Marketplace Data Product Analysis Using Web Scraping

L. Ahmad Syamsul Irfan Akbar [1], Gusti Bagus Indra Permana [1], Cipta Ramadhani [1]
[1]Department of Electrical Engineering Faculty of Engineering University of Mataram, Indonesia

ARTICLE INFO

Article history :

Received August 29 ,2024
Revised August 31, 2024
Accepted August 31, 2024

Keywords:

Web Scraping;
Beautiful Soup;
Selenium;

ABSTRACT

The development of the internet in providing fast information has a positive impact on business people to offer or sell products via the internet. One of the benefits obtained is that we can find out or search for various products needed via the internet. This has encouraged the emergence of many online stores in Indonesia, so an application is needed that can help users search and collect product data on the internet. The purpose of this study is to create a product data analysis from the results of web scraping from the market place website page (Tokopedia) using the BeautifulSoup4 (Bs4) method. Website design is done using the streamlit module from python. The results of this study are data analysis from the results of scraping the Tokopedia website page using the Beautiful Soup method with the following specified categories: price range, highest and lowest ratings, highest and lowest sales, highest and lowest prices, percentage of locations, percentage of seller stores and product recommendations.

Corresponding Author:

L. Ahmad S.Irfan Akbar, Jurusan Teknik Elektro, Fakultas Teknik,Universitas Mataram
Email: irfan@unram.ac.id

1. INTRODUCTION

In the current digital age, information technology and the internet have become essential components in the daily lives of teenagers and adults, significantly influencing their behavior and lifestyle choices [1]. According to data from the Indonesian Internet Service Providers Association (APJII), in 2020, the number of internet users in Indonesia reached 196.7 million people, accounting for approximately 73.7% of the total population, This widespread internet usage highlights the profound impact of digital connectivity on society [2]. The convenience offered by internet services has greatly benefited businesses, enabling them to efficiently offer and sell products online. Consumers can easily search for and purchase a wide variety of products from the comfort of their homes, leading to a surge in the number of online stores in Indonesia. Marketplaces, in particular, have become popular platforms for running online stores, providing a centralized space for buyers and sellers to interact [3].

In recent years, the integration of business intelligence with e-commerce has revolutionized traditional retail practices, significantly contributing to economic growth by creating new opportunities for entrepreneurs and small businesses. The accessibility of online marketplaces has democratized commerce, allowing even the smallest vendors to reach a broad audience. This digital transformation has fostered innovative business models and marketing strategies, embedding the internet deeply into modern life. Studies have shown that intelligent technologies, such as AI and big data, are critical in enhancing the efficiency and effectiveness of e-commerce platforms, leading to sustainable development and competitive advantages in the digital economy [4].

Moreover, the rise of e-commerce has not only transformed traditional retail practices but also spurred economic growth by creating new opportunities for entrepreneurs and small businesses. The accessibility of online marketplaces has democratized commerce, allowing even the smallest vendors to reach a broad audience. This digital transformation has paved the way for innovative business models and marketing strategies, further embedding the internet into the fabric of modern life [5].

The Internet provides a variety of information in various formats such as numbers, text, images, audio, and video on web pages. This diversity often causes difficulties in retrieving relevant information because it does not always match the user's search needs [6]. Web scraping[7] is a technique for obtaining information from websites automatically without having to copy it manually. This technique focuses on data retrieval and extraction, so that the information obtained is more focused and easier to search. Web scraping can be a tool for online entrepreneurs in collecting product data in a short time.

This paper uses the Beautiful Soup methods, web scraping allows product data retrieval from various marketplaces efficiently. This technique makes it easy for entrepreneurs to access and analyze product information quickly and accurately, which can ultimately improve their business strategy. The resulting web scraping application is able to provide informative information to users by providing output in the form of several categories for a searched item, such as: price range, rating order, price sorting, percentage of seller locations and others.

2. METHOD

Figure 1 shows the overall design of the web scraping system. Web scraping uses frameworks and libraries such as Selenium, Beautifulsoup4 and Streamlit. The input used in this web scraping application is the URL of a product searched in the search column of a marketplace. The URL is then processed by the web application to obtain product output with various categories.

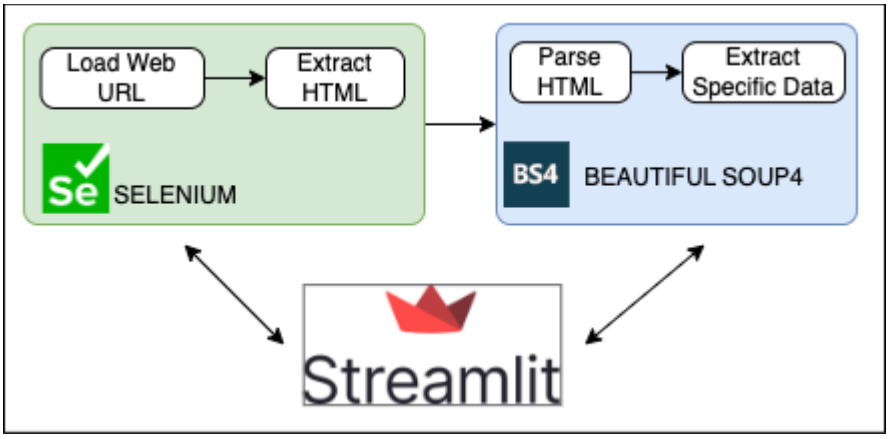

Figure 1. Overview of the system design.

2.1. Selenium

Selenium WebDriver is a widely-used tool for building web scraping applications[8] [9]. It plays a critical role in managing and controlling web pages within a browser, facilitating the collection of data from these pages[10]. Selenium has the ability to control and navigate web pages through a browser [11], so in this study selenium is used to collect a number of data based on certain objects in the market place. After the web page is fully loaded in the browser controlled by Selenium, selenium extracts or takes all HTML content from the page. This HTML is the source code of the web page that contains all the elements and structures that display content on the website.

2.2. Beautiful Soup

Beautiful Soup is a Python library specifically designed for extracting HTML and XML files. By using the extracted code, Beautiful Soup allows developers to parse different files more easily and efficiently. Beautiful Soup takes the HTML extracted by Selenium and parses it into a data structure that is easier to manipulate. Beautiful Soup creates a "parse tree" of the HTML, which is used to navigate and search for specific elements within the HTML. Once the HTML is parsed, Beautiful Soup is used to search for specific HTML elements based on tags, attributes, IDs, or classes. [12]

2.3. OpenPyXl

OpenPyXL is a Python library created to read and write Excel files, especially the .xlsx format introduced since Excel 2010. OpenPyXL can be used to process Excel files automatically without requiring Excel itself to be installed. The library also integrates well with NumPy and Pandas, making it a very flexible tool for generating Excel reports directly from data frames. [13]

2.4. Streamlit

Streamlit is a Python library used to create interactive web applications quickly and easily. Streamlit is designed for visualizing data, creating dashboards, or developing machine learning applications. Streamlit

can be used to transform Python scripts into functional web applications that will generate user interfaces. Streamlit also supports various integrations with other libraries such as Pandas, Matplotlib, and Plotly[14]

2.5. Proses Scraping Data Process

The scraping process begins by identifying the URL of the search results page on the marketplace. This URL will be used as input in the web scraping application. After the URL input is given and the program is run, Selenium will automatically open the Chrome browser to load the target web page. After the web page is loaded, the HTML content of the page will be extracted, and BeautifulSoup4 will parse the website elements. This parsing process allows BeautifulSoup4 to extract specific information from HTML elements, such as price, product name, number of products sold, product rating, location, name of the store selling the product, and a link to the product page. The data that has been successfully extracted will then be further processed using Pandas. The final result of this scraping will be displayed in the form of a table containing all the important information that has been collected. The web scraping process with beautifulsoup and selenium is shown in figure 2.

Algorithm 1 Search Page Scraping

```
Get user input: url,  number_of_pages (x)
if "Scrape" button is clicked then
    if url is not empty then
        Initialize Chrome WebDriver with options
        for each page in range(number_of_pages) do
            Load the page with the given url
            Wait for the page to load elements
            Parse page content with BeautifulSoup
            Extract data for each product on the page:
                Add product name to list_nama_produk
                Add product price to list_harga
                Add product rating to list_rating
                Add product sales to list_terjual
                Add product location to list_lokasi
        Create a DataFrame from the extracted data
        Convert DataFrame to Excel file format
        Encode Excel file to base64 for download
        Display download link for Excel file
```

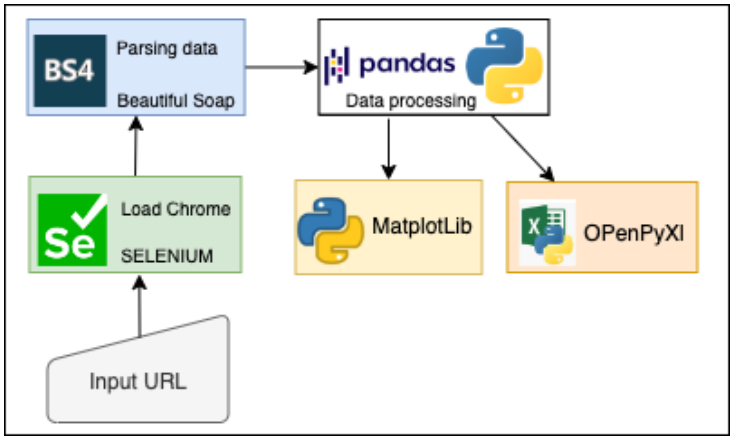

Figure 2. Web Scraping Process

3. RESULTS AND DISCUSSION

3.1. Analisa Struktur HTML

In Figure 3 is the stage of identification and analysis of HTML elements from the search page with the given product name keyword. BeautifulSoup4 will read the HTML element code from the marketplace web page and extract information from the desired element.

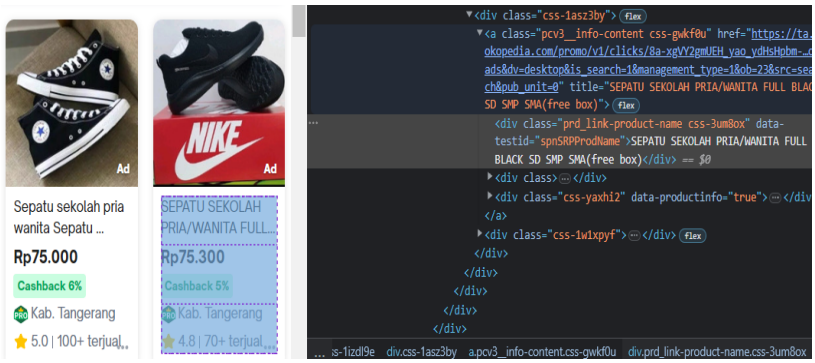

Figure3 . HTML Structure Analysis

The url results and element identification from the previous stage are added to the program, Beautifulsoup4 will perform extraction by parsing data on the specified HTML elements. Pandas will process data from the results of the data parsing carried out by Bs4 and will be displayed in table form, as shown in Figure 4.

	A	B	C	D	E	F	G
1	Nama Produk	Harga	Terjual	Rating	Toko	Lokasi	Link
2	Kanky Story Honjo - Sepatu	298.800	3.000	5	Sepatu Kanky	Kab. Bandung	https://ta.tokop
3	Sepatu Converse All Star Peti	75.000	30	5	briantara_shoes	Kab. Tangerang	https://ta.tokop
4	Spotec Sepatu Student Activ	252.900	80	5	Spotec Official S	Jakarta Utara	https://ta.tokop
5	TOMKINS Scool Wednesday	249.000	90	5	TOMKINS Offici	Bandung	https://ta.tokop
6	Sepatu Casual Kasogi Julio P	220.050	80	4.90	Kasogishoes	Surabaya	https://ta.tokop
7	Kanky Story Honjo - Sepatu	298.800	750	5	Sepatu Kanky	Kab. Bandung	https://www.tok
8	Sepatu Sekolah Warrior PX Sl	78.000	3.000	4.90	PASShoes	Kab. Bandung	https://www.tok
9	Dr. Kevin Sepatu Sneakers Sp	119.900	100	4.90	Dr Kevin Shoes	Bekasi	https://www.tok
10	Kanky Story Honjo - Sepatu	298.800	3.000	5	Sepatu Kanky	Kab. Bandung	https://www.tok
11	Kanky Hiro Musashi - Sepatu	259.920	500	4.90	Sepatu Kanky	Kab. Bandung	https://www.tok
12	Kanky Back To School Pria 3S	196.920	250	5	Sepatu Kanky	Kab. Bandung	https://www.tok
13	Leedoo Sepatu Sneakers Pria	99.000	3.000	4.70	Leedoo	Tangerang	https://www.tok
14	Sepatu Warior/Warrior Sparta	109.900	1.000	4.90	Mitra Sepatuu	Kab. Tangerang	https://www.tok
15	Sepatu Converse Allstar Alsta	69.999	500	4.90	project_sneakers	Kab. Tangerang	https://www.tok
16	SEPATU ANDO HITAM POLOS	109.800	100	4.90	Toko Abadi Shoe	Tangerang	https://www.tok
17	Homyped Anak Sepatu Seko	99.900	100	4.90	Homyped Offici	Tangerang	https://ta.tokop
18	sepatu sekolah Warior tinggi	150.000	1	5	Ds_Storee21	Jakarta Timur	https://ta.tokop
19	TOMKINS Scool Sanditon - H	249.000	100	5	TOMKINS Offici	Bandung	https://ta.tokop

Figure 4 Web Scraping Result

3.2. Antar Muka Web Scraping

Figure 5 is a web scraping interface page. On this page, the user will be asked to enter the URL address of the marketplace website page to be scraped, then the user enters the number of pages to be scraped. After that, the user can press the "Scrape" button and the system will run the scraping process. The website will display the scraping results, then the user can download the scraping results in excel format by pressing the "Download Excel" button. Some of the information displayed on the web page can be shown in Figure 6.

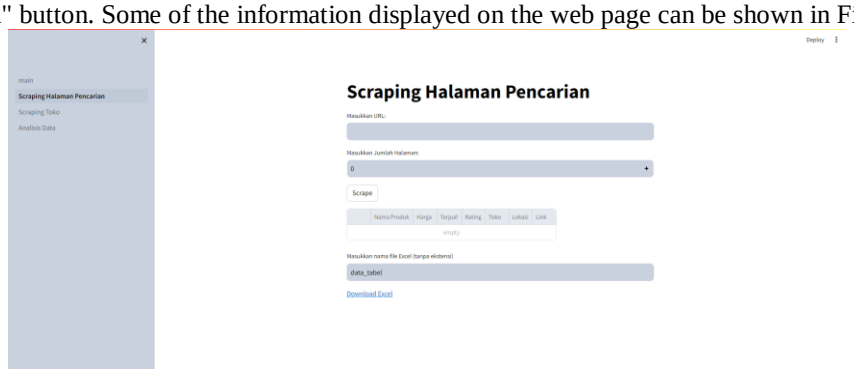

Figure 5. Web Scraping Interface

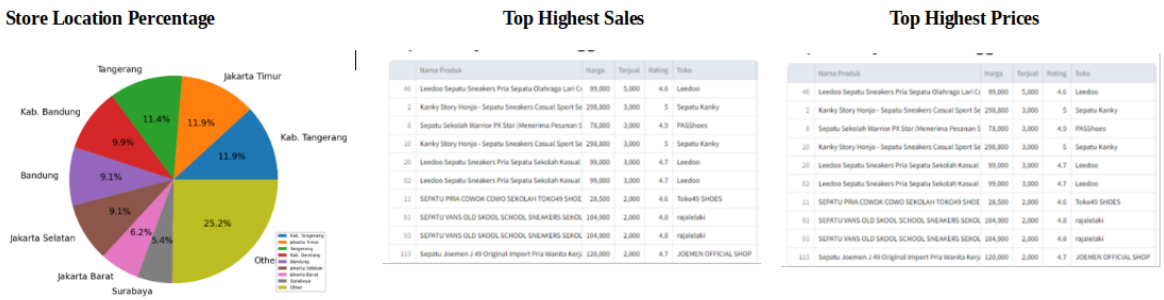

Figure 6. Web Scraping Data Analysis Result

4. CONCLUSION

The data output from the web scraping application for the market place website is grouped based on categories including: the price range of a product, the highest and lowest rating data sequence, the highest and lowest price data sequence, the highest and lowest sales data sorting, the percentage of seller locations, the percentage of seller stores and product recommendations to make it easier for users to choose the best product.

REFERENCES

[1] S. C. Joshi and G. Rose, "Information Technology, Internet Use, and Adolescent Cognitive Development," 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), Bengaluru, India, 2018, pp. 22-28, doi: 10.1109/CSITSS.2018.8768780.

[2] Asosiasi Penyelenggara Jasa Internet Indonesia. (2023, April 17). Survei APJII:. Retrieved from https://apjii.or.id/berita/d/survei-apjii-pengguna-internet-di-indonesia-tembus-215-juta-orang.

[3] Kemp,S. (2023, January 26). Digital 2023: Global Overview Report. DataReportal. Retrieved from https://datareportal.com/reports/digital-2023-global-overview-report.

[4] C. -L. Pan, X. Bai, F. Li, D. Zhang, H. Chen and Q. Lai, "How Business Intelligence Enables E-commerce: Breaking the Traditional E-commerce Mode and Driving the Transformation of Digital Economy," 2021 2nd International Conference on E-Commerce and Internet Technology (ECIT), Hangzhou, China, 2021, pp. 26-30, doi: 10.1109/ECIT52743.2021.00013.

[5] Costa J, Castro R. SMEs Must Go Online—E-Commerce as an Escape Hatch for Resilience and Survivability. Journal of Theoretical and Applied Electronic Commerce Research. 2021; 16(7):3043-3062. https://doi.org/10.3390/jtaer16070166

[6] S. Mehak, R. Zafar, S. Aslam and S. M. Bhatti, "Exploiting Filtering approach with Web Scrapping for Smart Online Shopping : Penny Wise: A wise Tool for Online Shopping," 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, Pakistan, 2019, pp. 1-5, doi: 10.1109/ICOMET.2019.8673399.

[7] Simon Munzert; Christian Rubba; Peter Meissner; Dominic Nyhuis, "Scraping the Web," in Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining , Wiley, 2015, pp.219-294, doi: 10.1002/9781118834732.ch9. keywords: {Voting;IEEE Sections;Task analysis;Standards;Data collection;Automation;XML}

[8] Selenium WebDriver. Available online: https://www.selenium.dev/documentation/webdriver/ (accessed on 2 August 2024).

[9] Wendt, H.; Henriksson, M. Building a Selenium-Based Data Collection Tool. Bachelor's Thesis, 16 ECTS, Information Technology, Linköping University, Linköping, Sweden, May 2020.

[10] Gojare, S.; Joshi, R.; Gaigaware, D. Analysis and design of selenium WebDriver automation testing framework. Procedia Comput. Sci. 2015, 50, 341–346.

[11] Naing I, Aung ST, Wai KH, Funabiki N. A Reference Paper Collection System Using Web Scraping. Electronics. 2024; 13(14):2700. https://doi.org/10.3390/electronics13142700

[12] Wendt, H., & Henriksson, M. (2020). *Building a Selenium-based data collection tool* (Bachelor's thesis, 16 ECTS, Information Technology). Linköping University. LIU-IDA/LITH-EX-G--20/066--SE.

[13] PyXLL. (n.d.). *Tools for Working with Excel and Python*. Retrieved July 29, 2024, from https://www.pyxll.com/blog/tools-for-working-with-excel-and-python/

[14] Tyler Richards, Getting Started with Streamlit for Data Science: Create and deploy Streamlit web applications from scratch in Python , Packt Publishing, 2021.