DOI 10.31891/2307-5732-2022-305-1-50-52
УДК 004.93, 004.8

КРИВЕНЧУК Ю. П.
https://orcid.org/0000-0002-2504-5833
e-mail: Yurii.P.Kryvenchuk@lpnu.ua
БУРАК М. Т.
https://orcid.org/0000-0002-8979-3347
e-mail: burakmarko@gmail.com
Національний університет "Львівська політехніка"

ПОРІВНЯЛЬНИЙ АНАЛІЗ ЕФЕКТИВНОСТІ НАДБУДОВ SELENIUM ТА BEAUTIFULSOUP

На сьогодні кількість цифрової інформації у всесвітній мережі збільшується експоненціально з кожним роком. Тому зріс попит на аналіз даних з вебресурсів. Проте для проведення операцій з даними їх потрібно спочатку отримати з джерела. Існує дуже багато інструментів, написаних під мову python для роботи з видобутком інформації, з яких selenium та beautifulSoup вважаються найпопулярнішими. Попри те, що обидві надбудови працюють досить добре, розробники прикладного програмного забезпечення намагаються обрати найбільш оптимальну з них. Тому виникла необхідність перевірки цих двох пакетів на ефективність. У роботі розглянуто тривалість роботи парсерів для пошуку тегів на вебсторінці за допомогою різних методів та платформ. Дослідження проведено на основі онлайн-платформ для продажу товарів. Результати показали, які саме інструменти та функції найкраще використовувати для знаходження товару на інтернет-магазинах.

Ключові слова: парсер, selenium, beautifulSoup, python, інтернет-магазин, тег, вебсторінка, пошук.

YURII KRYVENCHUK, MARKO BURAK
Lviv Polytechnic National University

COMPARATIVE ANALYSIS OF SELENIUM AND BEAUTIFULSOUP EFFICIENCY

Nowadays, the amount of digital information on the World Wide Web is growing exponentially every year. Therefore, the demand for data analysis from web resources has increased. However, to perform data operations, information must first be obtained from the source. Today almost every popular programming language has at least one library that can perform web scraping operations and extract data from websites, although some of them are hard to use or not compatible with the language of the projects, that this data is intended. Therefore, a lot of developers use python as the main tool for such projects. It can be used to build almost any platform and communicate with the parsers within a project. Also, this language is easy to use and has a huge community. There are many python-based tools for working with data mining, of which selenium and beautifulSoup are considered the most popular. Despite the fact that both add-ons work quite well, the developers strive to choose the most optimal one. Thus, there is a need to test these two packages for effectiveness.

The paper considers the duration of parsers to search for tags on a web page using different methods and platforms. The study was conducted on the basis of online platforms for the sale of goods. The results showed which tools and functions are the best choices to find products on online stores. The object of analysis was the website "Rozetka", which is the biggest and the most popular online store in Ukraine. This article has described the advantages and disadvantages of using these libraries especially for scraping data from online stores. To analyze these add-ons, there was a special program created which can open the website on a browser, find the search bar, and enter the name of the desired product, afterward using various methods of these libraries perform a product search. For each search time was recorded. The results showed that beautifulSoup can find tags faster than selenium in general, however, for searching and web scraping online stores selenium can perform better and is more suitable.

Keywords: web scraper, selenium, beautifulSoup, python, online store, tag, web page, search.

Постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями

Сьогодні сфера видобутку даних з вебсторінок досить добре розвинена і безліч теперішніх мов програмування мають надбудови, які з легкістю можуть знайти будь-які дані з цих джерел. Проте не завжди зручно використовувати певні мови для цих задач. Наприклад, для платформ написаних одною мовою не доречно використовувати парсер іншої – це накопичує надлишкову інформацію, ускладнює реалізацію і рідко є ефективним вирішенням проблеми. Зараз дуже популярною є мова python, адже на основі цієї мови можна зробити майже будь-який проєкт, тому її часто використовують для створення продуктів. Через таку популярність і зручність цієї мови, вона також стала хорошим інструментом для пошуку інформації на теренах інтернету.

Тим не менш, вибір мови – це не лише єдина проблема для розробників, найважчим є саме вибір найкращого інструменту серед пакетів для видобутку даних, адже усі вони здебільшого виконують однакові функції. Для вибору надбудов не лише потрібно знати, яка з них показує кращу ефективність, а й врахувати, з якого саме джерела видобувається інформація.

Позаяк онлайн-магазини містять великий об'єм даних, саме вони часто є джерелом інформації. Оскільки кількість таких магазинів дуже зросла за останні роки, а попит на інформацію з них надалі буде лише збільшуватися, саме інтернет-платформи стали об'єктом нашого дослідження.

Визначення ефективності кожного з інструментів включало не лише сам пошук даних, але й ініціалізацію модулів, які також впливають на час виконання програми.

Аналіз досліджень та публікацій

Згідно з статтею Рейчел Чапман [1] Selenium – це API сервіс, який дає змогу використовувати інтерактивні елементи браузера, створювати автоматичні тести чи знаходити дані на вебресурсах. Також ця

надбудова дає можливість опрацьовувати код javascript, тому має більше функцій, ніж інші. Також це єдина бібліотека, що уможливлює пошук прихованих елементів на сторінках.

У статті [1] описано надбудову BeautifulSoup – це бібліотека для мови python, яка дає змогу знаходити дані з вебсторінок за допомогою допоміжної бібліотеки request. Пошук відбувається серед тегів HTML сторінки, яка формується після GET запиту до адреси. Надбудова дуже проста у використанні, її функції зрозумілі, з її допомогою можна опрацьовувати також XML документи, та все ж її можливості доволі обмежені.

Selenium часто використовують передовсім для створення послідовності дій і уже потім для пошуку елементів, у той час, як beautifulSoup вміє лише шукати елементи з конкретної сторінки.

Ці дві бібліотеки добре справляються з пошуком інформації з документів, але, вважається, що beautifulSoup працює повільніше через просту реалізацію алгоритмів. З іншого боку, ця бібліотека показує чудові результати з необ'ємними завданнями.

Також варто зауважити, що є можливість поєднювати ці бібліотеки, а саме використовувати selenium для проведення операцій з інтерактивнами елементами, тоді як beautifulSoup може забезпечити простий пошук елементів. Однак, слід зауважити, що цей метод не завжди ефективний через додаткову ініціалізацію парсера.

Тим не менш, не всі погоджуються з твердженням, що beautifulSoup, працює повільніше за selenium. Так Франк Адраде [2] зазначає, що beautifulSoup працює швидше завдяки швидшому опрацюванню HTTP запитів, у той час як Selenium є повільнішим, бо опрацьовує всі файл з вебсторінки. Зазначені статті датовані 2021 роком, а тому порівняльний аналіз ефективності описаних у них інструментів є дуже актуальним на сьогодні.

Слід також зазначити, що вже були спроби описати деякі алгоритми пошуку даних в онлайн-магазинах. Так у статті [3] описано, алгоритм знаходження даних для сайту amazon.com з використанням бібліотеки selenium. З цієї публікації видно, що надбудова selenium є надпотужна і вміщає безліч корисних функцій для створення парсера. Також у статті використано бібліотеку pandas, яка допомогла автору зберігати дані після знаходження їх на сайті.

Формулювання цілей статті

Метою роботи є порівняти час виконання завдання інструментами selenium і selenium та вивчити їхні методи пошуку елементів у HTML файлі на основі даних онлайн-магазину.

Виклад основного матеріалу

Для досягнення мети цього дослідження було створено програмну реалізацію парсерів мовою python, яка дала змогу взаємодіяти з сервісом онлайн-магазину та шукати теги продуктів на ньому за допомогою надбудов selenium та beautifulSoup.

Джерелом даних було обрано сайт "Rozetka" [4], який є набільшим сервісом в Україні серед онлайн-магазинів. Сайт дуже популярний серед українців, тому є чудовим об'єктом для аналізу. Предметом пошуку на платформі було обрано популярний на ринку мобільний телефон компанії Apple iphone 12 Pro Max.

На жаль, запит GET не надає html документу з продуктом, оскільки магазин блокує запит на нього. А отже єдиний спосіб, яким можна отримати цей елемент, – це можливість інтерактиву з javascript, що дає змогу вписати назву товару в пошукову стрічку, а відтак досліджувати вихідний документ. Єдиним інструментом, який може взаємодіяти з javascript, є selenium, а отже його ініціалізація є обов'язковою для обох надбудов для подальшого аналізу сторінки. У зв'язку з цим час на її активацію в нашому дослідженні буде опущено.

Дослідження включало два експерименти:

- порівняння часу лише для пошуку елемента;
- порівняння часу для пошуку елемента з урахуванням ініціалізації модуля beautifulSoup.

Отримані дані покликані допомогти зрозуміти, яка з надбудов працює швидше для пошуку елемента, а також показати доцільність використання модулів для видобування даних з інтернет-магазинів.

Для дослідження було створено спеціалізований програмний продукт, при реалізації якого спочатку викликається метод driver.get, який відкриває сайт магазину, опісля знаходиться пошукова стрічка і вводиться назва товару "iphone 12 Pro Max" за допомогою методу send_keys(). Після завантаження сторінки з усіма товарами програма отримує html сторінку і шукає бажаний елемент.

Для аналізу використано такі методи бібліотеки selenium:

- By.XPATH
- By.CLASS_NAME
- By.CSS_SELECTOR

Оскільки бібліотека beautifulSoup не може самостійно обробляти XPATH, було досліджено лише два методи, які є відповідниками до selenium:

- find_all(*, class_=)
- select()

Результати виконання програми для кожного з методів з пошуку елемента представлені в таблиці 1.

Підсумки апробації програми пошуку елемента з урахуванням ініціалізації надбудови beautifulSoup представлені в таблиці 2.

Таблиця 1

Швидкодія методів для пошуку елемента

Метод	Час, с
By.XPATH	0.343468
By.CLASS_NAME	0.259308
By.CSS_SELECTOR	0.039893
find_all(*, class_=)	0.014951
select()	0.033909

Таблиця 2

Швидкодія методів з урахуванням ініціалізації

Метод	Час, с
By.XPATH	0.343468
By.CLASS_NAME	0.259308
By.CSS_SELECTOR	0.039893
find_all(*, class_=)	0.463759
select()	0.569477

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямі

Під час виконання даного дослідження було проаналізовано два інструменти для мови python, а також перевірено на швидкодію їхні методи. Враховуючи усі результати з таблицях 1 та 2, можна стверджувати, що надбудова beautifulSoup, справді, знаходить елементи на html документах швидше за надбудову selenium: усі методи цього пакету показали кращий час для знаходження тегів, найкращим з них виявився метод find_all(), який знайшов інформацію удвічі швидше за метод select() та в рази перевершив виконання майже усіх функцій модуля selenium.

Однак в першому експеременті не було враховано час ініціалізації парсера, що суттєво змінює час виконання: з таблиці 2 видно, що швидкодія збільшилась у понад 10 разів, а це більше за будь-який метод надбудови selenium. Тому можна зробити висновок, що за можливості використання HTTP запитів, бібліотека beautifulSoup дійсно працює швидше для пошуку тегів з вебсторінок, але якщо такої можливості немає і потрібно інтерактивно працювати зі скриптами javascript, то доцільніше використовувати надбудову selenium. Позаяк дослідження було проведене на основі інтернет-магазинів, для яких здебільшого необхідно інтерактивно комунікувати з вебсторінкою, то краще використовувати бібліотеку selenium, яка зменшить час апробації програми та ефективніше виконання завдання.

Література

1. Порівняння між Selenium та BeautifulSoup: який найкращий? Limeproxiess. Порівняння між Selenium та BeautifulSoup: який найкращий? Limeproxies. URL: https://limeproxies.netlify.com/blog/selenium-vs-beautifulsoup

2. Андраде Ф. Веб-парсинг за допомогою beautifulSoup, Selenium чи Scrapy? Medium. 2021. URL: https://towardsdatascience.com/web-scraping-with-beautiful-soup-selenium-or-scrapy-62c6f3545de7

3. Бхатачарія С. Парсинг сайтів електронної комерції за допомогою Selenium та Python. Analytics Vidhya. 2020. URL: https://medium.com/analytics-vidhya/web-scraping-e-commerce-sites-using-selenium-python-55fd980fe2fc

4. Інтернет-магазин ROZETKATM: офіційний сайт найпопулярнішого онлайн-гіпермаркету в Україні. URL: https://rozetka.com.ua/ua/

5. Кхдер МА. Парсинг або веб-сканування: Сучасний стан, техніки, підходи та застосування. International Journal of Advances in Soft Computing and its Applications. 2021; 13(3):144–68.

6. Томас ДМ, Матур С. Дата аналіз для парсингу з використанням python. 2019. с. 450–4.

7. МакХенлі Р. Навчання: Текстова аналітика для моделювання за допомогою python. 2021. с. 68–82.

References

1. Comparison Between Selenium vs BeautifulSoup: Which Is the Best One? Limeproxiess. Comparison Between Selenium vs BeautifulSoup: Which Is the Best One? Limeproxies. URL: https://limeproxies.netlify.com/blog/selenium-vs-beautifulsoup

2. Andrade F. Web Scraping with Beautiful Soup, Selenium or Scrapy? Medium. 2021: URL https://towardsdatascience.com/web-scraping-with-beautiful-soup-selenium-or-scrapy-62c6f3545de7

3. Bhattacharya C. Web Scraping E-commerce sites using Selenium & Python. Analytics Vidhya. 2020. URL: https://medium.com/analytics-vidhya/web-scraping-e-commerce-sites-using-selenium-python-55fd980fe2fc

4. Online store ROZETKA: the official site of the most popular online hypermarket in Ukraine. URL: https://rozetka.com.ua/ua/

5. Khder MA. Web scraping or web crawling: State of art, techniques, approaches and application. International Journal of Advances in Soft Computing and its Applications. 2021; 13(3):144–68.

6. Thomas DM, Mathur S. Data Analysis by Web Scraping using Python. 2019. p. 450–4.

7. McHaney R. Tutorial: Text analytics for simulation with python. 2021. p. 68–82.

Рецензія/Peer review : 22.01.2022 р. Надрукована/Printed : 27.02.2022 р.