Что такое Webscraping?

Web Scraping — один из самых популярных методов считывания различных данных, расположенных на веб-страницах, для их обработки, анализа и так далее.

По сути, это можно назвать “парсингом сайтов”, где информация собирается в более удобный для пользователя формат будь то таблица текстовый документ, в любом читаемом формате.

Метод состоит в том, что специально обученный алгоритм заходит на главную страницу сайта и начинает переходить по всем внутренним ссылкам, тщательно собирая информацию заданную в алгоритме.

Для чего используется Web Scraping?

Сбор данных для маркетинговых исследований.

Для извлечения определенной информации (телефонов, е-мейлов, адресов) с различных сайтов для создания собственных списков.

Сбор данных о товарах для анализа конкурентов.

Сбор финансовых данных.

В работе HR для отслеживания резюме и вакансий.

Пример

Минимум кода, просто хочу рассказать и показать основные принципы скрапинга

Постановка задачи

Обычно, задача для веб-скрейпинга выглядит так: есть данные, доступные только на веб-страницах, и их надо оттуда вытащить и сохранить в неком удобоваримом формате. Конечный формат не важен. По большей части речь о том, чтобы открыть браузер, пройтись мышкой по ссылкам и скопипейстить со страниц нужные данные. Ну, или сделать то же самое скриптом

Инструменты Скрапинга

Для автоматического использования браузера из скрипта на Node.JS есть много популярных инструментов.

PhantomJS – браузер на движке WebKit. Он запускает сайты в реальном браузере, только не показывает пользователю страниц, а вместо пользовательского интерфейса использует скрипты на Javascript.

SlimerJS – грубо говоря это тот же PhantomJS, только не на WebKit, а на Gecko.

CasperJS – самая известная Node.js-обёртка над PhantomJS и SlimerJS.

Electron – ещё один инструмент для создания десктопных приложений на базе веб-технологий. с его помощью также можно делать приложения для скрейпинга с отличной индикацией. Кроме того, Electron легко можно использовать в качестве headless-браузера

Nightmare – самая известная обёртка над Electron. Несмотря на название производит очень хорошее впечатление. Спроектирован настолько гладко, что некоторые начинающие скрейперы умудряются использовать Nightmare вообще ничего не зная про Electron. По интерфейсу очень похож на Horseman.js (и почти так же хорош).

Puppeteer – это библиотека Node.js, поддерживаемая командой Chrome Devtools. Библиотека запускает экземпляр Chrome/Chromium и предоставляет набор высокоуровневых API.

Puppeteer используется для выполнения множества различных задач:

автоматизация сбора данных с веб-сайтов;

создание скриншотов и PDF-файлов;

тестирование расширений Chrome;

автоматизация тестирования веб-интерфейсов;

диагностика проблем производительности с помощью таких методов, как захват временной шкалы трассировки веб-сайта.

Парсинг

В нашем случае DOM-парсер отлично подойдёт. В настоящий момент среди DOM-парсеров под Node.js уверенно лидирует cheerio – серверная версия культового JQuery.

Трудности Скрапинга

Первая плохая новость – весь выбор и отображение данных на сайте происходит динамически на одной и той же странице. Выглядит это так: при обновлении показывается выпадающий список, после выбора пункта – ещё выпадающий список (а иногда после него и ещё один), а потом появляется карта выбранного региона.