Министерство образования и науки Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ

ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

**Анализ методов парсинга сайтов**

Дипломная РАБОТА

студента 4 курса 441 группы

направления 02.03.03 математическое обеспечение и администрирование информационных систем факультета КНиИТ

Турченков Павел Александрович

Научный руководитель

Старший преподаватель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Портенко М.С.

подпись, дата

Зав. кафедрой

Доцент, кандидат физико-математических наук \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Огнева М.В. подпись, дата

Саратов 2018

**Содержание**

[Введение 4](#_Toc9042894)

[Описание актуальности работы 5](#_Toc9042895)

[Отбор библиотек для парсинга 6](#_Toc9042896)

[Реализация на языке Java 7](#_Toc9042897)

[**Jsoup** 7](#_Toc9042898)

[**Selenium** 10](#_Toc9042899)

[**Unirest** 13](#_Toc9042900)

[Реализация на языке C# 21](#_Toc9042901)

[**HtmlAgilityPack** 21](#_Toc9042902)

[**Fizzler** 22](#_Toc9042903)

[**AngleSharp** 23](#_Toc9042904)

[Реализация на языке Python 26](#_Toc9042905)

[Сравнение скорости реализаций 27](#_Toc9042906)

[Описание визуализации результатов 28](#_Toc9042907)

[Описание выхода из специфических ситуаций при парсинге 29](#_Toc9042908)

[Заключение 30](#_Toc9042909)

[Список использованной литературы 31](#_Toc9042910)

[Приложение А. 32](#_Toc9042911)

[Приложение Б. 33](#_Toc9042912)

[Приложение В. 34](#_Toc9042913)

# Введение

# Описание актуальности работы

# Отбор библиотек для парсинга

# Реализация на языке Java

Во время практики были реализованы следующие способы парсинга контента сайтов на языке Java: библиотека Jsoup, Selenium и Unirest.

### **Jsoup**

Jsoup – это библиотека для анализа документа HTML. Jsoup предоставляет API для получения и манипулирования данными взятых либо с помощью URL, либо из HTML. У этой библиотеки есть методы похожие на DOM, CSS и JQuery для того, чтобы получать и обрабатывать данные. (<https://o7planning.org/ru/10399/jsoup-java-html-parser-tutorial>)

Для парсинга сайта с помощью библиотеки Jsoup были использованы следующие методы:

(<https://jsoup.org/apidocs/org/jsoup/Connection.html>) (<https://jsoup.org/>)

* connect(String url) – В качестве входного аргумента требует строку в виде URL. Создает новый объект типа Connection для заданного URL, и используется для получения и анализа HTML-страницы;
* get() – выполняет GET запрос и анализирует результат. Возвращает объект Document;
* post() – выполняет POST запрос и анализирует результат. Возвращает объект Document;
* userAgent​(String userAgent) – В качестве входного аргумента требует строку в виде агента пользователя, который выглядит следующим образом: «Mozilla/5.0 (Windows NT 6.2; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.1667.0 Safari/537.36». Данный метод устанавливает заголовок запроса агента пользователя;
* parse(String html) – Анализирует код HTML и возвращает объект Document;
* getElementsByClass​(String className) – В качестве входного аргумента требует строку в виде наименования класса. Находит элементы, которые имеют этот данный класс и возвращает их. Тип возвращаемого значения Elements;
* getElementsByTag​(String tagName) – В качестве входного аргумента требует строку в виде наименования html тэга. Находит элементы, которые имеют этот данный тэг и возвращает их. Тип возвращаемого значения Elements;
* attr(String key) – В качестве входного аргумента требует строку в виде наименования атрибута тэга. Возвращает содержимое атрибута;
* text() – Возвращает текст внутри основного тэга и дочерних элементов;
* select(String cssQuery) – В качестве входного аргумента требует строку в виде селектора. Возвращает элементы, которые соответствуют селектору.
* selectFirst(String cssQuery) – В качестве входного аргумента требует строку в виде селектора. Возвращает первое соответствие заданному селектору. Аналогичный результат будет при select(String cssQuery).first(), но этот вариант менее эффективный, так как первый метод будет искать все совпадения, что будет затратно по времени, и только потом второй метод вернёт первый элемент.

(<https://jsoup.org/apidocs/org/jsoup/nodes/Element.html#select-java.lang.String->)

При построении логики парсинга сайта использовалась панель разработчика в браузере Google Chrome, чтобы не писать селекторы или XPath самостоятельно. Использование возможностей панели разработчика в браузере Google Chrome изображено на Рисунке 1.

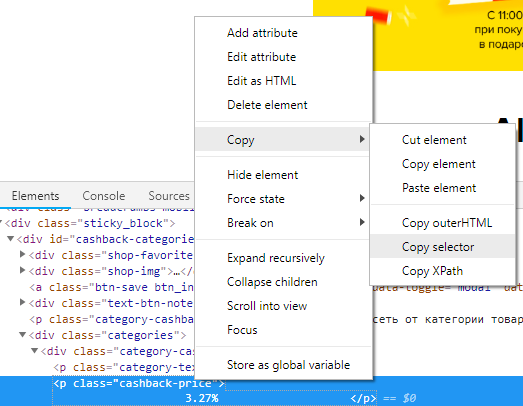


Рисунок 1 – использование возможностей панели разработчика в Google Chrome

Плюсы библиотеки Jsoup:

* является интуитивно понятной библиотекой, которой легко пользоваться;
* содержит множество методов, которые облегчают парсинг сайтов;
* присутствует имитация пользователя через userAgent;
* имеется возможность взаимодействия с cookies.

Минусы библиотеки Jsoup:

* с помощью библиотеки Jsoup можно парсить только статические страницы, то есть нельзя работать с динамически подключаемым контентом;
* Jsoup не поддерживает XPath. Но селекторы являются альтернативой;
* среди методов, которые возвращают элементы по классу, тэгам и так далее нет методов для возвращения только первого элемента. Всегда происходит полный обход html. Приходится пользоваться методов first().

### **Selenium**

Selenium – это набор различных программных инструментов, каждый из которых имеет свой подход к поддержке автоматизации тестирования. Весь набор инструментов обеспечивает широкий набор функций тестирования, специально предназначенных для тестирования веб-приложений всех типов. Эти операции очень гибкие, что позволяет использовать множество опций для определения местоположения элементов пользовательского интерфейса. Одной из ключевых функций Selenium является поддержка выполнения тестов на нескольких платформах браузера.

В случае с парсингом сайтов Selenium предоставляет возможность полностью имитировать поведение пользователя на сайте, то есть переход по ссылкам, использование активных объектов, заполнение некоторых полей и так далее.

Для работы с Selenium нужно скачать вебдрайвер и подключить его через setProperty(String key, String value), где key – имя драйвера, то есть «webdriver.chrome.driver», а value – путь к самому драйверу. В данном проекте использовался chromedriver.exe. Также нужно подключить библиотеку selenium для взаимодействия с браузером.

Для парсинга сайта с помощью Selenium были использованы следующие методы:

* get(String URL) – В качестве входного аргумента требует строку в виде URL. С помощью этого метода происходит навигация по URL;
* navigate().to(String URL) - В качестве входного аргумента требует строку в виде URL. С помощью этого метода происходит навигация по URL. Разница этого метода и метода get(String URL) небольшая, но она есть. Используя navigate(), можно использовать такие методы, как back(), forward(), refresh();

(<https://www.oodlestechnologies.com/blogs/How-To-Navigate-A-URL-Using-Selenium-Web-Driver/>)

* findElement(By by) – В качестве параметра принимает объект типа By, который задает критерий поиска. Возвращает элемент типа WebElement, который соответствует заданному критерию;
* findElements(By by) - В качестве параметра принимает объект типа By, который задает критерий поиска. Возвращает список всех элементов типа WebElement, которые соответствуют заданному критерию;
* By – представляет из себя локатор, который уникально идентифицирует элемент страницы. Selenium предоставляет несколько способов использования локаторов для поиска элементов:

1. By.id – в качестве локатора используется атрибут id элемента страницы;
2. By.name – в качестве локатора используется атрибут name элемента страницы;
3. By.xpath – используется для поиска элемента по XPath выражению;
4. By.tagName – поиск по имени HTML тега;
5. By.className – поиск по классу элемента;
6. By.cssSelector – используется для поиска элемента по селектору;
7. By.linkText – поиск ссылки с указанным текстом. Текст ссылки должен быть точным совпадением;
8. By.partionalLinkText – поиск ссылки по части с указанным текстом.

* getAttribute(String s) –В качестве параметра принимает строку в виде наименования атрибута. Возвращает текстовое представление найденного атрибута.

Помимо методов для поиска элементов используются такие методы, как click(), который позволяет взаимодействовать с активными объектами такими как кнопки и ссылки. Также метод isDisplayed (), который проверяет отображение элемента.

(<https://www.guru99.com/find-element-selenium.html>)

(<http://internetka.in.ua/selenium-webdriver-findelement-by/>)

(<https://www.seleniumhq.org/docs/01_introducing_selenium.jsp>)

При построении логики парсинга сайта использовалась панель разработчика в браузере Google Chrome, чтобы не писать селекторы или XPath самостоятельно, как в примере с Jsoup. Использование возможностей панели разработчика в браузере Google Chrome изображено на Рисунке 1.

Плюсы Selenium:

* имитация пользователя;
* методы для поиска как одного элемента, так и множества;
* поиск может осуществляться множествами локаторами.

Минусы Selenium:

* использует реальный браузер для работы, из-за чего скорость работы падает.

### **Unirest**

Unirest - это набор облегченных HTTP-библиотек, доступных на нескольких языках, созданный и поддерживаемый Mashape, который также поддерживает API-интерфейс с открытым исходным кодом Gateway Kong.

(<http://unirest.io/java.html>)

Заголовки и параметры передаются с помощью API header() и fields(). Запрос вызывается при вызове метода asJson(). Кроме этого метода есть другие варианты, например asBinary(), asString() и asObject().

Чтобы передать несколько заголовков или полей, можно либо создать Map и передать их в .headers (<String, Object>) и .fields (<String, String>) соответственно, либо постепенно добавлять через .header(<String, Object>) и .field(<String, Object>) поле за полем.

Следующий код показывает, как с помощью запросов программа получает json с сайта [Kopikot.ru](https://www.kopikot.ru/), который в последствии обрабатывается.

Пример с использование Map:

1. HttpResponse<String> response = null;
2. Map<String, String> header = new HashMap<>();
3. header.put("User-Agent", "PostmanRuntime/7.11.0");
4. header.put("Accept", "\*/\*");
5. header.put("Cache-Control", "no-cache");
6. header.put("Postman-Token", "ac6a192d-72f5-4238-9321-55c1308e9846,09ec761f-7c53-4696-bf5a-dc44918e73c5");
7. header.put("Host", "d289b99uqa0t82.cloudfront.net");
8. header.put("accept-encoding", "gzip, deflate");
9. header.put("Connection", "keep-alive");
10. header.put("cache-control", "no-cache");
11. for (int i = 0; i <= 2000; i += 100) {
12. String url = "https://d289b99uqa0t82.cloudfront.net/sites/5/campaigns\_limit\_100\_offset\_" + i + "\_order\_popularity.json";
13. try {
14. response = Unirest.get(url)
15. .headers(header)
16. .asString();
17. } catch (UnirestException e) {
18. log.error(e + " : " + i);
19. break;
20. }

Пример без использования Map:

1. HttpResponse<String> response = null;
2. for (int i = 0; i <= 2000; i += 100) {
3. String url = "https://d289b99uqa0t82.cloudfront.net/sites/5/campaigns\_limit\_100\_offset\_" + i + "\_order\_popularity.json";
4. try {
5. response = Unirest.get(url)
6. .header("User-Agent", "PostmanRuntime/7.11.0")
7. .header("Accept", "\*/\*")
8. .header("Cache-Control", "no-cache")
9. .header("Postman-Token", "ac6a192d-72f5-4238-9321-55c1308e9846,09ec761f-7c53-4696-bf5a-dc44918e73c5")
10. .header("Host", "d289b99uqa0t82.cloudfront.net")
11. .header("accept-encoding", "gzip, deflate")
12. .header("Connection", "keep-alive")
13. .header("cache-control", "no-cache")
14. .asString();
15. } catch (UnirestException e) {
16. log.error(e + " : " + i);
17. break;
18. }
19. HttpResponse<String> finalResponse = response;
20. if (finalResponse != null) {
21. futures.add(pool.submit(() -> parseElements(finalResponse.getBody())));
22. }
23. }

(<https://www.codeflow.site/ru/article/unirest>)

Для упрощения работы с запросами использовалось приложение Postman. Используя это приложения, можно отправлять запросы на сайты с нужными заголовками и полями. В примере выше можно заметить такие заголовки, как User-Agent, Cache-Control и прочие. Некоторые заголовки Postman ставит автоматически.

Как можно заметить из примера сверху, URL, который передаётся в запрос отличается от URL самого сайта. Такие различия объясняются тем, у запрос должен передаваться не по адресу самого сайта, а другому, который будет указан в панели разработчика. При отправке запроса на адрес самого магазина в ответе будет не html код страницы, а javaScript страницы.

Чтобы узнать URL, по которому передаются данные, нужно воспользоваться панелью разработчика и найти XHR – XMLHttpRequest, который отвечает за добавление данных на страницу. В случае с сайтом [Kopikot.ru](https://www.kopikot.ru/), запрос будет отправлен при нажатии кнопки «Больше». После этого в панели разработчика в разделе Network будет нужный XHR. Результат изображён на Рисунке 2.

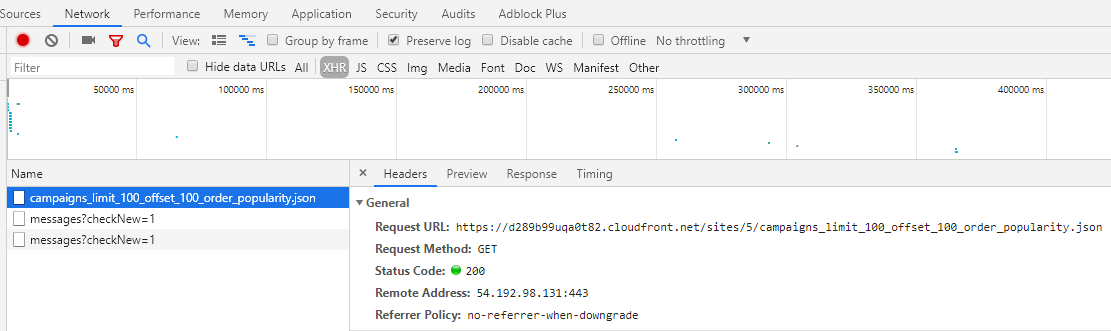


Рисунок 2 – XHR для 100 магазинов с сайта [Kopikot.ru](https://www.kopikot.ru/)

Как видно на Рисунке 2, в разделе Headers указан Request URL. Так как при отображении магазинов было выбрано показывать по 100 магазинов, в URL видно это число: «offset\_100». При нажатии кнопки ещё раз в Network появляется ещё один XHR, но уже с числом 200. Кроме самого URL важно какой метод использует запрос: GET или POST. Посмотреть это можно в «Request Method» Результат виден на Рисунке 3.

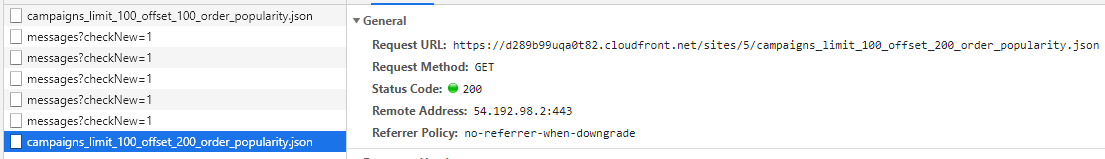


Рисунок 3 – XHR для 200 магазинов с сайта Kopikot.ru

Практически было выяснено, что таких XHR всего 14 штук: от 0 до 13.

Используя программу Postman, через этот URL отправляется запрос. В качестве ответа в разделе Body появляется JSON файл с нужной информацией. Результат запроса на Рисунок 4.

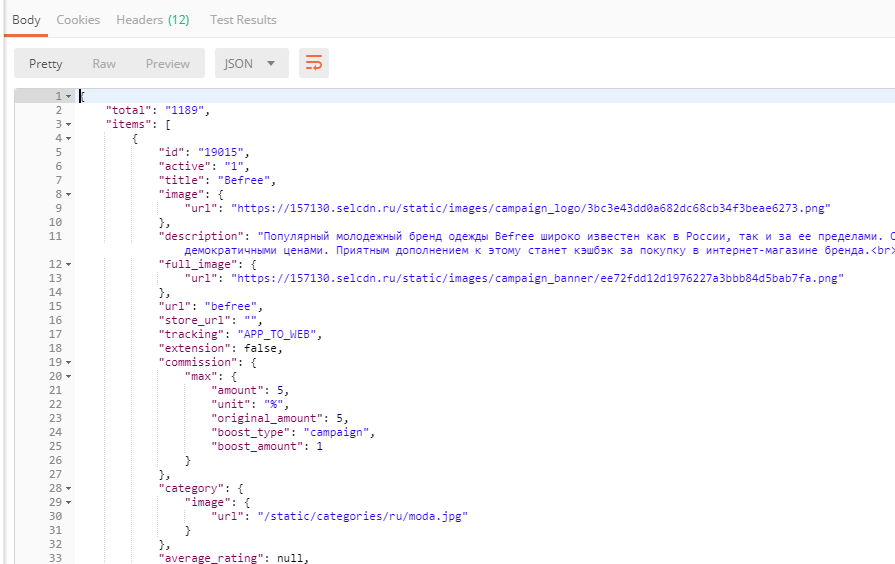


Рисунок 4 – результат запроса к URL через Postman

При просмотре какие запросы происходят на сервере важно обращать внимание на ту информацию, которая показана в Headers. Пример изображён на Рисунок 5.

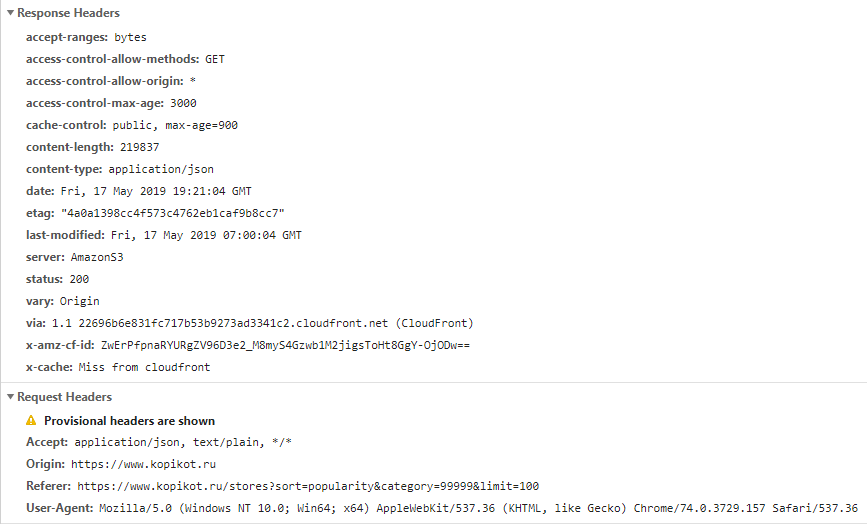


Рисунок 5 – Headers сайта Kopikot.ru

На этом рисунке можно увидеть, например, в каком формате будет ответ на запрос – поле content-type.

Также имеются случаи, когда при запросе передаются некоторые параметры. Для примера приведён Headers сайта [Cash4Brands.ru](https://cash4brands.ru/) на Рисунок 6 и Рисунок 7.

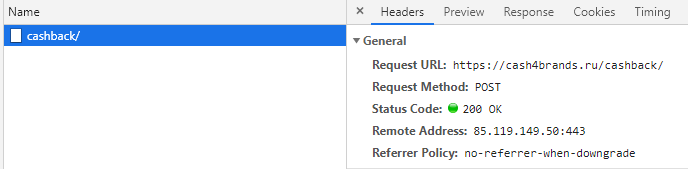


Рисунок 6 – Headers сайта Cash4Brands.ru часть 1

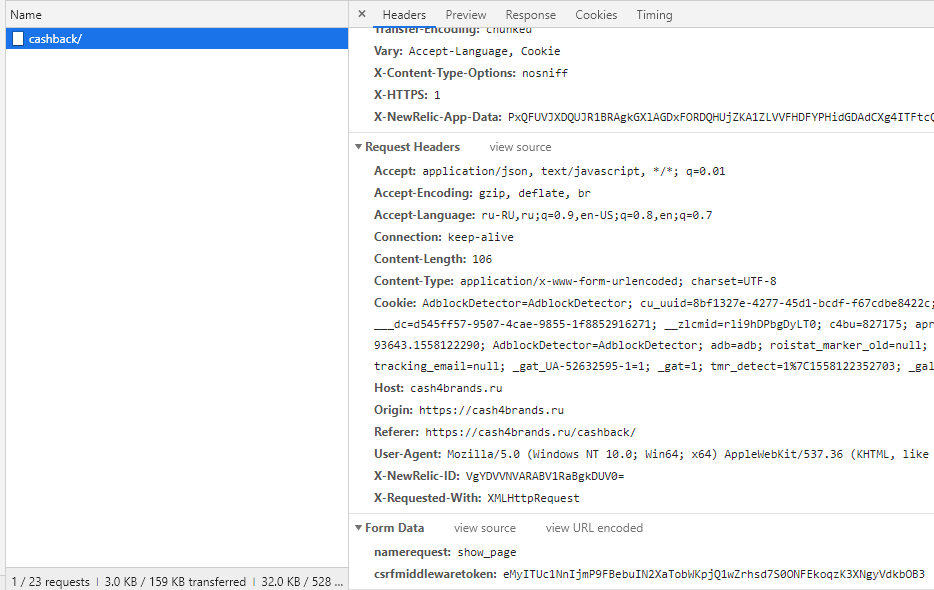


Рисунок 7 – Headers сайта Cash4Brands.ru часть 2

На Рисунке 7 в «Form Data» можно заметить 2 параметра, которые передаются: namerequest и csrfmiddlewaretoken. В данном случае нужен лишь первый параметр, так как второй отвечает за cookies. У поля с именем csrfmiddlewaretoken значением является хэш для идентификатора сессии и секретный ключ. Поле namerequest и его значение show\_page надо добавить в Body. Пример приведён на Рисунок 8.( <https://djbook.ru/ch14s05.html>)

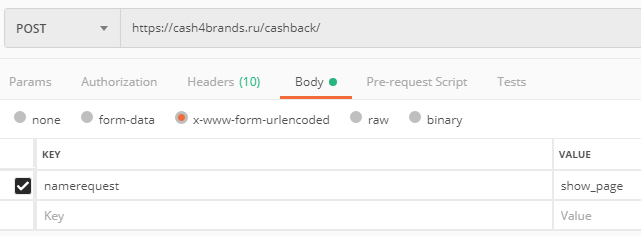


Рисунок 8 – параметры для запроса на Cash4Brands.ru

Результатом данного запроса является JSON, который в дальнейшем обрабатывается.

Программа Postman предоставляет возможность сгенерировать код для запроса. Для этого нужно нажать кнопку Code. В окне можно выбрать язык программирования и какие библиотеки использовать. Пример изображен на Рисунок 9.

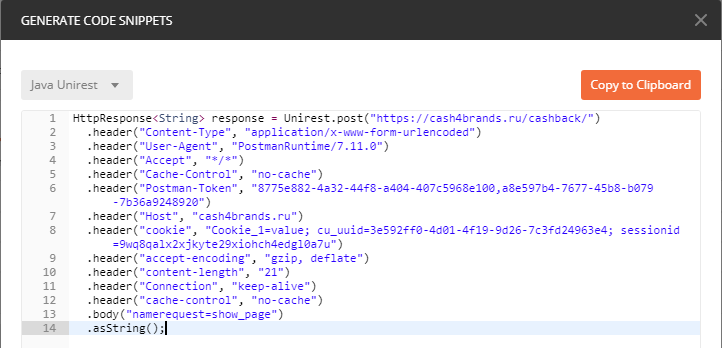


Рисунок 9 – генерирование кода в Postman

После получения JSON с запроса, он обрабатывается с помощью парсера для JSON. Разбор JSON осуществляется с помощью библиотеки JsonPath. Для примера взят код для обработки данных с Kopikot.ru.

1. List<Object> items = JsonPath.read(response, "$..items[\*]");

В этой строчке в список items записываются все элементы, которые содержатся в тэге «items». Далее элементы этого списка передаются методам, которые извлекают из этих элементов нужную информацию.

1. String label = JsonPath.read(item, "$.commission.max.unit");

В этой строчке кода из метода, который извлекает валюту кэшбэка. В данному случае это процент. Через точку происходит обращение к последующим элементам для извлечения информации.

Плюсы Unirest:

* быстрая скорость выполнения.

Минусы Unirest:

* поиск нужных запросов;
* иногда появляется неудобный формат ответа.

# Реализация на языке C#

Во время практики были реализованы следующие способы парсинга контента сайтов на языке Java: парсер HtmlAgilityPack, библиотека Fizzler, библиотека AngleSharp, CsQuery, RestSharp, Selenium.

### **HtmlAgilityPack**

HtmlAgilityPack – это HTML-анализатор, написанный на C# для чтения и записи DOM, и поддерживает обычный XPATH или XSLT – eXtensible Stylesheet Language Transformations.

(<https://habr.com/ru/post/273807/> , <https://habr.com/ru/post/112325/> , <https://html-agility-pack.net/documentation> )

Для парсинга сайта с помощью парсера HtmlAgilityPack были использованы следующие методы:

* SelectNodes(string xpath) – В качестве входного аргумента требует строку в виде XPath выражения. Возвращает HtmlAgilityPack.HtmlNodeCollection, которая содержит коллекцию узлов, соответствующих запросу XPath выражения, или ноль, если ни один узел не соответствует XPath выражению. Метод выбирает список узлов, соответствующих XPath выражению;
* SelectSingleNode(string xpath) – В качестве входного аргумента требует строку в виде XPath выражения. Возвращает первый HtmlAgilityPack.HtmlNode, соответствующий запросу XPath или нулевой ссылке, если не найдено ни одного соответствующего узла. Метод выбирает первый HtmlNode, соответствующий XPath выражению;
* Load(string url) – В качестве входного аргумента требует строку в виде URL. Возвращает объект класса HtmlDocument, который в дальнейшем обрабатывается. Метод получает HTML-документ из интернет-ресурса;
* GetAttributeValue(string name, string def) – В качестве первого входного аргумента требует строку в виде имени атрибута, второго аргумента – строку в виде значения, которое будет возвращено если не будет найден атрибут. Возвращает строку в виде значения атрибута. Метод ищет совпадение по названию атрибута в старшем элементе. Если совпадений нет, то вернётся второй аргумент.

Помимо методов используется свойство InnerText для извлечения текста между начальным и конечным тегами объекта.

Плюсы HtmlAgilityPack:

* средняя скорость выполнения.

Минусы:

* работает только с XPath выражениями.

### **Fizzler**

Fizzler – Библиотека .NET для выбора элементов из дерева узлов на основе селектора CSS. Реализация по умолчанию основана на HTMLAgilityPack и выбирает из документов HTML. (<https://code.google.com/archive/p/fizzler/>).

Для парсинга сайта с помощью библиотеки Fizzler были использованы следующие методы:

* Load(string url) – В качестве входного аргумента требует строку в виде URL. Возвращает объект класса HtmlDocument, который в дальнейшем обрабатывается. Метод получает HTML-документ из интернет-ресурса;
* QuerySelector(string selector) – В качестве входного аргумента требует строку в виде селектора. Возвращает элемент типа HtmlNode. Метод осуществляет поиск элемента по селектору.
* QuerySelectorAll(string selector) – В качестве входного аргумента требует строку в виде селектора. Возвращает список элементов типа HtmlNode. Метод осуществляет поиск элементов по селектору.
* GetAttributeValue(string name, string def) – В качестве первого входного аргумента требует строку в виде имени атрибута, второго аргумента – строку в виде значения, которое будет возвращено если не будет найден атрибут. Возвращает строку в виде значения атрибута. Метод ищет совпадение по названию атрибута в старшем элементе. Если совпадений нет, то вернётся второй аргумент.

Помимо методов используется свойство InnerText для извлечения текста между начальным и конечным тегами объекта.

Плюсы Fizzler:

* быстрая скорость выполнения;
* имеет все те же методы, что и HtmlAgilitypack;
* работает с селекторами.

Минусы Fizzler:

* работает только с XPath выражениями и селекторами.

### **AngleSharp**

AngleSharp – это библиотека .NET, которая позволяет анализировать гипертексты такие как HTML, SVG и MathML. XML также поддерживается библиотекой. Важным аспектом AngleSharp является то, что CSS также может быть проанализирован. Включенный анализатор построен на основе официальной спецификации W3C – World Wide Web Consortium. Также присутствуют стандартные функции DOM, такие как querySelector или querySelectorAll.

(<https://github.com/AngleSharp/AngleSharp>)

Для парсинга сайта с помощью библиотеки AngleSharp были использованы следующие методы:

* ParseDocument(string source) – В качестве входного аргумента требует строку в виде html. Метод парсит переданный аргумент и возвращает результат.
* GetElementsByClassName(string classNames) – В качестве входного аргумента требует строку в виде имени класса. Метод возвращает список уникальных элементов, которые имели класс с именем, которое было передано в виде строки.
* GetElementsByTagName(string tagName) – В качестве входного аргумента требует строку в виде имени тэга. Метод возвращает список уникальных элементов, которые имели тэг с именем, которое было передано в виде строки.
* GetAttribute(string name) – В качестве входного аргумента требует строку в виде имени атрибута. Метод возвращает значение искомого атрибута, либо null, если атрибут не найден.
* QuerySelectorAll(string selectors) – В качестве входного аргумента требует строку в виде селектора. Возвращает элемент типа IElement. Метод осуществляет поиск элемента по селектору.
* QuerySelectorAll(string selector) – В качестве входного аргумента требует строку в виде селектора. Возвращает список элементов типа IElement. Метод осуществляет поиск элементов по селектору.

Помимо методов используется свойство InnerText для извлечения текста между начальным и конечным тегами объекта и свойство InnerHtml для извлечения html кода элемента.

Для получения HTML используется класс WebClient, с помощью которого происходит подключение к страницам сайта.

1. WebClient webClient = new WebClient
2. {
3. Encoding = Encoding.UTF8
4. };
5. string page = "https://letyshops.com/shops?page=" + i;
6. string html = webClient.DownloadString(page);
7. HtmlParser parser = new HtmlParser();
8. var result = parser.ParseDocument(html).GetElementsByClassName("b-teaser");

Плюсы AngleSharp:

* быстрая скорость выполнения;
* при отсутствии элемента возвращает null, а не выбрасывает ошибку.
* работает с селекторами

Минусы AngleShrap:

* выбор по локаторам возвращает только коллекцию

### **CsQuery**

CsQuery – это порт jQuery для .NET 4. Он реализует все селекторы CSS2 и CSS3, все методы манипулирования DOM в jQuery и некоторые служебные методы.

Для парсинга сайта с помощью CsQuery были использованы следующие методы:

* CreateFromUrl(string url) – В качестве входного аргумента требует строку в виде URL. Возвращает HTML. Метод создаёт новое DOM дерево из файла html.
* GetAttribute(string name) – В качестве входного аргумента требует строку в виде имени атрибута. Метод ищет атрибут с заданным именем и возвращает его значение.

# Реализация на языке Python

# Сравнение скорости реализаций

# Описание визуализации результатов

# Описание выхода из специфических ситуаций при парсинге

# Заключение

# Список использованной литературы

# Приложение А.

# Приложение Б.

# Приложение В.