

**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: В. М. Филиппов  
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

Требуется выбрать корпус документов, который будет использоваться при дальнейшем выполнении лабораторных работ, проанализировать HTML-страницы, привести примеры поисковых запросов к выбранному корпусу документов

## 1 Корпус документов

В качестве корпуса документов для выполнения лабораторных работ были выбраны статьи из категории «Спорт» английской версии Википедии, а также материалы спортивной тематики с сайта The Guardian.

Выбор Википедии обусловлен простотой автоматизированной обработки (парсинга). Страницы ресурса были классифицированы на два типа: «статьи» и «категории». Алгоритм обработки следующий: если текущая страница идентифицирована как категория, то все ссылки с неё добавляются в очередь на обработку; если же это статья, то её HTML-код сохраняется в базу данных без добавления новых ссылок в очередь. Дополнительными преимуществами Википедии являются отсутствие агрессивных механизмов защиты от роботов (CAPTCHA, частые редиректы) и простая структура DOM-дерева, что облегчает работу библиотек по извлечению текста.

The Guardian — одно из старейших английских изданий с обширным архивом статей о спорте, в частности о футболе. Сайт газеты также отличается лояльностью к поисковым роботам: отсутствуют принудительные проверки CAPTCHA, блокировки (ошибки 403) и сложные цепочки перенаправлений.

## 2 Примеры документов

### 2.1 The Guardian

Я скачал HTML-страницу с сайта The Guardian с заголовком: "Semenyo a January target for Manchester United as well as Liverpool and City". В сыром виде она весит 295 КБ. После извлечения текста размер составил всего 14 КБ.

Документ представляет собой валидный HTML5 с явным указанием языка (`lang="en"`). Архитектура страницы указывает на использование **Server Side Rendering (SSR)** и компонентного подхода (вероятно, React), о чем свидетельствуют кастомные теги `<gu-island>` и хешированные имена классов.

Верстка выполнена с использованием семантических тегов HTML5:

- **Каркас:** `<header>`, `<main>`, `<article>`, `<footer>` и `<aside>`.
- **Текст:** Заголовок статьи обернут в `<h1>`. Основной текст разбит на параграфы `<p>`, однако все они имеют идентичный хешированный класс (например,

dcr-130mj7b).

- **Лид:** Вводная часть (*standfirst*) реализована через маркированный список `<ul>`.

## 2.2 Wikipedia

В качестве примера документа я выбрал статью посвящённую футбольному клубу Манчестер Юнайтед. В сыром виде документ весит 1.1 МБ, после извлечения текста 145 КБ.

Документ представляет собой валидный HTML5 с указанием языка (`lang="en"`). В отличие от предыдущего примера, данная страница сгенерирована движком MediaWiki (версия 1.46.0-wmf.7), что подтверждается мета-тегом `generator`.

Раздел `<head>` содержит стандартный набор мета-информации, характерный для информационных ресурсов:

- **Подключение ресурсов:** CSS и JavaScript загружаются через специальный обработчик `load.php` (ResourceLoader), который объединяет модули в один запрос для оптимизации.
- **Структурированные данные:** Присутствует блок `application/ld+json` (Schema.org), описывающий сущность `Article` и `Organization`, что помогает поисковым системам идентифицировать объект статьи.

Структура страницы классическая для MediaWiki и отличается высокой степенью вложенности контейнеров:

1. **Основной контейнер:** `<main id="content" class="mw-body">`. Это главный селектор для извлечения содержимого.
2. **Заголовок:** `<h1 id="firstHeading">`. Уникальный ID позволяет мгновенно найти название статьи.
3. **Информационная карточка (Infobox):** Ключевая особенность страниц Википедии. Реализована через таблицу:

```
<table class="infobox vcard">
```

Это критически важный элемент для парсинга фактологических данных (год основания, тренер, стадион).

4. **Текст:** Основной контент находится внутри `<div id="mw-content-text">`. Текст разбит на параграфы `<p>`, заголовки разделов `<h2>`, `<h3>` и списки `<ul>`.
5. **Таблицы данных:** Для отображения статистики и составов используется класс `wikitatable`, который имеет стандартизированные стили.

## 2 Примеры поисковых запросов

Ниже, на рисунке 1, приведены примеры поисковых запросов в Google с ограничением сайтов. Видим, что при запросе к обоим сайтам в топе выдачи находится только Википедия, первая статья на The Guardian только на 8 месте.

На рисунке 2 представлены запросы к Яндексу. В случае Википедии два первых результата одинаковы: заглавная статья клуба, а также статья, посвященная этому сезону. В случае The Guardian на первом месте в выдаче стоит титульная статья клуба, а далее последние новости, которые у Google поактуальнее. В случае общего запроса первые две страницы в выдаче одинаковые, а вот дальше Яндекс выдаёт нам уже статьи с новостного портала вместо Википедии.

При запросе по Википедии в Яндексе в топ выдачи попали странные страницы «2024 FA Cup final» и «2023 EFL Cup final», которые никак не связаны с клубом Манчестер Юнайтед. У Google получилась более релевантная выдача. При запросе с theguardian оба поисковика в целом выдали, что требуется: последние новости, расписание матчей, однако у Google выдача получась более актуальной по времени. С задачей поиска по двум сайтам Яндекс справился лучше, на мой взгляд: можно найти как новости с портала, так и титульную статью на Википедии. У Google выдача получилась однобокой.

В настоящее время поисковики используют ИИ модели для разбора запроса. Также видно, что они стараются актуализировать выдачу, чтобы дать самые новые новости пользователю.

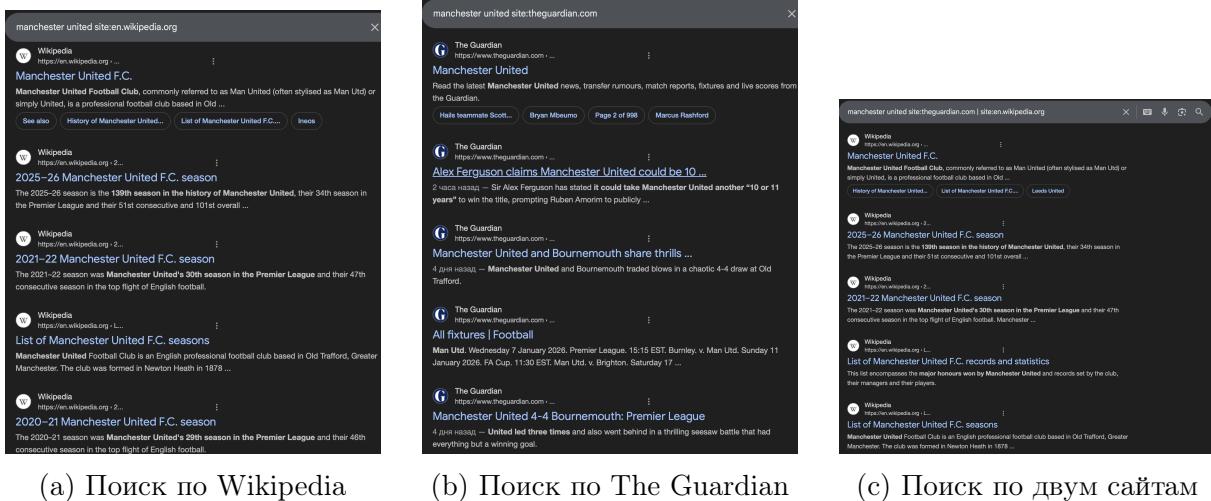


Рис. 1: Примеры поисковых запросов к Google

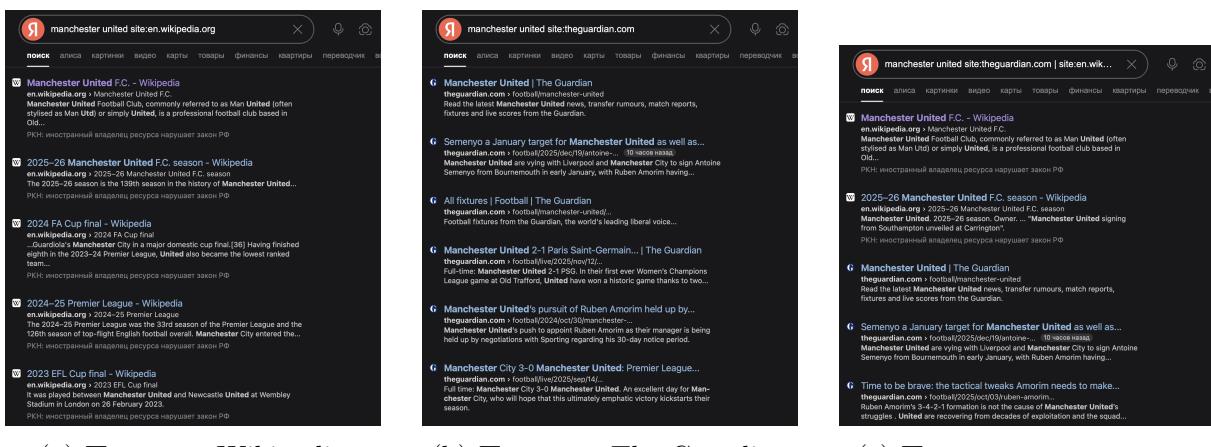


Рис. 2: Примеры поисковых запросов к Яндексу

### **3 Выводы**

При выполнении первой лабораторной работы по информационному поиску, я больше узнал про структуру HTML-документов, поскольку раньше с ними никогда не работал. Узнал про href блоки, селекторы и прочее. Познакомился с тем, каким образом поисковые движки выдают результаты запроса, а также больше узнал про структуру самого запроса.

## Список литературы

- [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] Селектор — что такое // Skyeng. URL: <https://skyeng.ru/magazine/wiki/it-industriya/chto-takoe-selektor/> (дата обращения: 11.12.2025).