IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

WEB SCRAPER APPLICATION FOR E-COMMERCE

[1]Prashant Chavan, [2]Dhiraj Holkar, [3] Pranay Bandichode

Computer Engineering,
[1]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. Proposed 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.

IndexTerms **– Web Scraper, Data Extraction, Web Scraping**

1. 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 e-commerce industry. Web scraper applications are specialized tools designed to extract valuable data and information from various e-commerce 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.

2. 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.

3. 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:

3.1 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.

3.2 Non-Functional Requirement

.

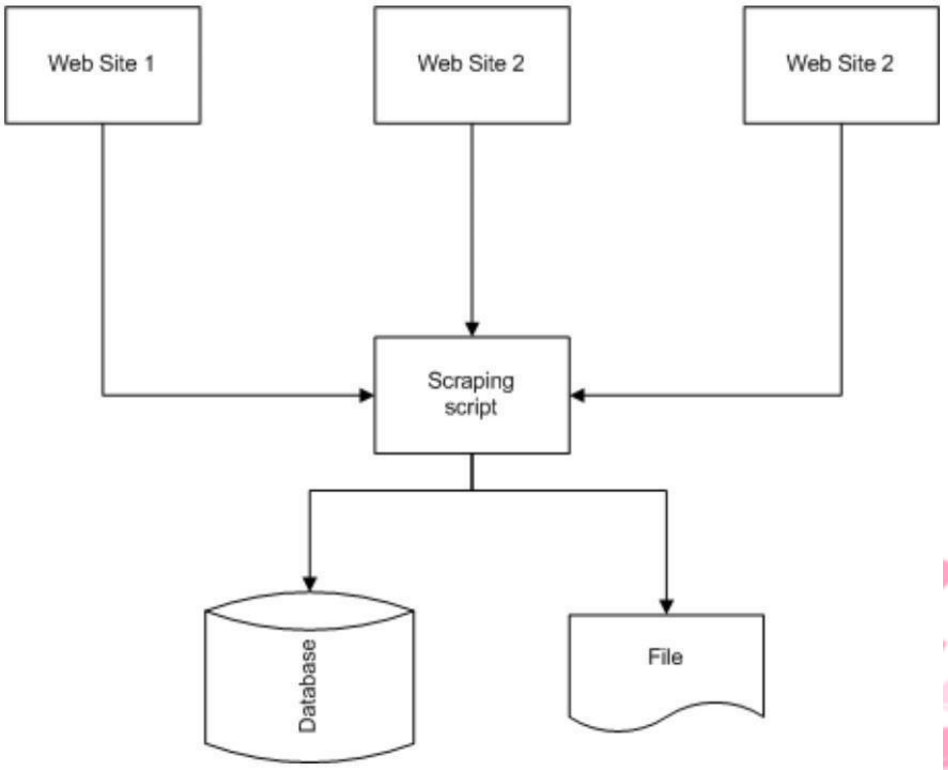
1. Performance: The system must exhibit low-latency response times, ensuring quick data retrieval and user interactions.
2. Reliability: The system should be available 24/7, with minimum downtime for maintenance or upgrades.
3. Scalability: The system architecture should support horizontal scaling to accommodate increased data processingrequirements.
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.

3.3 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 prouct using the applications user interface then the web scraper scrapes the product infornation 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:

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



3.4 ADVANTAGES

- Dynamic Content Handling: The web scraping tool must effectively interact with websites featuring dynamic content and AJAX requests. It should provide options for automating interactions, such as clicking buttons, filling forms, and scrolling.
- Scalability: The system's ability to scale and handle large volumes of data and increased user loads ensures that it can grow with the needs of the business or research project.
- Security and Compliance: The project includes mechanisms for data security, ensuring that sensitive information is protected, and compliance with legal and ethical standards in web scraping, enhancing user trust.

3.5 LIMITATIONS

- Website Structure Changes: E-commerce websites frequently update their structures, making it challenging to maintain consistent scraping functionality. Changes in HTML/CSS elements can break existing scraping scripts, necessitating constant monitoring and adjustments.
- Robots.txt and Legal Constraints: Some websites may explicitly disallow web scraping in their robots.txt files. While the project respects these rules, adherence to website terms of use and legal regulations is essential. Violations could result i n legal repercussions.

3.6 CONCLUSIONS

In embracing the formidable challenges of modern web data extraction, proposed project successfully integrates Superagent and Puppeteer to craft a comprehensive web scraping solution. The combination of Superagent'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, proposed 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.

4. REFERENCES

- https://ieeexplore.ieee.org/document/9885689
- https://ieeexplore.ieee.org/document/9915892
- https://ieeexplore.ieee.org/abstract/document/9833640
- https://ieeexplore.ieee.org/document/9753358