International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Web Scraping Based Site for Product Analysis

Ms. C. E. Rajaprabha[1], B. Noufia Firthous[2], C. Suresh Krishna[3], S. Nisha[4], R. Srikrishnan[5]

[1]ME.(Ph.D), Department of Computer Science and Engineering, Hindusthan Institute Of Techonology Coimbatore, India.
[2,3,4,5]Department of Computer Science and Engineering, Hindusthan Institute Of Technology, Coimbatore, India.

Abstract:
Gaining a competitive edge in the modern e-commerce world necessitates a thorough comprehension of competition tactics, customer habits, and market dynamics. This project presents an advanced way for methodically extracting, analyzing, and deriving useful insights from e-commerce platforms by using web scraping techniques. Compliant with IEEE standards, the goal of this project is to provide enterprises with strategic information based on extensive product data. The process includes creating a sophisticated web scraping program that can quickly and effectively explore target e- commerce websites. It specify ethical scraping procedures and legal issues, to the letter, it pulls relevant product facts, such as pricing, descriptions, availability, and customer reviews. The research concludes in a user interface that allows for smooth interaction with the studied data.

Keywords: Price Comparison, Product Analysis, Review Analysis, Low Cost Product.

1. INTRODUCTION

The scope of this project is creating a web scraping based site to compare products on various e-commerce platforms. The focus is on selecting target websites, developing a robust scraping module for efficient data extraction, and capturing essential product details. The system will employ a structured data collection and data organizing techniques for product analysis. Advanced analytic tools will provide valuable insights, supported by user-friendly visualization features. Automation for scheduled updates, security measures, and considerations for scalability and external system integration are key aspects. Comprehensive documentation, rigorous testing, and user training will ensure a reliable and effective web scraping-based site for product analysis.

Web scraping based site for product analysis aims to gather and contrast product details from many e-commerce sites, allows people to make instantaneous comparison. It also monitors competitor prices, analyse product catalogue and customer reviews, identify market trends, assess brand visibility, and uncover cross- selling opportunities. It also involves monitoring user engagement metrics, and staying informed about new product launches. These objectives collectively provide valuable insights for strategic decision-making and maintaining competitiveness in the e-commerce market.

2. LITERATURE REVIEW

A literature survey reveals significant advancements in the web scraping methodologies.

Swetha Srinivasan (2021) proposed that in the current data-driven world, it becomes increasingly essential that big data techniques are applied and analyzed for organizational growth. More specifically, with the large availability of data on the Web, whether from social media, websites, online portals, or platforms, to name but a few, it is important for organizations to know how to mine that data in order to extract useful knowledge.

Tyagi and Sharma (2022) proposed web scraping for e-commerce website. web scraping is basically an interactive method for website and some other online sources to browse for and access data. To delete a replica of the information and save it in an external archive for review, it uses software engineering technology and custom software programming to extract data or any other content of on-line sources. Web scraping is often called automatic data gathering, database discovery, database crawling, or content management mining.

Arman Shaikh (2023) proposed e-commerce price comparison for webscrapingPrice comparison sites are designed to compare the price of goods and services from a range of providers, which will help consumers in making decision to chooseproducts that will save their money through online. Considering the customers' busy lifestyle especially those who are living in the city area, most of the consumers prefer to buy their needs through the internet because it saves their time.

3. METHODOLOGY

Web scraping is a data extraction technique employed to gather information from websites. The process typically begins by fetching the HTML content of a web pageusing tools like the 'requests' library in Python. Once the HTML is obtained, the next step involves parsing it to extract the desired information. This parsing is oftenfacilitated by libraries such as Beautiful Soup, lxml, or html5lib, which provide functionalities for navigating and manipulating the HTML structure. For instance, using Beautiful Soup in Python, you can easily navigate through theHTML and extract specific elements. The library allows you to target elements using CSS selectors or XPath, providing a flexible means to pinpoint the data of interest.

Responsible web scraping involves employing techniques such as respectful crawling, which includes spacing out requests and adhering to the website's guidelines. This ensures that the scraping process doesn't overload the website's server or trigger protective measures like IP blocking. Additionally, it's essential to consider the legality of web scraping for a particular website, as unauthorized scraping may lead to legal consequences.

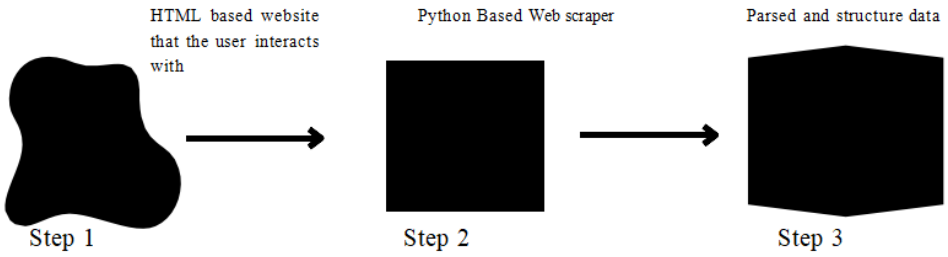

Fig. 1. Basic Pipeline of Web scraping

Required Skills for a Web Scraping Programming Languages.

Python: Python is one of the most popular programming languages for web scraping. Libraries like Beautiful Soup and requests make it easy to fetch and parse HTML content. Other libraries such as Scrapy provide more advanced capabilities for building web scrapers.

HTML and CSS Understanding

HTML: Knowing how to read and understand HTML is crucial for web scraping. You need to identify the HTML elements containing the data you want to extract.

CSS: Understanding CSS selectors is important for targeting specific HTML elements efficiently during web scraping.

XPath Knowledge: XPath is a powerful language for navigating XML and HTML documents. Knowing how to use XPath expressions is beneficial, especially when dealing with complex HTML structures.

Web Development Basics: Having a basic understanding of how websites are built and how web servers work is helpful. Knowledge of HTTP, status codes, and response headers is valuable for handling web requests effectively.

Regular Expressions (Regex): Regular expressions are handy for pattern matching and extracting specific information from text. They can be useful when dealing with complex or varied data formats.

Significance Of Web Scraping

Market Research and Competitive Analysis :

Businesses can use web scraping to gather information about competitors, market trends, and consumer sentiments. Analysing data from various sources helps in making informed decisions and staying competitive.

Price Monitoring and Product Tracking :

E-commerce businesses can utilize web scraping to monitor prices of products across different websites. This information is valuable for adjusting pricing strategies and staying competitive in the market.

Lead Generation :

Web scraping is commonly used for extracting contact information from websites, helping businesses build targeted email lists and generate leads for sales and marketing purposes.

Financial Data Analysis :

In the finance industry, web scraping is employed to collect and analyze financial data, stock prices, economic indicators, and news. This information is crucial for making investment decisions and risk assessments

Content Aggregation :

News aggregators, content curators, and similar platforms use web scraping to collect and organize information from various sources, providing users with a centralized location for news and updates.

4. SYSTEM ANALYSIS

4.1 OVERALL DESCRIPTION

- The proposed project involves the development of an advanced system designed for in-depth analysis of e-commerce products through web scraping techniques.
- The primary goal is to extract, organize, and analyze crucial product information from various e-commerce websites, offering businesses valuable insights for informed decision-making.

- The system's core functionalities include the selection of target e-commerce platforms, the implementation of a robust web scraping module for efficient data extraction, and the collection of key product details such as prices, descriptions, customer reviews, and availability.
- The extracted data will be stored in a structured database, incorporating data cleaning and preprocessing techniques to ensure accuracy and reliability.
- Advanced analytical methods, including statistical analysis, trend identification, and sentiment analysis, will be applied to derive meaningful insights from the collected data.
- The system will feature a user-friendly interface with visualization tools, enabling businesses to interact seamlessly with the analyzed information.
- Automation will be integrated for scheduled updates, ensuring that the system remains current with the dynamic nature of e-commerce data.
- Considerations for scalability and compatibility with external systems will be integral, allowing the system to adapt to future changes.
- Comprehensive documentation, rigorous testing, and user training will be prioritized to ensure the system's reliability, ease of use, and effectiveness in delivering actionable insights for strategic decision-making in the competitive e-commerce landscape.

4.2 EXISTING SYSTEM

Google Shopping: This platform operates on a pay-per-click model, which can result in considerable expenses, particularly for smaller businesses with limited budgets.

Crowd Sourcing: This sourcing approach involves gathering data and services, including ideas, from a sizable, open, and frequently rapidly changing community of internet users.

On demand quoting: This method entails asking the specific websites to supply their information.

Adding Data Manually: This method entails entering the different data by hand into the rows or columns.

DISADVANTAGES :

- Google Shopping, while offering significant advantages for both merchants and shoppers, is not without its drawbacks. One notable disadvantage lies in the costs associated with participation for merchants. The platform operates on a pay-per- click model, which can result in considerable expenses, particularly for smaller businesses with limited budgets. Additionally, the competitive nature of Google Shopping, with merchants bidding for ad placements, can lead to increased competition and higher bid amounts, posing challenges for smaller enterprises.
- Managing campaigns on Google Shopping can be complex, requiring a nuanced understanding of the platform and effective campaign management skills. This learning curve may be a hurdle for merchants new to online advertising. Moreover, the platform's emphasis on paid listings limits organic exposure for products, necessitating investment in advertising for increased visibility.
- Data Accuracy and Reliability: Because online scraping depends on the structure of the target websites, modifications to the site's design or content may cause errors in the data or cause the scraping process to malfunction, producing data that is not trustworthy.
- Maintenance Difficulties: Because e-commerce websites change their layouts and content all the time, the scraping code has to be updated often to stay efficient. This requires ongoing upkeep and observation.

- Dependency on Third-Party Websites: The stability and accessibility of the intended e-commerce websites determine how reliable the product comparison model is. The efficiency of the scraping procedure is directly impacted by any outages or modifications to the source websites.

4.2 PROPOSED SYSTEM

This model gathers and arranges information from various online stores, including prices, features, reviews, and specs. In order to give customers a thorough and unbiased evaluation of products in the e-commerce industry With the use of this product comparison model, buyers can choose products with greater knowledge. The combined data makes it easier to compare products side by side and gives users the flexibility to assess features, costs, availability, and user feedback from several platforms. This improves the overall buying experience by giving customers the ability to choose the best value, quality, and fit for their preferences.

ADVANTAGES:

- Price Comparison
- Low Cost Product
- Content Aggregation
- Financial Data Analysis
- Market Research and Competitive analysis
- Lead Generation
- Business Insights
- Review Insights
- Price Comparison

4.3 SYSTEM ARCHITECTURE

The system architecture comprises a web scraping module that retrieves the target site's response for the ethical extraction of diverse product data. The HTML parser extracts necessary data from the page response and the data undergoes preprocessing before being sent to the user. The analytical engine applies statistical methods for insights, and the user interface integrates visualization tools. An automation module enables scheduled updates, a security layer addresses ethical considerations, and the system is designed for scalability and integration. Comprehensive documentation, rigorous testing,and user training ensure a robust and efficient system.

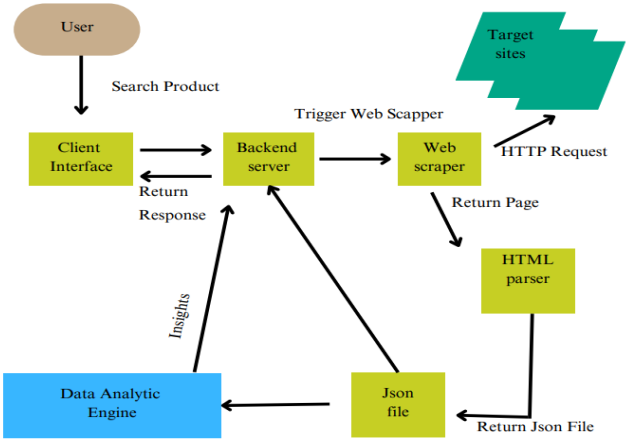

Fig. 2. System Architecture

5. MODULES

5.1 WEB SCRAPPING MODULE

5.1.1 Getting HTTP response

Data extraction is an important part of web scraping. An HTTP request is sent tothe target sites to get a response. The requests library in Python is used for making HTTP requests to a specific URL and returns the response. It provides inbuilt functionalities for managing both the request and response. The 'requests.get()' python function takes an e-commerce site's URL as its first argument and returns a response object. The Response object contains all the information that was returnedby the server in response to your request. This includes the status code, the headers, and the body of the response. We can access the status code of the response using the ".status_code property:response.status_code". You can access the response ofthe headers using the ".headers property: response.header" and you can access the body of the response using the ".text property: response.text".

5.1.2 Extraction of data

The HTML structure is analyzed and the necessary tags are identified. To extractour required content, the BeautifulSoup library is used. It is a Python library that is commonly used for Extraction of HTML Tags. It is an HTML parser that builds a parse tree. It allows you to interact with HTML in a similar way to how youinteract with a web page using developer the HTML content that is scraped earlier, as its input. The second argument, "html.parser", makes sure that you use the appropriate parser for HTML content.

5.2 DATA ANALYSIS

In the realm of web scraping, data analysis assumes a pivotal role, serving as the linchpin for extracting meaningful insights from the harvested information. The practice extends to recognizing patterns, applying sentiment analysis to textual data like customer reviews, and dissecting pricing dynamics.

5.2.1 PANDAS

Pandas allows for importing and exporting tabular data in various formats, such as CSV or JSON files. This is because pandas are used in conjunction with otherlibraries that are used for data science. It is built on the top of the NumPy library which means that a lot of structures of NumPy are used or replicated in Pandas. The data produced by Pandas are often used as input for plotting functions of Matplotlib, statistical analysis in SciPy. Pandas generally provide two data structures for manipulating data. They are Series and Dataframes.

SERIES

A Pandas Series is a one-dimensional labeled array capable of holding data of any type (integer, string, float, python objects, etc.). The axis labels are collectively calledindexes. Pandas Series is nothing but a column in an Excel sheet. Labels need not be unique but must be a hashable type. The object supports both integer and label-based indexing and provides a host of methods for performing operations involving the index.

Matplotlib

Data Analysis and Visualization

Once you have extracted relevant data, you might want to perform some analysis or create visualizations to better understand the patterns or trends within the data. Matplotlib can be used

for this purpose.

Integration with Pandas

If you're working with tabular data extracted during web scraping, you might use Pandas to organize the data in a DataFrame and then use Matplotlib to create visualizations.

5.3 SENTIMENT ANALYSIS

Sentiment Analysis is the way of computationally determining whether a piece of writing is positive, negative, or neutral. It's also referred to as information extraction because it involves assessing a presenter's point of view or mindset. Opinion Mining is another name for sentiment analysis.

5.3.1 VADER

VADER has the advantage of assessing the sentiment of any given text without the need for previous training as we might have to for Machine Learning models. The result generated by VADER is a dictionary of 4 keys neg, neu, pos and compound: neg, neu, and pos meaning negative, neutral, and positive respectively. Their sum should be equal to 1 or close to it with float operation.

5.3.2 VADER Sentiment Analysis

Apply VADER sentiment analysis to assess the sentiment of the extracted text. The Sentiment Intensity Analyzer class from the nltk.sentiment module can be used forthis purpose. The sentiment_scores dictionary contains values for "neg" (negative), "neu" (neutral), "pos" (positive), and "compound" (overall compound score).

5.3.3 WORD CLOUD

Word clouds are indeed powerful tools for analyzing text data, offering a visually intuitive way to identify recurring themes and prominent keywords within a body of text. By generating a word cloud, we can quickly discern which words are most prevalent, thereby shedding light on the central ideas or subjects discussed in the text. This analysis enables us to uncover trends, patterns, and potentially even sentiments expressed throughout the text.

In addition to revealing the most frequently occurring words, word clouds allow for the identification of both overarching topics and more specific subtopics within the text. By observing the relative sizes of the words in the cloud, we can gauge the emphasis placed on certain concepts or terms, providing valuable insights into the content's focus and significance.

Moreover, word clouds serve as effective tools for summarizing and condensing large volumes of text into easily digestible visual representations. By presenting the most pertinent words in a visually striking format, they facilitate quick comprehension and interpretation, making complex textual data more accessible and understandable.

6. SYSTEM SOFTWARE AND IMPLEMENTATION

6.1 FUNCTIONAL REQUIREMENTS

The functional requirements of a web scraping tool encompass its capability to extract structured or unstructured data from web pages, navigate the HTML Document Object Model (DOM) for precise content identification, and handle dynamic elements using techniques such as JavaScript rendering. The tool should support data transformation and cleaning, utilizing regular expressions for advanced pattern matching. It should incorporate mechanisms for concurrencyand parallel processing, ensuring efficiency in scraping multiple pages simultaneously, while also featuring robust error handling for scenarios like network errors or changes in page structure. The ability to manage authentication,

handle sessions, and integrate with proxy servers for IP rotation is essential for accessing restricted or personalized content.

6.2 NON-FUNCTIONAL REQUIREMENTS

The non-functional requirements of a web scraping tool encompass aspects beyond its specific functionalities, emphasizing factors that contribute to its overall performance, reliability, and user experience. These include considerations such as performance efficiency, where the tool should operate with minimal latency and resource usage, ensuring timely and responsive data extraction. Reliability is crucial, requiring the tool to maintain consistent functionality under varying conditions, handle errors gracefully, and provide accurate results. Scalability is another non-functional requirement, necessitating the tool's ability to handle an increasing workload and adapt to changing data volumes.

7. DEPLOYING MODEL

Deploying a web scraping model involves making the model accessible and operational in a production environment, allowing it to handle real-world web scraping tasks. This process typically includes several key steps. First, the machine learning model used for web scraping should be trained on relevant data to ensure its effectiveness in making predictions or classifications. Once trained, the model needs to be integrated into a deployment framework, such as Flask or FastAPI, which allows it to be hosted on a web server.

Deploy to a production environment, ensuring that it's accessible to clients and can handle the expected workload. Consider scalability requirements and design your deployment architecture to accommodate increases in traffic or data volume over time. Once trained, it's essential to evaluate the performance of your model using validation data or cross-validation techniques.

8. CONCLUSION AND FUTURE ENHANCEMENT

8.1 CONCLUSION

Our Site combines product information from several platforms, the web scraping- based product comparison model for e-commerce sites expedites the decision- making process for customers. It gives consumers the power to make knowledgeable decisions and gives companies insightful data for tactical decision-making. Despite its benefits, issues like data integrity and upkeep require attention to be effective over time. All things considered, it's a dynamic tool that improves the purchasing experience for customers and companies in the cutthroat world of online commerce. Users of this website will be able to compare costs on a variety of e-commerce shopping websites in order to choose which website offers the best combination of low cost and a good deal on the product they are interested in purchasing. The purchasers are going to unquestionably appreciate the time and effort that this saves them. In the end, this will help buyers shop online by bringing together tactics, the greatest offers and deals from all of the biggest online retailers, and by providing customers with an easier way to shop online.

8.2 FUTURE ENHANCEMENT

- Enhance the scraping model to support additional e-commerce websites and extract more detailed product information.

- Implement advanced features such as price tracking, user reviews analysis, and recommendation systems.
- Optimize the Flask application for performance and scalability to accommodate a larger user base.
- Integrate user authentication and personalized user experiences for registered users.

REFERENCES

1. Hilem, Judith \“Web scraping for food reasearch.\” British Food Journal (2019)
2. Pedro, Zayani Dabbabi, Miruna-Mihaela Mironescu, Olivier Thonnard, Alysson Bessani, Frances Buontempo, and Ilir Gashi. \"Detecting Malicious Web Scraping Activity: a Study with Diverse Detectors.\" In 2018 IEEE 23rd Pacific Rim International Symposium on Dependable Computing (PRDC), pp. 269-278. IEEE, 2018.
3. Bruni, Renato, and Gianpiero Bianchi. \"Website categorization: A formal approach and robustness analysis in the case of e-commerce detection.\" Expert Systems with Applications 142 - 2020.
4. Kunang, Y.N. and Purnamasari, S.D., 2018, October. Web scraping techniques to collect weather data in South Sumatera. In 2018 International Conference on Electrical Engineering and Computer Science (ICECOS) (pp. 385- 390). IEEE 2018.
5. Scarnò, Marco, and Y. Seid. \"Use of artificial intelligence and Web scraping methods to retrieve information from the World Wide Web.\" Int. J. Eng. Res. Appl. 8, no. 1 (2018): 18-25.
6. Akrianto, M.I., Hartanto, A.D. and Priadana, A., 2019, November. The Best Parameters to Select Instagram Account for Endorsement using Web Scraping. In 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE) -2019.
7. Himawan, Arif, Adri Priadana, and Aris Murdiyanto. \"Implementation of Web Scraping to Build a Web-Based Instagram Account Data Downloader Application.\" IJID (International Journal for Development) 9, no. 2 (2020).
8. Sundaramoorthy, K., Durga, R. and Nagadarshini, S., 2017, April. Newsone - an aggregation system for news using web scraping method. In 2017 International Conference on Technical Advancements in Computers and Communications (ICTACC) (pp. 136-140) IEEE -2017.
9. Soujanya, R., Goud, P.A., Bhandwalkar, A. and Kumar, G.A., 2020. Evaluating future stock value asset using machine learning. Materials Today: Proceedings, 33,(2020)
10. Vargiu, E. and Urru, M., 2013. Exploiting web scraping in a collaborative filtering-based approach to web advertising. Artif. Intell. Res., 2(1), pp.44-54, (2013).