IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, May 2022
Impact Factor: 6.252

Data Extraction from E-commerce Website using Web Scrapping with Python and Selenium

Unnati N. Changole, Sandesh B. Harne, Prof. Ashwini P. Ghatol
Students, Department of Computer Science & Engineering
Sipna College of Engineering & Technology, Amravati, Maharashtra, India

Abstract: *The internet provides access to a vast array of information and data gathered by people. It will, however, include a vast array of disparate and poorly structured data, which will be difficult to obtain through physical means and tricky to use in mechanical operations. Procedures and numerous outfits have been established in the recent past to allow data collection and transformation into ordered information by B2C and B2B systems. Beginning with a basic introduction and a brief review of various online scraping applications and tools, this article will focus on various elements of web scraping. We also went through the process of web scraping, as well as the numerous types of web scraping techniques, before concluding with the benefits and drawbacks of web scraping, as well as a detailed discussion of the various sectors where it may be used. The possibilities for utilizing these data are vast, including Open Government Data, Big Data, Business Intelligence, aggregators and comparators, as well as the development of new apps and mash ups, to name a few.*

Keywords: Web scrapping, Data Extraction, Business Intelligence, Automation, Selenium, HTML Parsing

I. INTRODUCTION

The online world is currently massive in terms of websites with vast amounts of explanatory materials available in various designs such as text, graphical, audio video, and so on, which will focus on the inconsistency in fact repossession due to the insignificance of the fact user is seeing. Only a web browser can view the data that the websites display. They don't let you save a copy of the information for your own use. The only other alternative is to physically copy and paste the data displayed in the browser into our computer's hard drive, which is time-consuming. Web scraping is helpful in this case. Web scraping (also known as Screen Scraping, Web Data Extraction, and Web Harvesting, among other terms) is a method of extracting web data without having to copy it manually. It's a method for extracting useful information from websites' HTML and storing it in a central database or spreadsheet. This is accomplished by using the website's URL. Web scrapers use specifically written programs to carry out this task. It may be built in the usual way for a specific website or it can be easily arranged to operate with any website. A Web scraper's main purpose is to convert unstructured data into structured data and store it in databases.

HTTP programming, DOM parsing, and HTML parsers are a few Web scraping techniques. The information gathered is then utilized for retrieval and analysis. Web scraping is now used for various purposes, such as online pricing comparison, weather data monitoring, website change detection, Web mash-up, Web research, and Web data integration. Web scraping may also be associated with the duration of usage of a limited number of websites.

II. LITERATURE SURVEY

Approximately 70% of the information gathered on the internet comes through PDF documents, which are unstructured and difficult to work with. Furthermore, a web page is a structured format (containing of HTML code), but one that is not reusable. The "tyranny of PDF" is a phrase used to describe the large amount of information that is delivered yet constrained by this sort of architecture. Its ordered nature reproduces the potentials uncovered using scraping processes, ensuring the key capability of this document (HTML documents). Web scraping techniques and procedures rely on organization and HTML language assets. Scraping for tools and robots will be possible as a result, freeing people from the arduous, error-prone task of physical data retrieval. To summarize, these technologies provide data in a variety of forms for enhanced distribution and integration, including JSON, XML, CSV, XLS, and RSS. Data collection a handcrafted technique is

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, May 2022
Impact Factor: 6.252

accurately incompetent in searching, copying and pasting data in Spreadsheet for processing. This is a monotonous, annoying and frustrating technique.

III. The Necessity to Scrape Websites

Approximately 70% of the information gathered on the internet comes through PDF documents, which are unstructured and difficult to work with. Furthermore, a web page is an orderly format (containing of HTML code), but one that is not reusable. The "tyranny of PDF" refers to a large amount of information transmitted yet constrained by this sort of design. Its ordered nature reproduces the potentials uncovered using scraping processes, ensuring the essential capability of this document (HTML documents). Scraping techniques and tools rely on the organization as well as HTML language assets. This will make scraping for tools and robots possible, freeing people from the arduous and error-prone tasks of physical data extraction.

Methods

The system is used to create a website that uses online scraping and crawling tactics to search for product categories using an HTML DOM-based architecture.

A. Working Process

The web application is built using the procedures listed below:

1. Install Python libraries.
2. Get the URL using the request and selenium libraries and save it in a temporary variable
3. In a temp variable, parse the HTML and transform it to JSON format.
4. Remove the product label, price, specifications, and image.
5. Compare product costs.
6. Sort scraped data by price.

The first two processes are known as web crawling, while the third and fourth steps are known as web scraping.

B. Web Scrapping/Crawling Implementation using python

We chose Python as the scripting language for this project because it provides quick and strong libraries, as well as community support for web scraping and crawling. For different phases, we utilized the Web Selenium Driver python libraries.

Selenium:

Selenium is a unique testing framework that supports a wide range of browsers, including Google Chrome, Firefox, and Internet Explorer. It offers a variety of web application testing activities. Selenium is a web driver that can handle dynamic web pages, i.e., a page whose components can change without the page refreshing [15]. Although it was designed to provide web tests for web applications, it may also be used on sites that contain JavaScript. The Python Web Driver Selenium package has it.

Data Extraction through Selenium:

Web scraping is all about dealing with large volumes of data, and Python is one of the best languages for doing so since it has a low learning curve and a large library and framework ecosystem, including NumPy, CSV, Web driver, and others. Selenium and other Python-based web scraping technologies have advantages. Selenium is a website testing system that uses a web-driver package to take control of the browser and simulate real-world human behaviour. Selenium provides a simple way to extract data using Scrappy selectors to collect HTML code, which is useful because most websites are JavaScript rich.

Web scraping is a useful approach for obtaining large volumes of data from a variety of websites. After that, the data may be saved in any file or database. Online scraping is the process of extracting the HTML code from a web page. ; Selenium is employed to achieve the goal here. Selenium is a web browser automation tool and an automated testing tool in general. It may, however, be used for web scraping as well.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, May 2022
Impact Factor: 6.252

C. Website Module

I. User Interface

We created a user-friendly interface that is straightforward to utilize. By entering the required product name into this interface, the user may query. According to the information source, the query's results include product specifications, price, and name.

II. Business Logic

The application layer is in charge of getting data from the website. It parses data using the Beautiful Soup package and is built on custom Python scripts. The web application is in charge of engaging with users and providing them with the information they need. Figures 1 and 2 depict the working flow of the web application and data scraping. Following the discovery of instances, web scraping is used to retrieve data and metadata such as price, label, and product specification for price comparison.

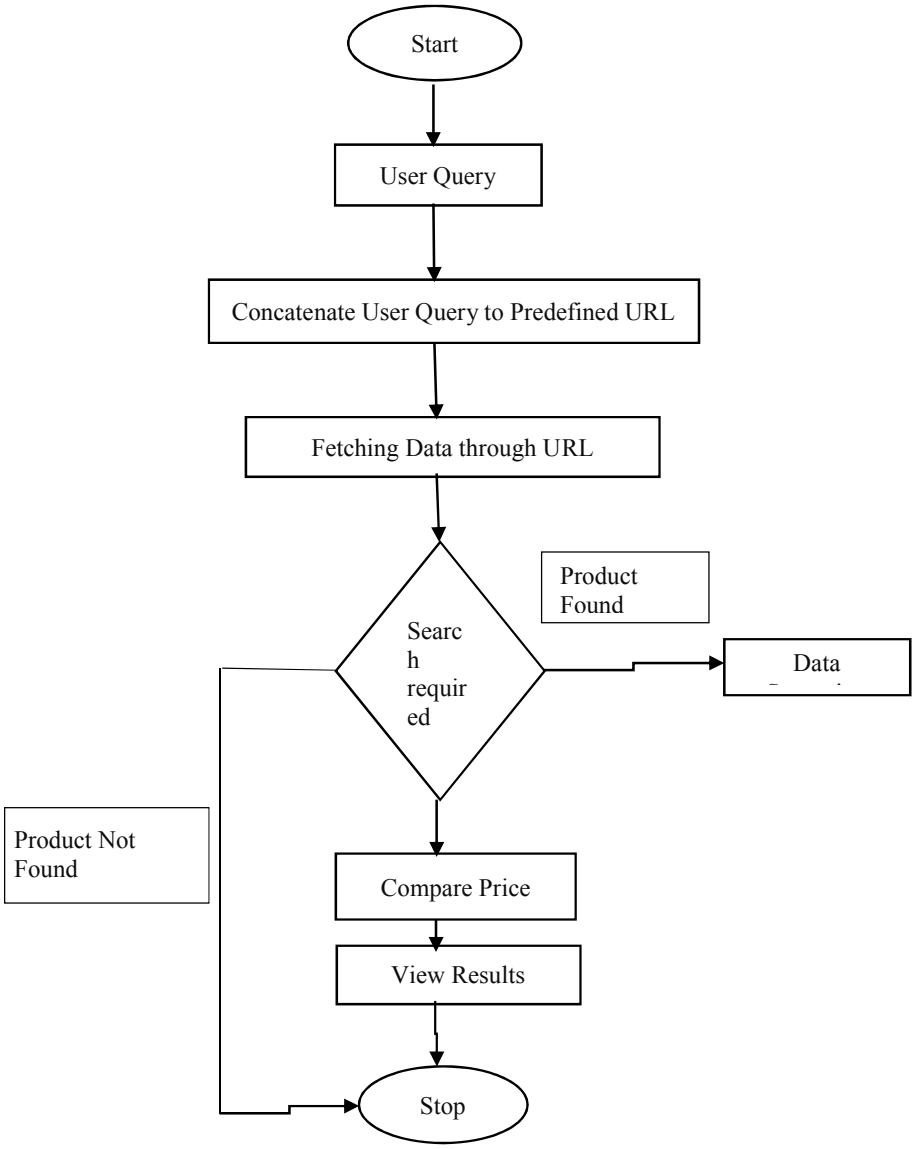

Fig : Data Scraping Flow

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, May 2022
Impact Factor: 6.252

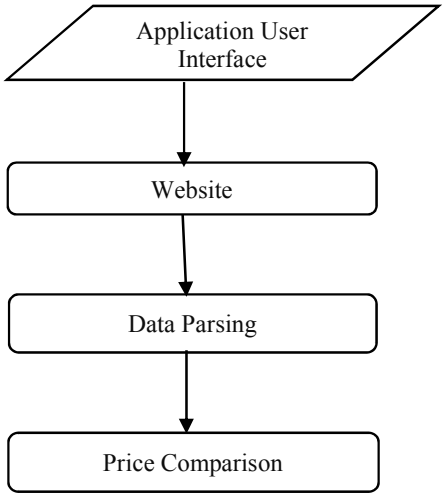

Fig :- System Overview

IV. RESULTS AND DISCUSSION

This research creates a PHP-based online application that allows users to view and select the lowest-priced 5 examples of a product from several web domains promoting that product. For product price comparison, users can optionally choose specific web domains that they feel to be the best seller in terms of quality and price. This application is timely because in today's digital world, various online businesses sell the same type and brand of product at varying costs. Due to the demands of work and the human race, people do not have enough time to visit several websites in order to obtain a product at a low price. Legality considerations must be considered before moving further with the implementation of this sort of online application. This service uses data from several online domains to compare costs of identical goods that are being offered to customers for purchase. The question of whether it is permissible to access their data arises, but in terms of legality, this program only accesses data that is on the web interface of e-commerce sites, and given that it is free for anybody who visits this site, it is evident that no infringement has occurred. Another essential consideration is the administrators' updating of web domains. The online domains that have been viewed are updated on a regular basis by their administrators. In the meanwhile, the technique that has been employed is to access a web page every time a user queries for a certain product. So that the most recent page is used for subsequent steps. This method improves our program by lowering memory requirements because a local database is not required to store data on a regular basis. Along with this, product availability errors are decreased, i.e., if we use a database and a person looks for a product that is out of stock on the web domain but is available in our database, the results are incorrect and create uncertainty.

Because a brief query like "Samsung J5" will yield mobiles as well as accessories related to this inquiry, the scraped information from each online site for a transaction started by a user may contain various things. However, our main concern is to show only relevant mobiles on our web application. To accomplish this, we use string matching, which means that if a user selects a category of mobiles, our system scrapes details of each product from a web page and performs string matching on product titles to identify required products from each web domain.

Applications of Web Scrapping:

- E-commerce
- Finance
- Data Science
- Social Media
- Sales

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, May 2022
Impact Factor: 6.252

V. CONCLUSION

Web scraping may be used to collect various sorts of data from websites for commercial or personal needs, and there are several methods for doing so. However, you must be aware of the load that your web scraper places on the page, since there may be consequences to careless web scraping. Consider running a script through the first 100 pages; this would be an aggressive scraper, and we would be putting an unreasonably large strain on the website servers, potentially disrupting their operation. Web scraping is a violation of certain websites terms and conditions, and the website is likely to take action against you in such cases

REFERENCES

[1]. Shakra Mehak, Rabia Zafar, Sharaz Aslam, Sohail Masood Bhatti "Exploiting Filtering approach with Web Scrapping for Smart Online Shopping" 2019 International Conference on Computing, Mathematics and Engineering Technologies – iCoMET 2019

[2]. Vidhi Singrodia, Anirban Mitra, Subrata Paul "A Review on Web Scrapping and its Applications" 2019 International Conference on Computer Communication and Informatics (ICCCI -2019), Jan. 23 – 25, 2019, Coimbatore, INDIA

[3]. Sarah Fatima , Shaik Luqmaan , Nuha Abdul Rasheed "Web Scraping with Python and Selenium" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 23, Issue 3, Ser. II (May – June 2021), PP 01-05

[4]. K Usha Manjari, Syed Rousha, Dasi Sumanth, Dr. J Sirisha Devi "Extractive Text Summarization from Web pages using Selenium and TF-IDF algorithm" Proceedings of the Fourth International Conference on Trends in Electronics and Informatics (ICOEI 2020) IEEE Xplore Part Number: CFP20J32-ART; ISBN: 978-1-7281-5518-0

[5]. Sanjay Kumar Malik1 , SAM Rizvi2 "Information Extraction using Web Usage Mining, Web Scrapping and Semantic Annotation" 2011 International Conference on Computational Intelligence and Communication Systems