УДК 004.415.25

В. Ю. Геновська
О. Р. Бойко

ПРОГРАМНИЙ ДОДАТОК НА ОСНОВІ БІБЛІОТЕКИ PYTHON REQUESTS

Вінницький національний технічний університет

Анотація
Розглянуто спосіб реалізації методу web-scrapping на основі бібліотеки Python Requests
Ключові слова: web-scrapping, Python, Requests, проксі-служби.

Abstract
The method of implementation of web-scrapping based on Requests Library in Python.
Keywords: web-scrapping, Python, Requests, proxy services.

Вступ

Підприємства, які не покладаються на дані, мають дуже низький шанс на успіх у світі, керованому даними. Одним з кращих джерел даних є дані, доступні публічно в Інтернеті на різних веб-сайтах, для отримання яких потрібно використовувати техніку, яка називається Web Scraping або Data Scraping.

Результати дослідження

В результаті аналізу літературних джерел [1-3] виділимо такі основні завдання, які потрібно виконати для успішного процесу збирання, обробки, зберігання та оновлення релевантних даних з потрібних нам веб-сайтів:

1. Розробка web-scrapper та установка серверів
2. Запуск web-scrapper
3. Зберігання даних
4. Обхід чорного списку та CAPTCHA
5. Перевірка якості даних
6. Технічне обслуговування

Для великомасштабних проектів доцільно розглядати такий фреймфорк як Scrapy, написаний мовою Python. Хоча він є безумовним лідером, має безліч переваг, серед яких асинхронність, гнучкість та надійність. Проте для невеликих проектів, які потребують опрацювання невеликої кількості сторінок такий інструмент буде надто громіздким та ресурсозатратним. Альтернативою є HTTP-бібліотека Python – Python Requests, яка є не лише більш легкою в розумінні та простішою у використанні, а також вимагає менше процесорної роботи.

Складність реалізації web-scrapping полягає в тому, що велика кількість веб сайтів застосовують анти-web-scrapping заходи. Якщо будь-який з цільових веб-сайтів має будь-який тип блокування на основі IP-адреси, IP-адреса серверів буде занесена в чорний список, і сайт не буде відповідати на запити наших серверів.

Рішення такої проблеми є використання динамічних IP адрес, проксі. Є багато проксі служб, серед яких було використано Intoli. Intoli - це розумна проксі-служба, яка може похвалитися великим пулом проксі-серверів. Він автоматично обертає проксі-сервер на кожному HTTP-запиті, а також дозволяє повторно використовувати той самий проксі-сервер, та використовувати беззаголовочний браузер для відображення динамічних сторінок JavaScript. Помилка запитів автоматично повторюється, і вони використовують інтелектуальні алгоритми маршрутизації, щоб уникнути виявлення.

Висновки

Під час аналізу виявлено, що для web-scrapping може застосовується велика кількість бібліотек та фреймворків, проте для невеликих проектів ідеальним рішенням є бібліотека Requests, яка написана на мові Python. Також було розглянуто проблеми, які виникають в процесі збирання інформації сайтів та було знайдено рішення - використання проксі служб.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Документація бібліотеки Requests [Електронний ресурс] – Режим доступу: http://docs.python-requests.org/en/master/user/intro/#philosophy
2. ScrapeHero – How to rotate proxies and ip-adresses (Як змінювати проксі та IP адреси) [Електронний ресурс] – Режим доступу: https://www.scrapehero.com/how-to-rotate-proxies-and-ip-addresses-using-python-
3. Как обойти серверную блокировку [Електронний ресурс] – Режим доступу: https://m.habr.com/ru/company/ods/blog/346632/

Геновська Влада Ю.— студентка групи 1СІ-15б, факультет комп'ютерних систем та автоматики, Вінницький національний технічний університет, Вінниця, e-mail: vhenovska@gmail.com

Науковий керівник: ***Бойко Олексій Р.*** – кандидата технічних наук, доцента кафедри автоматизації та інтелектуальних інформаційних технологій, Вінницький національний технічний університет, м. Вінниця.

Henovska Vlada Y. — Faculty of Computer Systems and Automatics, Vinnytsia National Technical University, Vinnytsia, email : vhenovska@gmail.com

Supervisor: ***Boyko Oleksiy R.*** - candidate of technical sciences, Docent of Automation and Intellectual Information Technologies Department, Vinnytsia National Technical University, the city of Vinnitsa