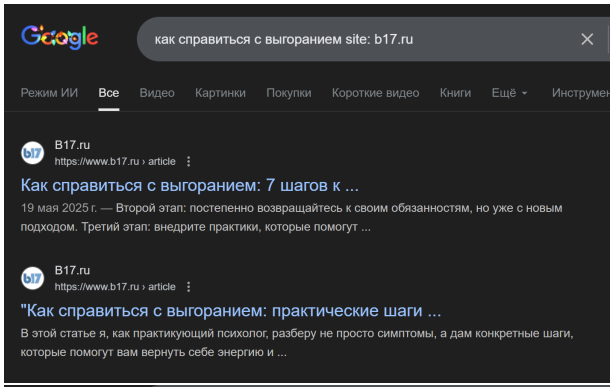
- *Размер сырого HTML*: 240 Kb
- *Извлеченый текст*: 22 Kb
- *Структура*: Документ имеет простую структуру: заголовок (<h1>), основной текст (itemprop="articleBody").

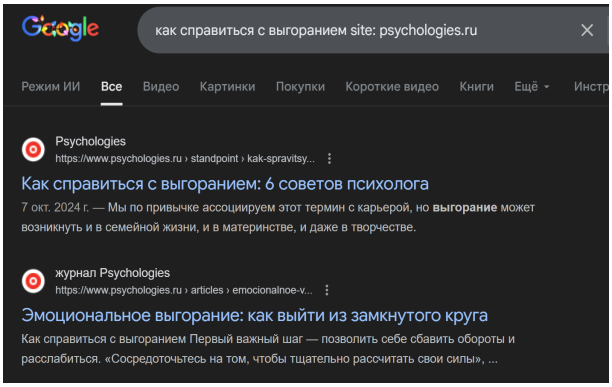
Средний результат документов с **psychologies.ru**:

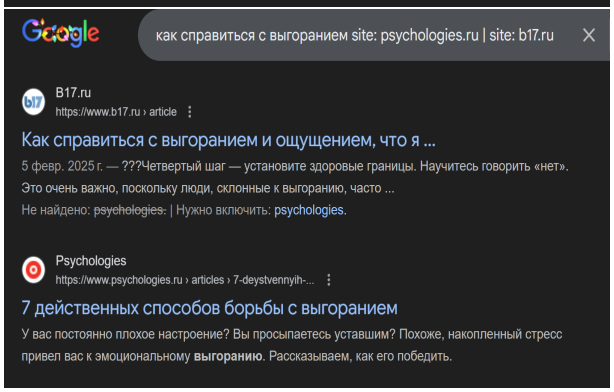
- *Размер сырого HTML*: 235 Kb
- *Извлеченый текст*: 14 Kb
- *Структура*: Структура документа сложнее чем у b17, потому что сайт предлагает авторам больше возможностей для оформления (квизы, цитаты, картинки). Это заставляет более тщательно продумывать логику парсинга.

Поисковые запросы и анализ выдачи

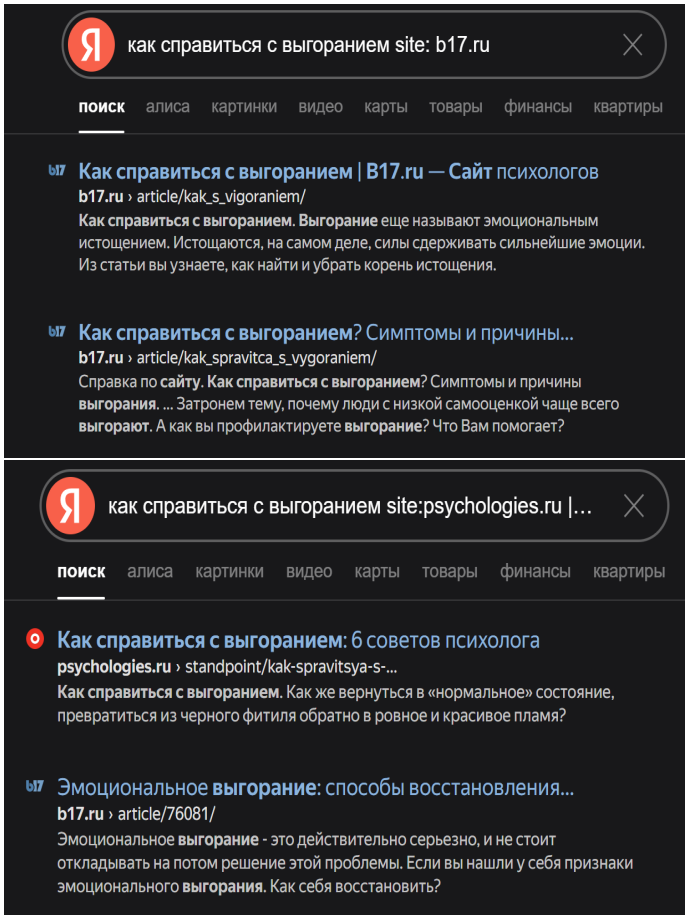
Для анализа были использованы Google и Яндекс. Чтобы задать конкретные ресурсы в поиске был использован оператор *site:*.
Google - запрос к сайту b17.ru, запрос к сайту psychologies.ru, запрос к обоим сайтам:







Аналогичные запросы с помощью Яндекса:



Оба поисковика выдали примерно те же результаты, которые бы с высокой вероятностью соответствовали ожиданиям пользователя.

Вывод

В ходе выполнения лабораторной работы был собран и проанализирован корпус документов на основе психологических порталов **b17.ru** и **psychologies.ru**. Была изучена структура HTML-страниц статей, выделены ключевые смысловые блоки (заголовок, основной текст). Определено, что для полноценного анализа необходимо не просто извлекать весь текст, а научиться выделять эти структурированные блоки отдельно.
Подготовленный корпус документов является релевантным, тематически однородным и достаточно объемным для выполнения последующих лабораторных работ по информационному поиску, таких как токенизация, стемминг, проверка закона Ципфа и построение булева поиска.

Литература

[1] Маннинг, Рагхаван, Шютце *Введение в информационный поиск* — Издательский дом «Вильямс», 2011. Перевод с английского: доктор физ.-мат. наук Д. А. Клюшина — 528 с. (ISBN 978-5-8459-1623-4 (рус.))

[2] b17.ru: Сайт психологов №1`https://b17.ru`

[3] psychologies.ru: Онлайн-журнал про психологию `https://psychologies.ru/`.