iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 11 **Issue:** V **Month of publication:** May 2023

**DOI:** https://doi.org/10.22214/ijraset.2023.53467

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 11 Issue V May 2023- Available at www.ijraset.com*

# Web Scraping: Extracting Insights from the DigitalLandscape

Srikanth Kulkarni[1], Ayush Buradkar[2], Pratiksha Ghadge[3], Srushti Khainar[4]
*D.Y. Patil College of Engineering Akurdi Pune*

***Abstract:*** ***Web scraping is a potent method that makes it possible to scrape useful data from websites, making it an essential tool in a number of fields like data analysis, market research, and competitive intelligence. The objective of this project is to provide a web scraping solution that is reliable and effective, automates the process of gathering data from internet sources, and offers valuable insights for decision-making. For subsequent analysis, the gathered data is saved in a structured format, such as CSV, JSON, or a database. To improve the quality and utility of the retrieved data, the project also contains methods for cleaning and preparing the data. The data is then examined statistically, visually, and with the use of machine learning algorithms to find patterns, trends, and insights that might aid in making well-informed decisions.***
***Index Terms: Data extraction, Scrapy, web-scraper, searching,user-interface***

## I. INTRODUCTION

In the era of data science engineering, collecting informa- tion from websites for analysis is completely anticipated. By learning how to scrape website pages, you can save time and money. While we must trawl through various websites to obtain information in an organised configuration, certain organisations, such as Twitter, do offer APIs to access their data in a gradually assembled fashion. The fundamental idea behind web scratching is to retrieve information that is alreadythere on a website and transform it into a format that can be used for analysis. One of the most often used programming languages for data science projects is Python. Scraping the web is easier when Beautifulsoup is used with Python. In this paper, we will get a detailed but fundamental explanation of how to use BeautifulSoup to scratch data in Python. This will make it simple and energy-efficient for information researchersto collect and store information from website pages.

## II. MOTIVATION OF PROJECT

*1)* As the number increases, it becomes more difficult to collect the data in a structural manner; therefore, this application can be utilised to obtain ready-made structuraldata.
*2)* This can be helpful for information inquiry, analysis, or building a database.
*3)* Web scraping can assist in obtaining contact details from websites, including email addresses, phone numbers, or social network profiles.

## III. ALGORITHM

Let v and w represent the rightmost nodes of F1 and F2, respectively, and let F1 and F2 be ordered forests with adistance metric cost function on nodes. The recursion yields the tree edit distance: The following is the underlying premise. We consistently contrast the forests' rightmost nodes, v and w. We branch for each of the three instances that need to be looked into when comparing the nodes: remove v, insert w, and relabel v to w. Since v is now accounted for, we removeit from its forest in the delete branch.

$$
\begin{aligned}
\delta(\theta,\theta) &= 0 \\
\delta(F_1,\theta) &= \delta(F_1 - v,\theta) + \gamma(v \to \lambda) \\
\delta(\theta,F_2) &= \delta(\theta,F_2 - w) + \gamma(\lambda \to w) \\
\delta(F_1,F_2) &= \min \begin{cases} \delta(F_1 - v,F_2) + \gamma(v \to \lambda) \\ \delta(F_1,F_2 - w) + \gamma(\lambda \to w) \\ \delta(F_1(v),F_2(w)) + \delta(F_1 - T_1(v),F_2 - T_2(w)) \\ \quad +\gamma(v \to w) \end{cases}
\end{aligned}
$$

Fig. 1. Mathematical Model

W is also taken out of its forest in the insert branch. Relabeling nodes causes us to branch twice, and the pair of relabeled nodes is then included in the mapping. This indicates that nodes descending from v can only map to nodes descending from w in order to comply with the mapping restrictions.

As a result, it is necessary to contrast the left forestof v with the left forest of w. The algorithm uses dynamic programming since the lemma states that the tree edit distancecan be determined by combining answers to subproblems. Since the result is computed from the bottom up, each potentialsubproblem requires its own table entry. The nodes of the subproblems always have consecutive indices since the forestsare given postorder indices.

## IV. LITERATURE SURVEY

1) Vidhi Singrodia, Anirban Mitra, and Subrata Paul, "A Review on Web Scrapping and its Applications (IEEE 2019)" Beginning with a basic introduction and a succinct review of various software and tools for web scraping, this paper will concentrate on various elements of web scraping. It also described the steps involved in web scraping and elaboratedon the various web scraping strategies before concluding with a discussion of the advantages and disadvantages of web scraping and a thorough explanation of the numerous applications it may be used for. Drawback: This paper reviews numerous Web scraper features, tools, and software, however it does not address the implementation process.
2) Web Scraping: Current State and Application Areas (2019) - Mamadou Bousso, Babiga Birregah, Edouard Ngor, Rabiyatou Diouf, and Usmane Sall The available frameworks, methods, groups, and tools for web scraping are reviewedin this study along with their advantages and disadvantages. Two major sections make up the study. The summary of web scraping is in Section 2. Drawback: The earlier web scraping systems are reviewed in this study, but no new techniques are presented.
3) Rizul Sharma's Web Data Scraping (2020) The overviewof the data extraction technique and its implementation using Python are the main topics of this essay. Drawback: Different challenges in the area of extracting information from hidden webs and possible solutions
4) Web scraping for data analysis using Python (IEEE 2020)Sandeep Mathur and David Mathew Thomas A database is developed that gathers all the unstructured data from multiple sources, analyses it according to its specifications, then appliesmodels and algorithms to the assembled, organised, cleaned, and re-analyzed data to get the desired results.
5) Algorithms for web scraping (2011), Kongens Lyngby In this thesis, the author investigates the potential of using approximate tree pattern matching based on the tree edit distance and constrained derivatives for web scraping.
6) Patents and Publications Web Scraping Sushitha Vijay- alakshmi S Katti, Sowmya H N, Samanvita N In this the proposed system has an option to fetch more information and do more work efficiently as Web Crawlers are involved whichdoes the task on behalf of the humans , making it accessibleto use anywhere with a click of button
7) An Optimal Data Entry Method, Using Web Scraping andText Recognition (2021) Roopesh N; Akarsh M S; C.Narendra Babu In this paper, a method to collect and preprocess thedata which is especially useful in training such chatbots. This approach employs machine learning methods such as webscraping and text recognition techniques to prepare the data
8) Utilizing Web Scraping and Natural Language Processingto Better Inform Pedagogical Practice(2020) Stephanie Lunn; Jia Zhu; Monique Ross This research full paper describes howweb scraping and natural language processing can be utilized to answer complex questions in computer science education and the application of web scraping and NLP are useful in obtaining and analyzing pertinent information from internet source. Drawback: This can expedite manual tasks, and canbe used with other techniques for additional validation.
9) Web-Browsing Application Using Web Scraping Tech- nology (2021) Won-Chi Jung,Jinsu Kim The paper discusson a network blocking-based network separation techniquethat converts data from the external network connected tothe Internet into symmetry data from which malicious code is removed through an agent and delivers it to the client of the internal networks. Drawbacks: This approach is expected to reduce the time required by reducing the process of rendering HTML information on the site in PDF and allowing access to information on the site immediately.

## V. GOALS AND MOTIVATION

### A. *Goal*

The goal of web data scraping is to extract and gatherspecific information from websites automatically. It involves using automated scripts or tools to crawl through web pages, locate relevant data, and extract it into a structured format for further analysis or use.

*B. Objective*

The objectives of web data scraping can vary depending on the specific needs and requirements of the project, but hereare some common objectives:

1) *Data Collection:* Web scraping allows you to collect largeamounts of data from various sources on the internet. This datacan be used for various purposes such as research, analysis, market intelligence, lead generation, or content aggregation.
2) *Competitor Analysis:* Scraping data from competitor websites enables businesses to monitor their competitors' activities,pricing strategies, product offerings, customer reviews, and other relevant information. This helps in identifying market trends, making informed decisions, and staying competitive.
3) *Market Research:* Web scraping provides a valuable source of data for market research. By scraping data from different websites, you can gather information on consumer preferences,pricing trends, product reviews, and market dynamics. This data can aid in understanding customer behavior, identifying opportunities, and developing effective marketing strategies.
4) *Content Aggregation:* Web scraping allows you to gather content from multiple websites and aggregate it into a single platform. This is useful for news aggregation, content duration, or building comprehensive databases for specific topics or industries.

## VI. SYSTEM ARCHITECTURE

The user interface provides an interface for users to inter- act with the data scraper system. This can be a web-based dashboard, a desktop application, or a command-line interface. Users can input their scraping requirements, configure scraping rules, monitor scraping progress, and access the extracted datathrough the user interface. The job management component handles the scheduling and execution of scraping tasks. It manages the queue of scraping jobs, assigns resources, and ensures efficient utilization of computing resources. Each scraping engine/node is responsible for visiting the target websites,navigating through web pages, and extracting thedesired data based on the configured scraping rules.

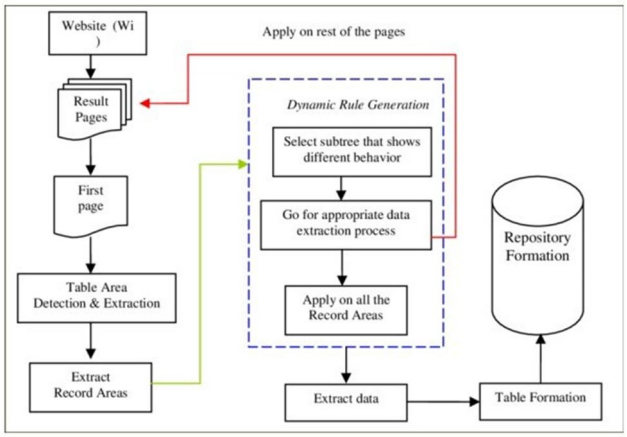



Fig. 2. System architecture

This component processes the raw data, cleanses and val- idates it, performs any necessary data transformations, and stores it in a structured format suitable for further analy- sis. Data processing may involve tasks such as removing duplicates, standardizing formats, and resolving inconsisten- cies. The processed data is typically stored in a database or data storage system for efficient retrieval and analysis. The system architecture includes the necessary infrastructure and resources to support the data scraper. This can include servers, cloud-based computing resources, storage systems, and net- work components. Scalability and reliability considerations are important to ensure that the system can handle large-scale scraping tasks and accommodating increased user demand.

## VII. OBJECTIVE

The objectives of web data scraping can vary depending on the specific needs and requirements of the project, but hereare some common objectives:

1) *Data Collection:* Web scraping allows you to collect largeamounts of data from various sources on the internet. This datacan be used for various purposes such as research, analysis, market intelligence, lead generation, or content aggregation.

2) *Competitor Analysis:* Scraping data from competitor web- sites enables businesses to monitor their competitors' activi- ties, pricing strategies, product offerings, customer reviews, and other relevant information. This helps in identifying mar- ket trends, making informed decisions, and staying competi- tive.
3) *Market Research:* Web scraping provides a valuable source of data for market research. By scraping data from different websites, you can gather information on consumer preferences, pricing trends, product reviews, and market dy- namics. This data can aid in understanding customer behavior,identifying opportunities, and developing effective marketing strategies.
4) *Content Aggregation:* Web scraping allows you to gather content from multiple websites and aggregate it into a single platform. This is useful for news aggregation, content duration, or building comprehensive databases for specific topics or industries.

## VIII. AREA OF PROJECT

The area of a web data scraping project can vary depending on the specific context and application. Here are some common areas where web data scraping is applied.

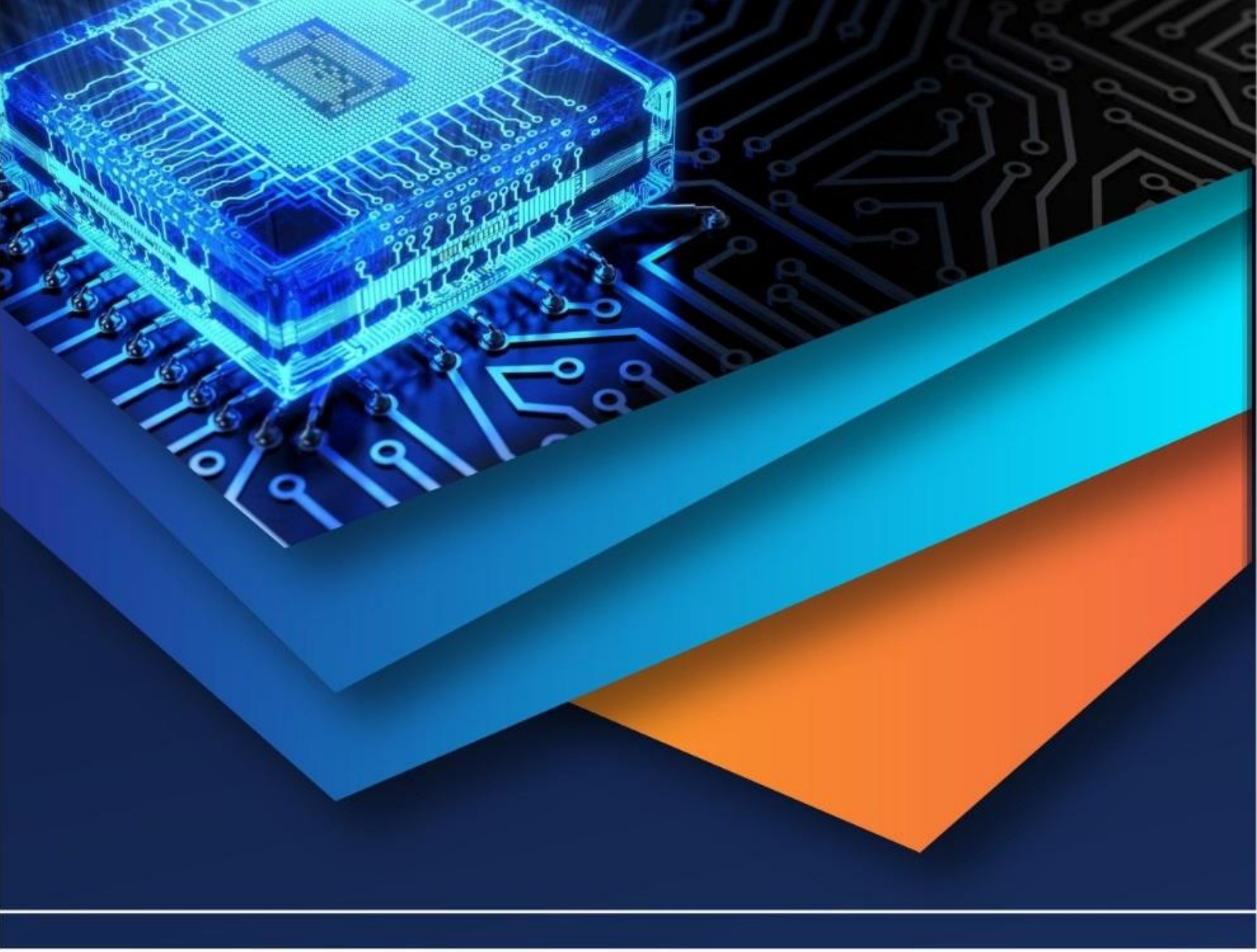
1. Market Research: Web data scraping is frequently employedin market research to compile data on rivals, consumer be- haviour, pricing trends, product reviews, and industry dynam- ics. This information aids companies in comprehending the market environment, spotting opportunities, and making wise decisions. 2. E-commerce and Price Monitoring: In the e-commerce sector, web scraping is used to track and keep track on product pricing on various platforms. Businesses can analyse pricing patterns, modify their own pricing plans, and remain competitive in the market by scraping data from a variety of websites. 3.Academic Research: Web scraping is fre-quently used by researchers to gather information for academicstudies, social science research, or data-driven analyses. They may collect huge amounts of data from online sources throughweb scraping, which they can then analyse for a variety of research projects. 4.Analysis of Social Media: Web scrapingis used to gather information from social media sites like Twitter, Facebook, and Instagram. Uses for this data include sentiment analysis, brand monitoring, social network analysis, and determining user patterns and online behaviour.

## IX. CONCLUSION

The proactive nature of the concept enables the user to an- ticipate changes in the product's pricing and plan ahead to pur-chase it in the future. Hourly based data scraping occasionally provides customers with reliable results because product pricesfluctuate greatly. The suggested solution successfully satisfies customer expectations while making product purchases byoffering the greatest deal on offer on e-commerce websites. By automating the web scraping process, more reliable data is produced while simultaneously reducing the need for manual labour.creating a user interface, a web or mobile application, and web extensions to make use of it simple for users. Model development and feature research for improved performance outcomes. increasing the number of testing models used to increase coverage and get rid of exceptions, defects, and errors.

## REFERENCES

[1] Renita Crystal Pereira, Vanitha T. "Web Scraping of So- cial Networks". International Journal of Innovative Research in Computer and Communication Engineering, vol. 3, pp.237-239, Oct. 7, 2018
[2] Patrick Hagge Cording, "Algorithms for Web Scraping",Kongens Lyngby 2011.
[3] Roopesh N, Akarsh M S, C. Narendra Babu, Senior Member, IEEE M S Ramaiah University of Applied Sciences, India "An Optimal Data Entry Method, Using Web Scraping and Text Recognition", 2021 International Conference on Information Technology (ICIT).
[4] Sushitha S, Vijayalakshmi S Katti, Sowmya H N, Samanvita N. "Patents and Publications Web Scraping", IJCSNInternational Journal of Computer Science and Network, Vol- ume 5, Issue 2, April 2016
[5] SARR, E. N., Ousmane, SALL., DIALLO, A. (2019, October). Fact Extract: "Automatic Collection and Aggrega- tion of Articles and Journalistic Factual Claims from Online Newspaper". In 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS) (pp. 336-341). IEEE.
[6] Saurkar, Anand V., Kedar G. Pathare and Shweta A. Gode, "An Overview On Web Scraping Techniques And Tools", International Journal on Future Revolution in Com- puter Science and Communication Engineering, pages 363- 367, 2018.
[7] Rahul Dhawani, Marudav Shukla, Priyanka Pu- var, Bhagirath Prajapati, A Novel "Approach to Web ScrapingTechnology", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue5, 2019.
[8] S. d. S. Sirisuriya, "A comparative study on web scraping", 8th International Research Conference KDU, pp. 135-140, November 2015.
[9] Holbert Ghazvinian and Viswanathan, "Simple WebScraping", Jun. 2015, [online] Available: https://seanholbert.wordpress.com/2011/07/15/scrappy-simple-webscraping/.
[10] Subrata Paul, Vidhi Singrodia, Anirban Mitra, "A Review on Web Scrapping and its Applications", 2019 In- ternational Conference on Computer Communication and In- formatics (ICCCI -2019), Jan. 23 – 25, 2019, Coimbatore, INDIA.
[11] Pontus Andersson, "Developing a Python based web scraper", A study on the de43 velopment of a web scraper for TimeEdit, Summer 2021.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY