

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

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

Кафедра информатики и программирования

Анализ методов парсинга сайтов

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 441 группы

направления 02.03.03 Математическое обеспечение и администрирование
информационных систем

факультета компьютерных наук и информационных технологий

Турченкова Павла Александровича

Научный руководитель

Старший преподаватель

подпись, дата

Портенко М.С.

Зав. кафедрой

Доцент, кандидат физико-математических наук

подпись, дата

Огнева М.В.

Саратов 2019

Введение

Актуальность темы. Интернет постоянно расширяется. Каждый день появляется все больше сайтов, на которых размещена различная информация. Человеку с каждым днем труднее найти информацию самостоятельно среди сайтов, поэтому роль парсинга сайтов становится все значительнее.

Парсинг – это метод получения контента, который состоит в том, что специально написанный алгоритм заходит на страницы сайта и собирает информацию в указанных тегах, либо обращается к серверу и берет информацию, размещенную на сайте.

На сегодняшний день существует достаточное количество приложений, которые парсят сайты и систематизируют данные. Например, Screaming Frog SEO Spider или ComparseR [1]. Существуют также компании, которые занимаются парсингом и анализом сайтов. Например, компания Meltwater, которая разрабатывает и продает программное обеспечение для мониторинга СМИ и бизнес-аналитики [2]. К списку компаний, которые разрабатывают программы для парсинга, можно отнести Google и Яндекс [3].

Цель бакалаврской работы – реализация программы для парсинга сайта на языках программирования Java, C# и Python, а также анализ приведенных реализаций.

Поставленная цель определила следующие задачи:

1. Рассмотреть средства для парсинга;
2. Реализовать программы, используя средства для парсинга;
3. Сравнить реализации по скорости выполнения парсинга и дать оценку;
4. Описание визуализации результатов.

Теоретическая значимость бакалаврской работы. В работе описаны такие средства для парсинга сайта, как:

1. Парсинг контента сайта через библиотеку для парсинга html кода;
2. Парсинг контента сайта через отправку запросов на сервер;
3. Парсинг контента сайта через инструменты тестирования.

Для каждого типа средств для парсинга приведены и описаны библиотеки, фреймворки и инструменты тестирования.

Для каждого средства представлены основные методы для парсинга контента сайта.

Описаны некоторые нюансы, которые могут возникнуть при парсинге сайтов.

Практическая значимость бакалаврской работы. В работе описан алгоритм обхода сайтов LetyShops.ru, MegaBonus.ru и Kopikot.ru. Алгоритмы дают представление о парсинге сайта, а также на примере показывают, как работать со средствами для парсинга.

Для каждого приведённого в теоретической части средства приведена реализация в виде программы для парсинга контента сайта с кэшбэками.

Структура и объём работы. Бакалаврская работа состоит из введения, 8 разделов, заключения, списка использованных источников и 13 приложений. Общий объем работы – 112 страниц, из них 65 страниц – основное содержание, включая 31 рисунок и список использованных источников информации – 48 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Виды парсинга» посвящен классификации средств для парсинга.

Данный раздел содержит в себе классификацию средств для парсинга:

1. Парсинг контента сайта через библиотеку для парсинга;
2. Парсинг контента сайта через отправку запросов на сервер;
3. Парсинг контента сайта через инструменты тестирования.

Также раздел содержит такие основные определения, как html, css и DOM дерево.

В конце описано насколько законен парсинг на территории РФ, а также какие ограничения накладываются на программиста во время разработки программы для парсинга сайтов.

Второй раздел «Общий обзор средств для парсинга» посвящен средствам и методам, которые рассматриваются в дипломной работе.

Данный раздел содержит перечисление средств для каждого из языков программирования:

1. Java – 2 библиотеки;
2. C# - 5 библиотек;
3. Python – 2 библиотеки и 1 фреймворк.

Также для всех языков программирования используется инструмент тестирования Selenium.

В приведённых средствах есть все указанные в первой главе типы средств для парсинга сайтов.

Третий раздел «Обзор средств для парсинга на языке Java» посвящен рассмотрению средств для парсинга на языке программирования Java.

Данный раздел содержит описание таких средств для парсинга, как библиотеки Jsoup, Unirest и инструмента для тестирования Selenium на Java.

Помимо описания приводятся основные методы приведённых средств для парсинга. Обзор каждого средства заканчивается плюсами и минусами.

Также описаны некоторые нюансы при парсинге сайтов с использованием данных средств.

В конце предоставлена ссылка на приложение, где реализуется рассмотренный метод парсинга.

Четвертый раздел «Обзор средств для парсинга на языке C#» посвящен рассмотрению средств для парсинга на языке программирования C#.

Данный раздел содержит описание таких средств для парсинга, как библиотеки HtmlAgilityPack, Fizzler, AngleSharp, CsQuery, RestSharp и инструмента для тестирования Selenium на C#.

Помимо описания приводятся основные методы приведённых средств для парсинга. Обзор каждого средства заканчивается плюсами и минусами.

Также описаны некоторые нюансы при парсинге сайтов с использованием данных средств.

В конце предоставлена ссылка на приложение, где реализуется рассмотренный метод парсинга.

Пятый раздел «Обзор средств для парсинга на языке Python» посвящен рассмотрению средств для парсинга на языке программирования Python.

Данный раздел содержит описание таких средств для парсинга, как библиотеки BeautifulSoup, Requests, фреймворка Scrapy и инструмента для тестирования Selenium на Python.

Помимо описания приводятся основные методы приведённых средств для парсинга. Обзор каждого средства заканчивается плюсами и минусами.

Также описаны некоторые нюансы при парсинге сайтов с использованием данных средств.

В конце предоставлена ссылка на приложение, где реализуется рассмотренный метод парсинга.

Шестой раздел «Алгоритмы парсинга сайтов» посвящен разбору основных алгоритмов, которые были использованы при реализации средств для парсинга.

Данный раздел содержит описание алгоритмов, которые были разработаны самостоятельно в процессе реализации средств для парсинга сайтов LetyShops.ru, MegaBonus.ru и Kopikot.ru. Цель алгоритма заключается в нахождении информации о магазинах, которые предоставляют кэшбэки клиентам. Под информацией понимается имя магазина, значение кэшбэка,

валюта кэшбэка, изображение логотипа магазина и ссылка на страницу магазина на сайте.

Для первого сайта использовалась библиотека на языке программирования C# Fizzler, которая работает с html кодом напрямую. Аналогичный алгоритм работает для остальных средств такого же типа.

Для второго сайта – инструмента тестирования Selenium на языке программирования C#. Аналогичный алгоритм работает для остальных средств такого же типа.

Для третьего сайта – библиотека на языке программирования Java Unirest, которая позволяет отправлять запросы на сервер. Аналогичный алгоритм работает для остальных средств такого же типа.

В конце раздела описаны нюансы, которые могут встретиться при парсинге иных сайтов.

Седьмой раздел «Сравнение скорости реализаций» посвящен рассмотрению результатов замеров времени выполнения реализаций средств парсинга.

Данный раздел содержит рисунки, демонстрирующие время выполнения реализаций для каждого из средств для парсинга.

Время выполнения измеряется в секундах.

Помимо замеров основных средств для парсинга сайтов, использовалась комбинация из инструмента тестирования Selenium и библиотеки для работы с html кодом.

В конце написан вывод, основанный на сделанных замерах.

Восьмой раздел «Визуализации данных» посвящен описанию средств для более комфортного ознакомления с данными, добытыми при помощи парсинга.

Данный раздел содержит описание средств для визуализации данных для всех трёх языков программирования, таких как, использование паттерна MVC, файл формата Excel, файл формата CSV.

Для реализации паттерна MVC были использованы следующие средства:

1. Для Java был использован Spring;
2. Для C# был использован ASP.NET;
3. Для Python был использован Django.

Для создания Excel файла были использованы специальные библиотеки:

4. Для Java был использован Apache POI;
5. Для C# был использован OfficeOpenXml;
6. Для Python был использован openpyxl.

Для создания CSV файла были использованы стандартные средства языков программирования для записи данных в файл.

Также представлены рисунки результатов данных визуализаций.

Заключение

В данной работе были рассмотрены библиотеки для парсинга сайтов на трех языках программирования и их основной функционал. Помимо этого, были написаны программы, которые, используя библиотеки для парсинга сайтов, собирают и систематизируют данные. Также произведен анализ скорости выполнения различных реализаций библиотек.

В ходе анализа скорости выполнения были сделаны выводы, что для использования Selenium лучше использовать версию на Java. Среди парсеров, которые реализуют библиотеки для парсинга, что быстрее всего работают библиотека Fizzler на языке программирования C# и библиотека Jsoup на языке программирования Java. Среди библиотек и фреймворков, которые работают с запросами, быстрее являются RestSharp и Unirest.

В ходе работы также были описаны способы решения проблем, возникающие при написании программы и самом парсинге сайтов и способы визуализации данных для лучшего ознакомления с ними.

Список использованной литературы

1. Обзор ТОП-6 парсеров сайтов [Электронный ресурс]: сайт. URL: <https://siteclinic.ru/blog/seo-instrumenty/obzor-parserov/> (дата обращения: 22.05.2019). Загл. с экрана. Яз. рус.
2. Meltwater [Электронный ресурс]: сайт. URL: <https://www.meltwater.com/?ucs> (дата обращения: 22.05.2019). Загл. с экрана. Яз. англ.
3. Парсинг сайтов: как с точки зрения закона выглядит один из самых полезных ИТ- инструментов по миру (и в России)? [Электронный ресурс]: сайт. URL: <https://habr.com/ru/post/340302/> (дата обращения: 22.05.2019). Загл. с экрана. Яз. рус.
4. SELENIUM / WEBDRIVER. Введение. [Электронный ресурс]: сайт. URL: <https://selenium2.ru/docs/introduction.html> (дата обращения: 22.05.2019). Загл. с экрана. Яз. рус.
5. Использование Java JSoup для анализа кода HTML [Электронный ресурс]: сайт. URL: <https://o7planning.org/ru/10399/jsoup-java-html-parser-tutorial> (дата обращения: 22.05.2019). Загл. с экрана. Яз. рус.
6. Unirest for Java [Электронный ресурс]: сайт. URL: <http://unirest.io/java.html> (дата обращения: 22.05.2019). Загл. с экрана. Яз. англ.
7. Распарсить HTML в .NET и выжить: анализ и сравнение библиотек [Электронный ресурс]: сайт. URL: <https://habr.com/ru/post/273807/> (дата обращения: 22.05.2019). Загл. с экрана. Яз. рус.
8. Beautiful Soup Documentation [Электронный ресурс]: документация. URL: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/> (дата обращения: 22.05.2019). Загл. с экрана. Яз. англ.
9. WEB SCRAPING С ПОМОЩЬЮ SCRAPY И PYTHON 3 [Электронный ресурс]: сайт. URL: <https://www.8host.com/blog/web-scraping-s-pomoshhyu-scrapy-i-python-3/> (дата обращения: 22.05.2019). Загл. с экрана. Яз. рус.

10.Python: модуль requests [Электронный ресурс]: сайт. URL:
<https://rtfm.co.ua/python-modul-requests/> (дата обращения: 22.05.2019).
Загл. с экрана. Яз. рус.