iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.60555

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Web Scraping Application for E-Commerce Website

Prashant Chavan[1], Dhiraj Holkar[2], Shivaji Vasekar[3]
Computer Engineering, Dr. D. Y. Patil Institute of Engineering Management and Research, Pune, India

Abstract: ***In the rapidly evolving digital world, efficient extraction of data from the web is essential for insights and decision making.***
This project introduces a versatile web scraping solution by integrating Superagent and Puppeteer. Superagent efficiently handles HTTP requests and static content, while Puppeteer excels in dynamic web page interactions. The combination of their capabilities forms a robust approach for navigating and scraping modern websites. The resulting tool enables seamless extraction of both static and dynamic content, providing researchers, analysts, and businesses with a powerful tool to acquire structured data from diverse web sources, facilitating data-driven decisions and research across multiple domains. The combined power of these libraries offers an end-to-end solution for automating data extraction and integration. It also emphasizes on ethical considerations and the importance of adhering to legal and site-specific scraping policies ensuring responsible data extraction practices.
Keywords: Web Scraper, Data Extraction, HTTP request.

I. INTRODUCTION

In the ever-evolving landscape of e-commerce, staying competitive and informed is crucial for success. The proliferation of online marketplaces, retailers, and products has made it increasingly challenging for businesses to keep a close eye on their competitors and market trends. To address this, the introduction of web scraper applications has emerged as a game-changing innovation in the ecommerce industry. Web scraper applications are specialized tools designed to extract valuable data and information from various ecommerce websites, providing businesses with a wealth of insights to inform their strategies. These applications utilize advanced algorithms to systematically navigate through web pages, capturing details such as product listings, pricing, customer reviews, and stock availability. The extracted data is then structured and organized in a user-friendly format for easy analysis.

II. NEED OF THE STUDY

If you are a person who owns a business however small or big. Web scrapper can be used for your business to increase sales and stay ahead of the competition. The way you can do this is by using web scrapper to extract prices of products that you are selling, from different ecommerce websites and comparing those prices with products on your website and providing consumers with more affordable prices by increasing or decreasing product prices as needed and hence making your business profitable. Similarly, users can use this web scraping tool to search for a product and compare the prices of this product on different websites and make the online shopping experience easy and profitable.

III. RESEARCH METHODOLOGY

The methodology section describes the strategic plan and methods employed to develop the web scraper tool, encompassing various crucial components. These include the libraries, browsers, different sources to extract data from and the analytical framework. The details are outlined as follows:

A. Functional Requirement

1) *Data Extraction:* The system must extract data from e-commerce websites, including product details, pricing information, and customer reviews.
2) *Dynamic Content Handling:* The web scraping tool must effectively interact with websites featuring dynamic content and AJAX requests.
3) *CAPTCHA Solving:* The system should implement CAPTCHA-solving mechanisms to bypass CAPTCHA challenges when encountered during web scraping.

4) *Data Parsing and Structuring:* The system should parse and structure the extracted data into a standardized format, such as JSON or CSV.
5) *Data Storage and Export:* Users should have the option to store scraped data in a variety of formats, including databases, cloud storage, or local files.

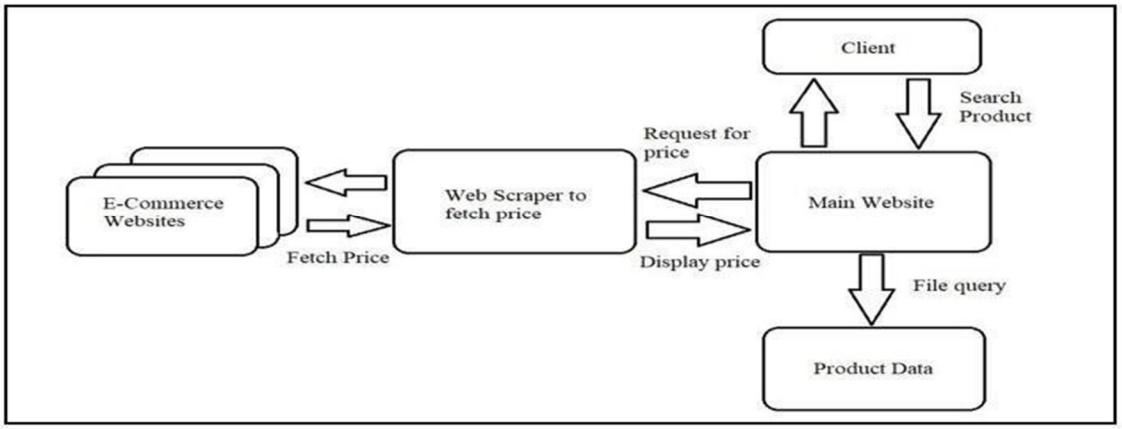
B. Non-Functional Requirement

1) *Performance:* The system must exhibit low-latency response times, ensuring quick data retrieval and user interactions.
2) *Reliability:* The system should be available 24/7, with minimum downtime for maintenance or upgrades.
3) *Scalability:* The system architecture should support horizontal scaling to accommodate increased data processing requirements.
4) *Usability:* Interface Design: The user interface should follow best practices for user experience, featuring a clean design and intuitive navigation.
5) *Compliance:* The system should comply with all relevant legal and regulatory requirements governing web scraping, data privacy, and user data protection.
6) *Resource Usage:* The system should utilize system resources efficiently, minimizing CPU and memory usage.
7) *Error Handling:* Error messages should be clear and user-friendly, aiding users in understanding and addressing issues.

C. Theoretical Framework

A web application for scraping data from ecommerce websites is a collection of programs that allows user to enter the name of desired product using the applications user interface then the web scraper scrapes the product information from the websites on the internet using tools like Superagent and puppeteer. This makes data scraping efficient. The web application typically consists of the following components:

1) A database of user profiles.
2) A front end that allows users to search the desired product for its information
3) A backend that scrapes data from websites.

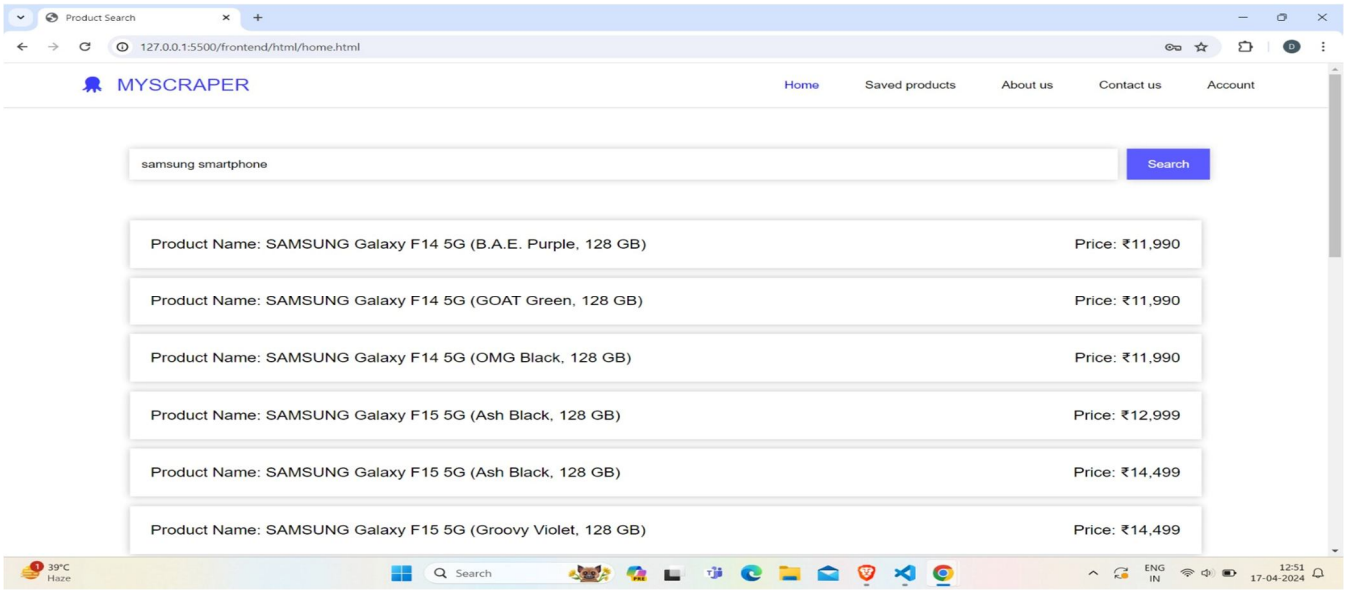


IV. RESULTS AND DISCUSSION

A web scraping application developed to extract data from e-commerce websites has generated important results and insights, facilitating competitive analysis, pricing information, and research. save the market. The following findings and reflections summarize the effectiveness and significance of the project:

1) *Data Extraction Performance:* The data collection application successfully extracted product details including price, description, customer reviews, and availability. The process is automated and scalable, allowing for continuous updates and comprehensive data collection.
2) *Data quality and Consistency:* The application maintains data integrity by handling different formats and structures across different e-commerce websites. Quality checking and error handling mechanisms ensure accurate and reliable data recovery.
3) *Information and Analysis:* Retrieved data provides valuable insights into competitors' pricing strategies, product categories and customer sentiment. This information enables strategic decisions and improved market understanding.

4) *Challenges and Limitations:* Although effective, data collection applications still encounter challenges related to dynamic web content, CAPTCHAs, and bitrate limits imposed by some websites. To overcome these obstacles, it is necessary to implement advanced techniques such as proxy rotation and user-agent randomization.
5) *Ethical Considerations:* Ethical concerns related to web scraping have been addressed by adhering to site terms of service, following robots.txt guidelines, and prioritizing privacy and security data. This application operates within legal boundaries and ethical standards.
6) *Future Improvements:* To further optimize the data collection process, future enhancements may focus on improving data normalization, real-time monitoring, and integration with analytics tools. Additionally, integrating machine learning algorithms for sentiment analysis and demand forecasting can add deeper analytical capabilities.
7) *Influencing Decision Making:* Recovered data directly affects pricing strategies, inventory management, and marketing campaigns. Real-time updates and trend analysis have enabled proactive responses to market dynamics and competitive changes.
8) *Scalability and Maintainability:* Application architecture supports scalability with minimal maintenance requirements. Regular updates and monitoring ensure ongoing reliability and optimized performance.



V. CONCLUSIONS

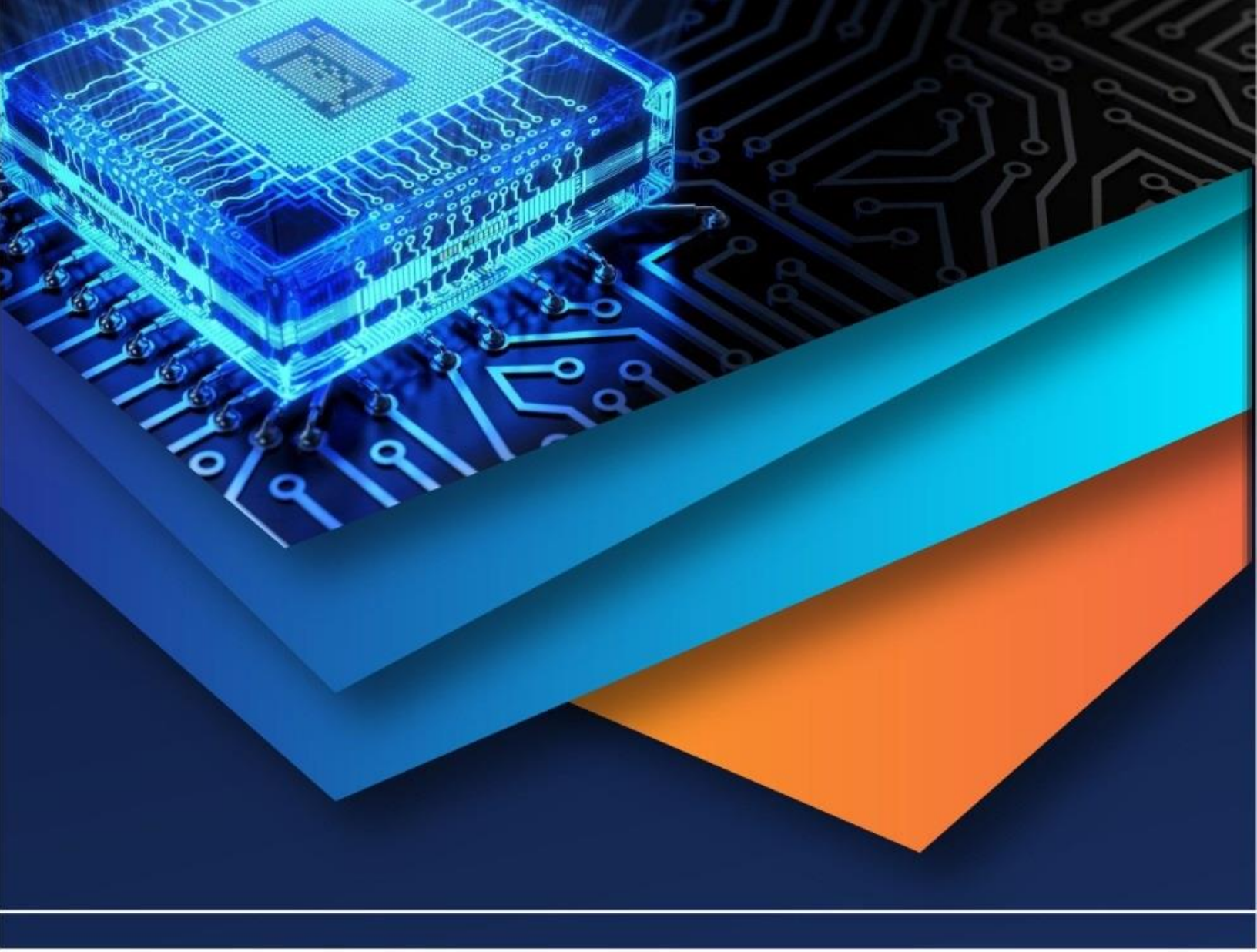
In embracing the formidable challenges of modern web data extraction, this project successfully integrates Superagent and Puppeteer to craft a comprehensive web scraping solution. The combination of Super agent's efficient HTTP requests and Puppeteer's dynamic content handling offers a versatile tool skilled at navigating diverse websites. With a focus on efficiency, adaptability, and ethical data acquisition, this project delivers a robust system capable of extracting both static and dynamic content. The seamless combination of these technologies empowers users across domains, ensuring informed decision-making and research through easy access to a wealth of web-derived information.

VI. ACKNOWLEDGMENT

We would like to thank Mr. Shivaji Vasekar, whose valuable suggestions and guidance were very important in completing our project "Web Scraping Application for E-commerce Website". Mr. Shivaji Vasekar's expertise, encouragement and tireless support played a vital role in transforming our ideas into a powerful and innovative force. We would also like to express our gratitude to our university for providing us with the space and resources necessary to help us complete this important project. Teacher support and learning support are important in our education system. We are grateful to the open-source community and dedicated developers behind the tools and libraries that support this project. Their spirit of collaboration and commitment to collaboration has been a constant source of inspiration during the development of our project.

REFERENCES

[1] Ajay Sudhir Bale Dept. of Electronics and Communication Engineering New Horizon College of Engineering Bengaluru,
[2] India. (IEEE Xplore, 19 September 2022) "Web Scraping Approaches and their Performance on Modern Websites."
[3] Harsh Khatter Department of Computer Science KIET Group of Institutions Ghaziabad, India. (IEEE Xplore, 14 October, 2022) Web Scraping based Product Comparison Model for E-Commerce Websites.
[4] Priya Matta Department of Computer Science and Engineering Graphic Era Deemed to be University, Dehradun, India. (IEEE Xplore, 20 April, 2022) Comparative Study of Various Scraping Tools: Pros and Cons.
[5] Ayush Asawa Sardar Patel Institute of Technology Mumbai, India. (16 June 2022) Co-Mart- A Daily Necessity Price Comparison Application.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY