|  |
| --- |
|  |
| МИНОБРНАУКИ РОССИИ |
| Федеральное государственное бюджетное образовательное учреждение  высшего образования  **«МИРЭА − Российский технологический университет»**  **РТУ МИРЭА** |
| Институт кибербезопасности и цифровых технологий |
| Кафедра КБ-14 «Цифровые технологии обработки данных» |

**ОТЧЕТ**

**по практической работе**

**«№1. АНАЛИЗ ИСХОДНОГО КОДА ВЕБ-РЕСУРСОВ И ФОРМИРОВАНИЕ CSS-СЕЛЕКТОВРОВ»**

**по дисциплине «Обработка данных с веб-ресурсов и поисковых машин»**

|  |  |  |  |
| --- | --- | --- | --- |
| Выполнил |  |  | Парамонов П.К.  *фамилия, имя, отчество* |
| шифр | 22Б0723 | группа | БСБО-05-22 |
|  |  |  |  |
| Проверил |  | к.т.н., доцент  *ученая степень, должность* | Ильин Д.Ю.  *фамилия, имя, отчество* |

**Москва 2024г.**

## Цель практической работы

Цель настоящей практической работы — научиться выполнять анализ исходных данных веб-страниц и осуществлять подготовку CSS-селекторов и запросов для извлечения данных из веб-источников.

## Задачи практической работы

Для достижения поставленной цели необходимо решить следующие задачи:

1. Определить вариант задания.
2. Найти подходящие источники данных
3. Рассмотреть исходный код веб-страниц и доступные способы извлечения данных.
4. Составить таблицу со списком источников и перечнем данных к извлечению.
5. Сформировать набор CSS-селекторов или запросов к API для извлечения данных.
6. Продемонстрировать, что представленные селекторы и запросы соответствуют задаче.
7. Подготовить ответы на контрольные вопросы.
8. Составить отчет о проведенной работе.

## Вариант задания

Вариант выполнения задания 91

## Ход выполнения практической работы

В ходе выполнения практической работы будут рассмотрены 5 сайтов с объявлениями квартир:

1. Домклик
2. Циан
3. Авито
4. Яндекс.Недвижимость
5. Юла

Обращаться к данным на всех сайтах, изложенных выше, будем с помощью селекторов.

Далее занесем всю информацию о сайтах и селекторах в соответствующие таблицы.

Таблица 1 – Селекторы для Домклик

|  |  |
| --- | --- |
| Цена | $$('div[data-e2e-id="product-snippet-price-sale"]>p') |
| Параметры | $$('a[data-test="product-snippet-property-offer"]') |
| Адрес | $$('span[data-e2-id="product-snippet-address"') |
| Информация для автомобилей(шоссе/км от мкада | $$('div[data-e2e-id="car-icon"]~div') |
| $$('div[data-e2e-id="car-icon"]~div~span') |

Таблица 2 – Селекторы для Циан

|  |  |
| --- | --- |
| Цена | $$('span[class\*=-popular-price-]') |
| Описание | $$('a[class\*=”-c-popular-tec-info-“]') |
| Метро рядом | $$('span[class\*=-metro-name-]') |
| Время до метро | $$(‘span[class\*=-metro-time-]') |

Таблица 3 – Селекторы для Авито

|  |  |
| --- | --- |
| Описание | $$('div[class^="iva-item-titleStep-"') |
| Цена | $$('span[class\*="price-root-"]') |
| Геолокация | $$('div[class^="geo-root-"') |
| Подробное описание | $$('div[class\*="item-description"]>p') |
| Дата публикации | $$('div[class^="iva-item-dateInfoStep-"]>div>p') |

Таблица 4 – Селекторы для Яндекс.Недвижимость

|  |  |
| --- | --- |
| Название комплекса | $$('h3[class="SiteSnippetSearch\_\_heading"]') |
| Адрес | $$('div[class\*="AddressWithGeoLinks"]') |
| Таблица расценок | $$('div[class="SnippetPriceTable SiteSnippetSearch\_\_table"]') |
| Подробное описание | $$('div[class\*="SiteSnippetSearch\_\_description\_fixed"]') |
| Застройщик | $$('div[class="SiteSnippetSearch\_\_developerBlock"]>a') |

Таблица 5 – Селекторы для Юлы

|  |  |
| --- | --- |
| Город | $$('div[data-test-component="Badges"]>div>div>span') |
| Цена | $$('p[data-test-block="ProductPrice"]') |
| Описание | $$('span[data-test-block="ProductName"]') |

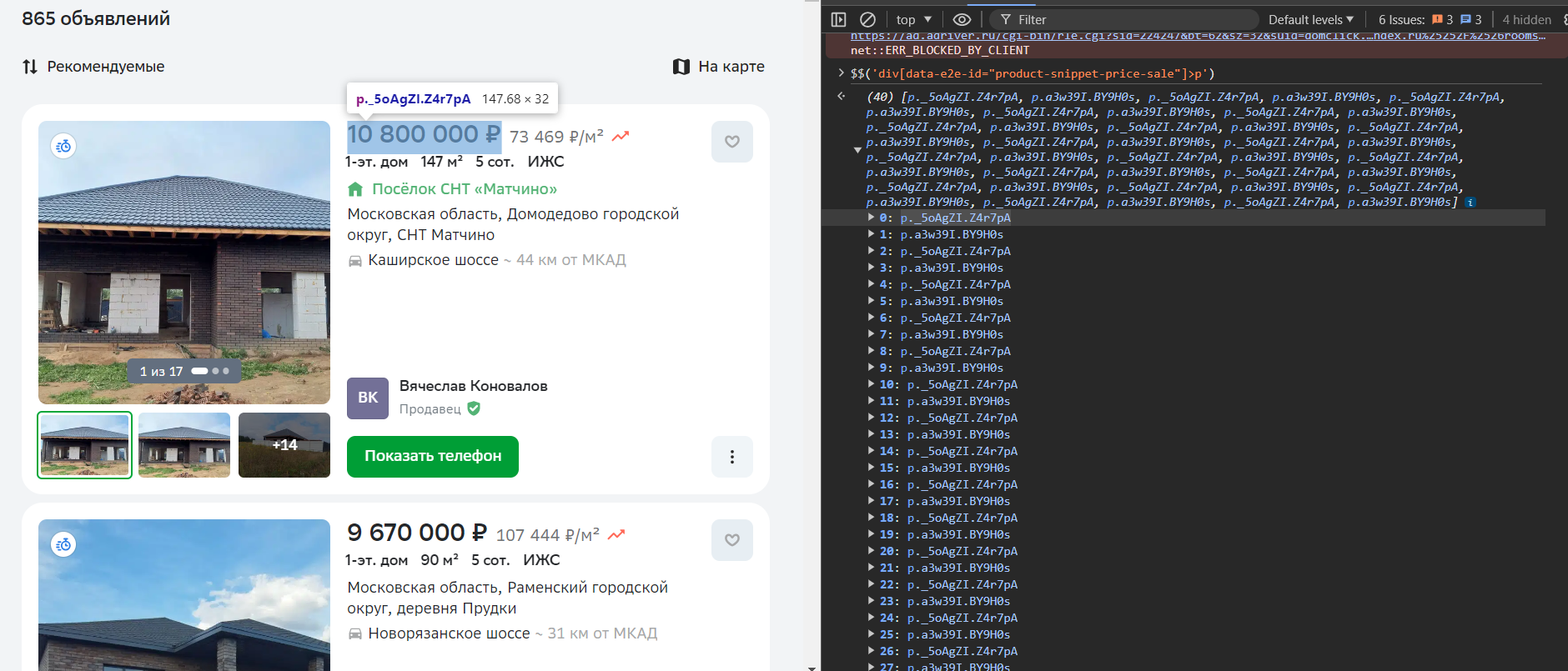


Рис 1 – Пример работы селектора для Домклик

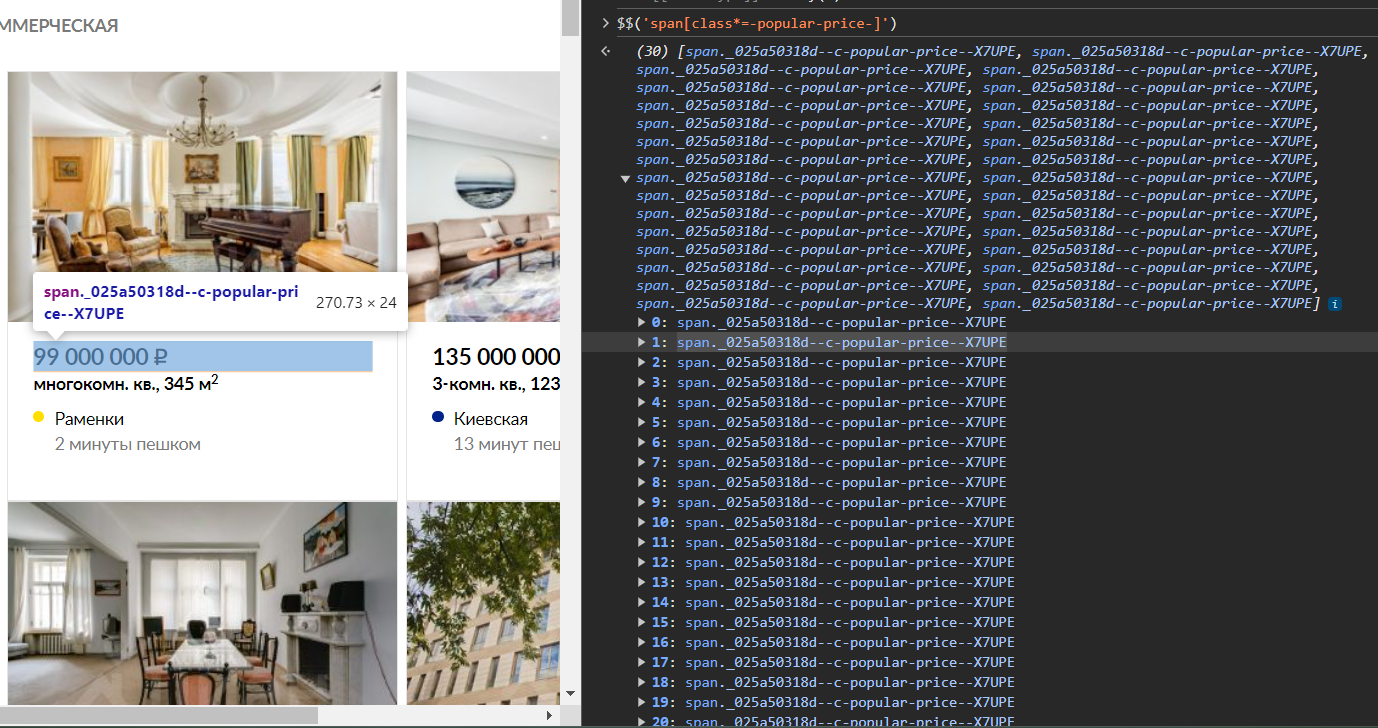


Рис 2 – Пример работы селектора для Циан

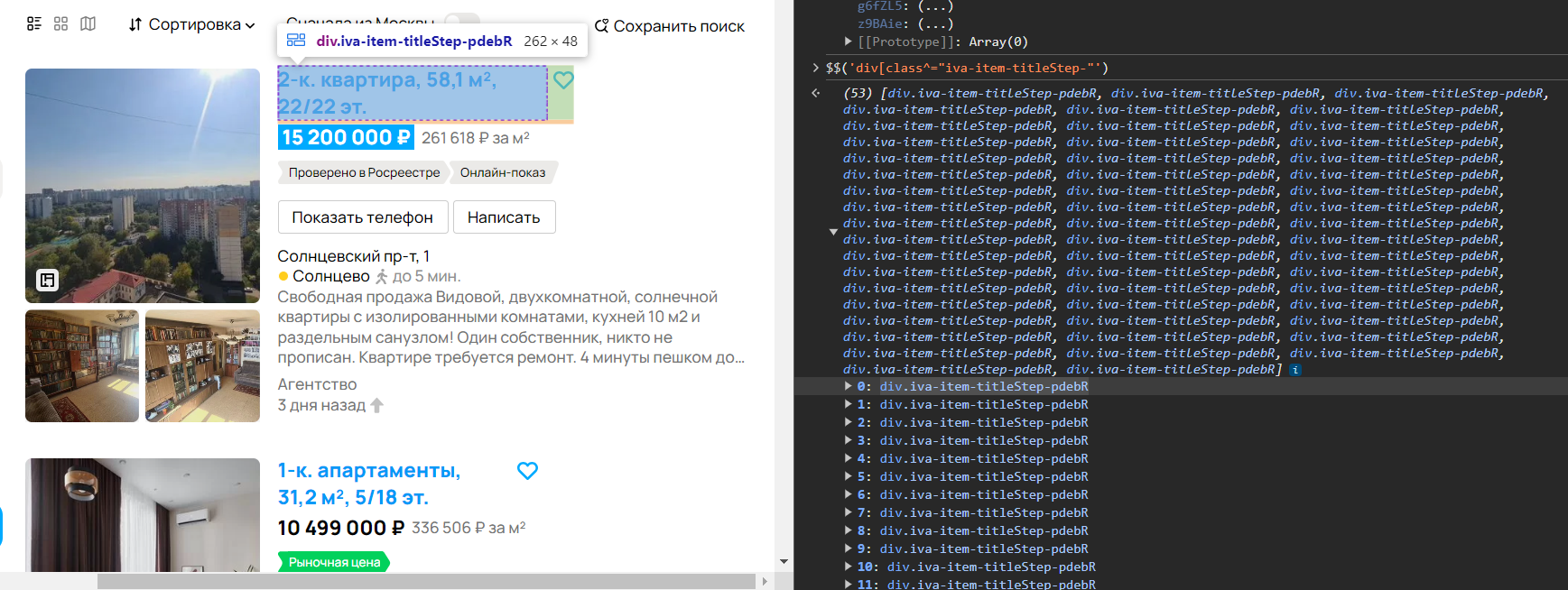


Рис 3 – Пример работы селектора для Авито

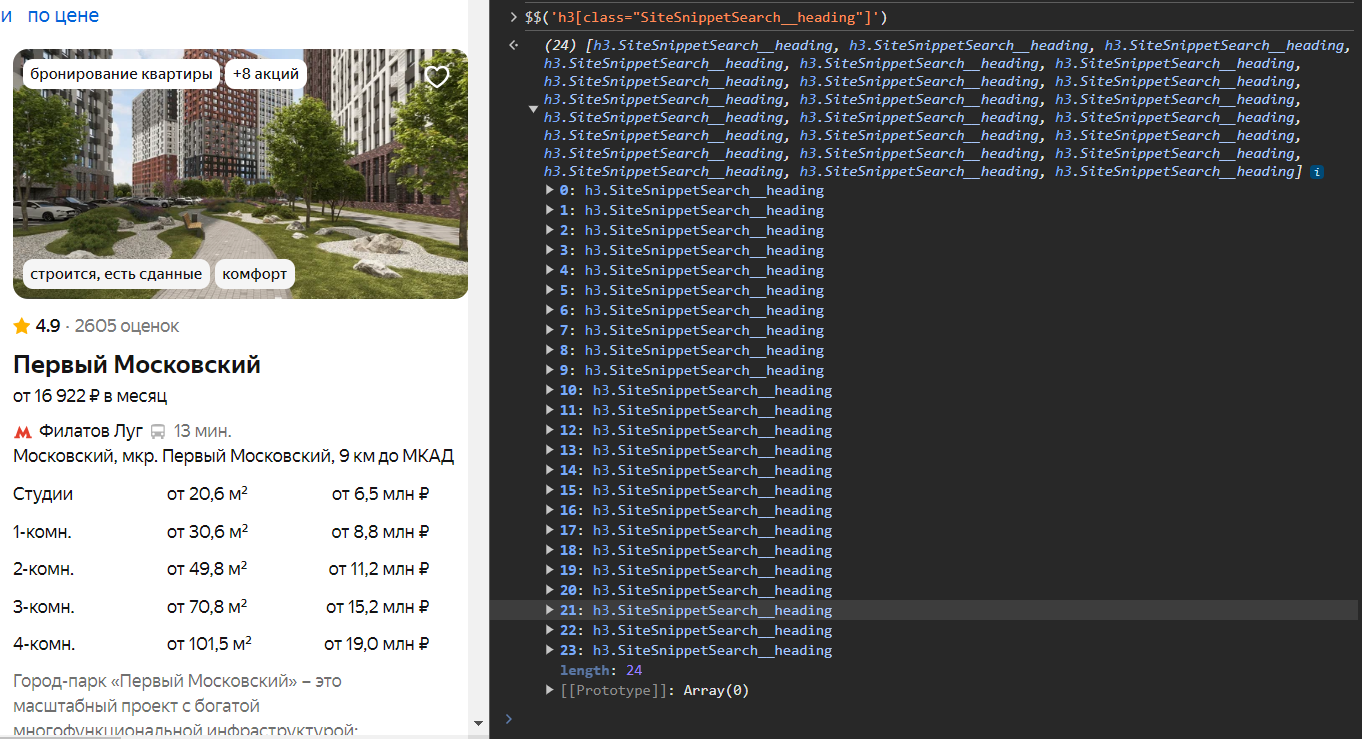


Рис 4 – Пример работы селектора для Яндекс.Недвижимость

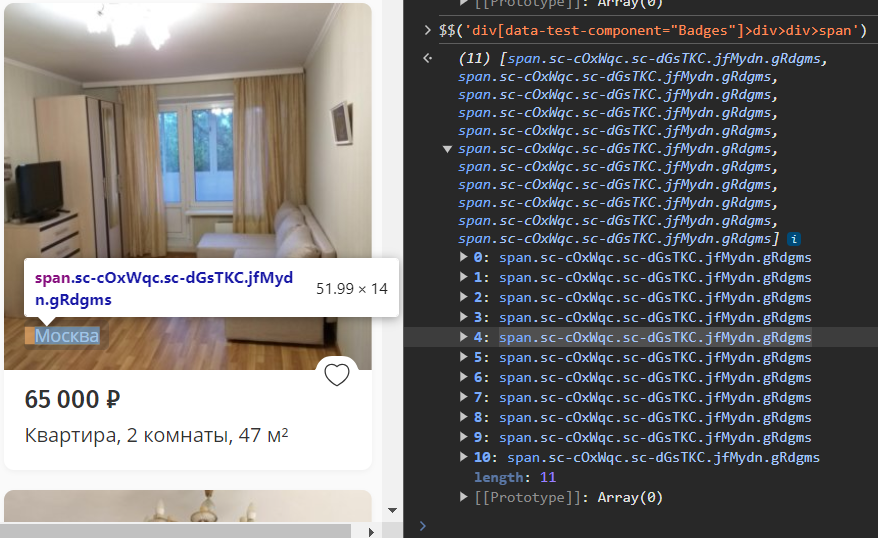


Рис 5 – Пример работы селектора для Юла

## Трудности в ходе выполнения практической работы

Без уточнения было непонятно что в задании нужно сделать. При выполнении основной трудностью было отбирать сайты с удобной для написания селекторов вёрсткой. В остальном проблем не было.

## Ответы на контрольные вопросы

1. Сбор данных может применяться для различных целей. К примеру, анализ данных, т.е. изучение тенденций и закономерностей. Мониторинг: отслеживание изменений в реальном мире, например, цены на товар.
2. API предоставляет структурированные данные напрямую с сервера, данные обычно возвращаются в формате JSON или XML. API более надежен, так как изменяется реже и предоставляет доступ к данным без необходимости парсинга HTML.

Данные с веб-страниц необходимо извлекать из HTML-разметки с использованием методов парсинга. Изменение структуры сайта может нарушить работу парсера. Веб-страницы не всегда структурированы, и доступ к данным может быть более ограниченным или неполным.

1. Наиболее распространенные форматы данных при работе с веб-ресурсами: HTML, JSON, XML, CSV
2. Для того, чтобы CSS-селектор не стал неактуальным после обновления веб-ресурса, стоит использовать уникальные и устойчивые селекторы, такие как ID элементов. Использовать атрибуты data-, поскольку они могут быть более стабильными
3. Существует 5 видов кодов ответа:

1xx: Informational (информационные):

* 100 Continue («продолжайте»);
* 101 Switching Protocols («переключение протоколов»);
* 102 Processing («идёт обработка»);
* 103 Early Hints («предварительный ответ»);

2xx: Success (успешно):

* 200 OK («хорошо»);
* 201 Created («создано»);
* 202 Accepted («принято»);
* 203 Non-Authoritative Information («информация не авторитетна»);
* 204 No Content («нет содержимого»);
* 205 Reset Content («сбросить содержимое»);
* 206 Partial Content («частичное содержимое»);
* 207 Multi-Status («многостатусный»);
* 208 Already Reported («уже сообщалось»);
* 226 IM Used («использовано IM»).

3xx: Redirection (перенаправление):

* 300 Multiple Choices («множество выборов»);
* 301 Moved Permanently («перемещено навсегда»);
* 302 Found («найдено»);
* 303 See Other («смотреть другое»);
* 304 Not Modified («не изменялось»);
* 305 Use Proxy («использовать прокси»);
* 306 — зарезервировано (код использовался только в ранних спецификациях);
* 307 Temporary Redirect («временное перенаправление»);
* 308 Permanent Redirect («постоянное перенаправление»).

4xx: Client Error (ошибка клиента):

* 400 Bad Request («неправильный, некорректный запрос»);
* 401 Unauthorized («не авторизован»);
* 402 Payment Required («необходима оплата») — зарезервировано для использования в будущем;
* 403 Forbidden («запрещено (не уполномочен)»);
* 404 Not Found («не найдено»);
* 405 Method Not Allowed («метод не поддерживается»);
* 406 Not Acceptable («неприемлемо»);
* 407 Proxy Authentication Required («необходима аутентификация прокси»);
* 408 Request Timeout («истекло время ожидания»);
* 409 Conflict («конфликт»);
* 410 Gone («удалён»);
* 411 Length Required («необходима длина»);
* 412 Precondition Failed («условие ложно»);
* 413 Payload Too Large («полезная нагрузка слишком велика»);
* 414 URI Too Long («URI слишком длинный»);
* 415 Unsupported Media Type («неподдерживаемый тип данных»);
* 416 Range Not Satisfiable («диапазон не достижим»);
* 417 Expectation Failed («ожидание не оправдалось»);
* 418 I’m a teapot («я — чайник»);
* 419 Authentication Timeout (not in RFC 2616) («обычно ошибка проверки CSRF»);
* 421 Misdirected Request;
* 422 Unprocessable Entity («необрабатываемый экземпляр»);
* 423 Locked («заблокировано»);
* 424 Failed Dependency («невыполненная зависимость»);
* 425 Too Early («слишком рано»);
* 426 Upgrade Required («необходимо обновление»);
* 428 Precondition Required («необходимо предусловие»);
* 429 Too Many Requests («слишком много запросов»);
* 431 Request Header Fields Too Large («поля заголовка запроса слишком большие»);
* 449 Retry With («повторить с»);
* 451 Unavailable For Legal Reasons («недоступно по юридическим причинам»).
* 499 Client Closed Request (клиент закрыл соединение);

5xx: Server Error (ошибка сервера):

* 500 Internal Server Error («внутренняя ошибка сервера»);
* 501 Not Implemented («не реализовано»);
* 502 Bad Gateway («плохой, ошибочный шлюз»);
* 503 Service Unavailable («сервис недоступен»);
* 504 Gateway Timeout («шлюз не отвечает»);
* 505 HTTP Version Not Supported («версия HTTP не поддерживается»);
* 506 Variant Also Negotiates («вариант тоже проводит согласование»);
* 507 Insufficient Storage («переполнение хранилища»);
* 508 Loop Detected («обнаружено бесконечное перенаправление»);
* 509 Bandwidth Limit Exceeded («исчерпана пропускная ширина канала»);
* 510 Not Extended («не расширено»);
* 511 Network Authentication Required («требуется сетевая аутентификация»);
* 520 Unknown Error («неизвестная ошибка»);
* 521 Web Server Is Down («веб-сервер не работает»);
* 522 Connection Timed Out («соединение не отвечает»);
* 523 Origin Is Unreachable («источник недоступен»);
* 524 A Timeout Occurred («время ожидания истекло»);
* 525 SSL Handshake Failed («квитирование SSL не удалось»);
* 526 Invalid SSL Certificate («недействительный сертификат SSL»).

## Вывод

В ходе выполнения работы "Анализ исходного кода веб-ресурсов и формирование CSS-селекторов" были рассмотрены аспекты анализа HTML-структуры веб-страниц и методы построения устойчивых CSS-селекторов. Также в ходе практической работы было усвоено, что правильное использование селекторов играет важную роль при сборе данных с веб-ресурсов.