УДК 004.056

ПАРСИНГ ДАНИХ З ВЕБ СТОРІНОК

Брітвін А., Ткачук Р.
Національний університет "Львівська політехніка", м .Львів
Львівський державний університет безпеки життєдіяльності, м .Львів

В роботі описано основні ключові фактори розробки автоматизованої системи парсингу інформації з веб сторінок. Ключовим фактором дослідження є усвідомлення необхідності оптимізації алгоритму, створення безперебійної системи з автоматичним відновленням стану, та мінімізацією шансу втрати даних.

Ключові слова: *аналіз веб-сторінки, великі дані, модель клієнт-сервер, побудова системи.*

The paper describes the main key factors in the development of an automated system for parsing information from web pages. A key factor in the study is the awareness of the need to optimize the algorithm, create a seamless system with automatic recovery, and minimize the chance of data loss.

Keywords: *web page analysis, big data, client-server model, system construction.*

З моменту впровадження великих обсягів даних у наші сучасні бізнес моделі потреба у вилученні, аналізі та обробці даних стає все більш важливою для компаній у всіх галузях промисловості. Зі збільшенням збору даних зростає потреба їх читати та розуміти [1].

Парсинг даних – це процес взяття даних в одному форматі та перетворення їх в інший формат. Цей процес є дуже важливий у сучасному світі, оскільки сьогоднішнє життя переходить на цифровий рівень.

Парсинг даних включає у себе лексичний і синтаксичний аналіз. Загалом його можна охарактеризувати як процес аналізу рядка символів у мові, що відповідає правилам формальної граматики. Аналіз з точки зору дослідження даних розширює це визначення у двоетапний процес, в якому синтаксичний аналізатор програмовано вказує, які дані читати, аналізувати або перетворювати. Результатом зазвичай є більш структурований формат.

Важливим аспектом придатності даних для цілей є структура, в якій вони знаходяться. Часто сама структура не підходить для потреб даних [2].

Веб парсинг – це метод отримання великого обсягу загальнодоступних даних з веб -сайтів. Він автоматизує збір даних і перетворює зібрані дані у формати на ваш вибір, такі як HTML, CSV, Excel, JSON, txt [3]. Цей просес насамперед складається з 3 частин:

– аналіз HTML сторінки;

– видобуток даних;

– зберігання даних.

Найважливіше у парсингу даних – це програмування. Через це багатьом компаніям потрібно наймати досвідчених розробників для сканування

веб-сайтів. Тоді як для тих, хто не має великого бюджету та не володіє навичками кодування, стануть у нагоді інструменти для скребку в Інтернеті. І парсингові мови програмування, і використання інструментів веб-парсингу мають деякі спільні переваги. Серед основних переваг парсингу даних можна виділити:

– видобуток даних автоматизований;
– висока швидкість;
– зібрана інформація є набагато точнішою ніж зібранна в ручну;
– отримання чистих та структурованих даних.

Після збору даних зазвичай відбувається їх очищення та реорганізація, оскільки зібрані дані не є структурованими та готовими до використання. Інструменти веб-скребку перетворюють неструктуровані та напівструктуровані дані у структуровані, а інформація веб-сторінки реорганізовується у презентабельні формати [3, 4].

Для того щоб створити сценарії парсингу даних з веб сторінок, нам необхідно зробити декілька кроків з налаштування нашого проекту. Першим кроком необхідно додати в проект бібліотеку selenium webdriver. Далі створимо екземпляр Chrome із шляхом до драйвера, який завантажили через веб-сайти відповідного браузера. Далі використовуючи метод .get () драйвера для завантаження веб-сайту. Ми також можемо завантажити локальний сайт розробки, оскільки цей процес еквівалентний відкриттю вікна Chrome на локальній машині, введенню URL-адреси та натисканню Enter. Метод .get () не тільки починає завантаження веб-сайту, але і чекає його повного візуалізації, перш ніж перейти до наступного кроку. Після успішного завантаження сторінки можна скористатися атрибутом .title для доступу до текстового заголовка веб-сторінки.

Тобто, для того щоб пропарсити дані з веб-сторінки нам необхідно створити драйвер браузера, який буде імітувати звичайний веб-браузер. Далі необхідно задати адресу потрібного нам сайту, та користуючись командами даної бібліотеки отримати необхідні елементи сторінки.

Література

1. Viktor Mayer-Schönberger. Big Data: A Revolution That Will Transform How We Live, Work /, Kenneth Cukier., 2014. – 280 с. – (3). – (2).

2. Cathy O'Neil, Rachel Schutt. Doing Data Science: Straight Talk from the Frontline / Cathy O'Neil, Rachel Schutt., 2013. – 510 с. – (O'Reilly Media, Inc.).

3. Selenium is an umbrella project for a range of tools and libraries that enable and support the automation of web browsers. [Електронний ресурс] // Selenium. – 2021. – Режим доступу до ресурсу: https://github.com/SeleniumHQ.

4. HTML [Електронний ресурс]. – 2016. – Режим доступу до ресурсу: https://developer.mozilla.org/ru/docs/Web/HTML.