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 3](#_Toc9010997)

[Описание актуальности работы 4](#_Toc9010998)

[Отбор библиотек для парсинга 5](#_Toc9010999)

[Реализация на языке Java 6](#_Toc9011000)

[**Jsoup** 6](#_Toc9011001)

[**Selenium** 9](#_Toc9011002)

[**Unirest** 10](#_Toc9011003)

[Реализация на языке C# 13](#_Toc9011004)

[Реализация на языке Python 14](#_Toc9011005)

[Сравнение скорости реализаций 15](#_Toc9011006)

[Описание визуализации результатов 16](#_Toc9011007)

[Описание выхода из специфических ситуаций при парсинге 17](#_Toc9011008)

[Заключение 18](#_Toc9011009)

[Список использованной литературы 19](#_Toc9011010)

[Приложение А. 20](#_Toc9011011)

[Приложение Б. 21](#_Toc9011012)

[Приложение В. 22](#_Toc9011013)

# 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.

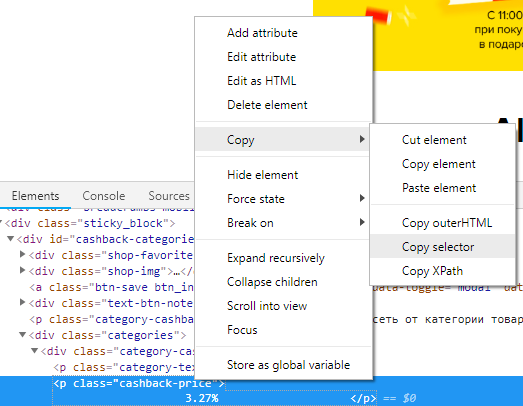


Рисунок 1 – использование возможностей панели разработчика в Google Chrome

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.

(<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.

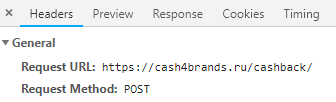
Пример с использование Map:

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 <= 1300; 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. }

Пример без использования Map:

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

Чтобы облегчить работу была использована программа Postman. Как это работает:



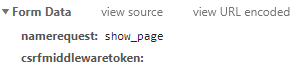
В Response Headers обращаем внимание на это



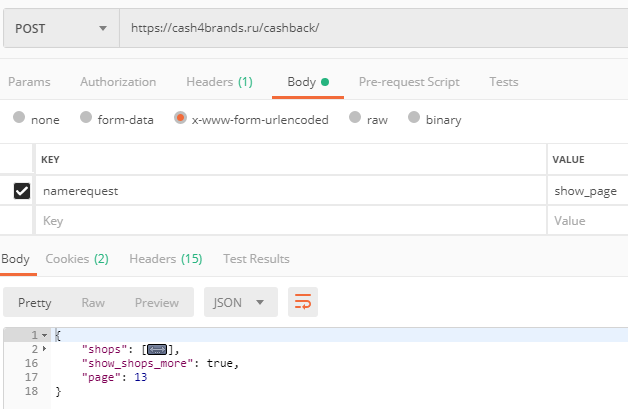
В Request Headers обращаем внимание на это



И на параметры, которые передаются

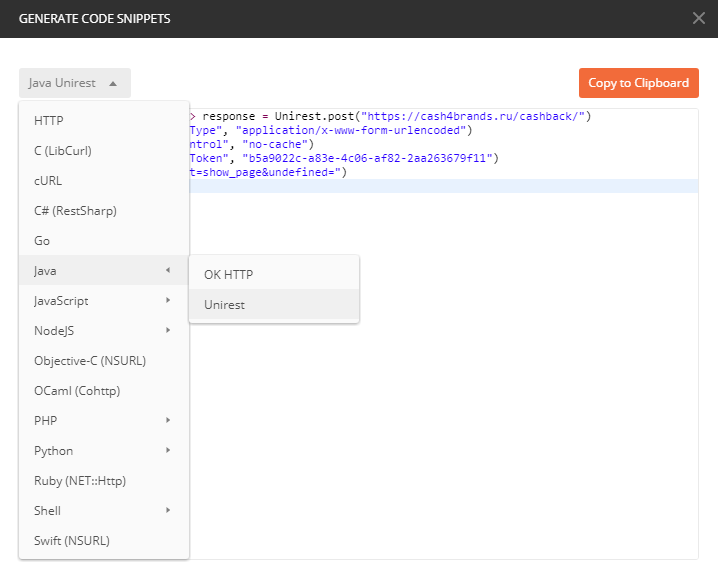


После настройки в Postman получаем результат в виде json:





Также с помощью Postman можно сразу же взять готовый код:



Далее с помощь цикла можно получить все возможные ответы с сайта, в которых содержится информация для парсинга. (В Cash4Brand я использую регулярные выражения, хотя можно использовать JSONPath, как при парсинге Kopikot).

# Реализация на языке C#

# 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.