International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 11 | November - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Web Scraping and its Applications

Mr. Saisharan Erlewad[1], Mr. Siddhant Bhosale[2], Mr. Rohit Chavan[3], Mr. Ganesh Giri[4], Prof, Himanshi G. Agrawal[5]
[1,2,3,4]Undergrad. Student, Dept. of Information Technology SKN Sinhgad Institute of Technology & Science, Lonavala, Maharashtra
[5]Asst Professor, Dept. of Information Tech, SKN Sinhgad Institute of Technology & Science, Lonavala, Maharashtra

Abstract- *In the age of information, the plenitude of data on the World Wide Web has come an inestimable resource for various industries and research domains. This research paper explores the world of web scraping and its different operations in our data-rich period. It delves into the elaboration, tools, and challenges of web scraping, emphasizing its ethical and legal confines. Through real- world case studies, we illustrate its vital part in disciplines similar to e-commerce, finance, social media, healthcare, journalism, and academia. While web scraping unlocks unknown data- driven perceptivity, it prompts conversations on sequestration and data power. This paper concludes by considering arising technologies and the transformative eventuality of web scraping in driving data-driven invention across industries*

Keywords- **web scraping, python, library, Selenium, Data Extraction.**

I. INTRODUCTION

Web scraping, the automated extraction of data from websites, has surfaced as a transformative technology in this landscape. It offers the means to unleash the hidden gems within the ever-expanding digital universe, enabling data-driven decision-making, innovation, and insights like never before. This research paper is a journey into the world of web scraping and its myriad applications, illuminating the ways, challenges, and ethical considerations that accompany this technological prowess. As we traverse the intricate web of web scraping, it's imperative to trace its evolution, from its rudimentary origins to the sophisticated results of today. This technology has evolved in tandem with the exponential growth of online data, adapting to the dynamic nature of the internet. In addition to discussing its evolution, we delve into the tools and libraries that empower web scrapers, making data extraction from websites more accessible and effective. Likewise, web scraping isn't just a technical marvel; it's a catalyst for transformative change across various sectors. This paper delves into the multifaceted applications of web scraping, showcasing real - world case studies and examples from e-commerce, finance, social media, healthcare, journalism, academia, and beyond. By scraping data from sources as different as online commerce, social networks, and academic databases, web scraping offers a regard into trends, patterns, and perceptivity that have remained hidden in the vast expanse of the web.

Fig1. Web Scraping Timeline

II. BACKGROUND STUDY

The internet's exponential growth has made data a precious resource. Web scraping, the automated extraction of data from websites, has emerged as a powerful tool. Initially used for search engine indexing, it has evolved with advanced techniques and tools. Web scraping finds applications across diverse fields, from e-commerce to healthcare and academia. However, ethical and legal concerns have arisen, necessitating responsible use. As industries and research increasingly rely on data, web scraping has emerged as a linchpin for harnessing the wealth of online information. This research paper explores the history, applications, and ethical dimensions of web scraping, highlighting its transformative potential in the digital age.

Web scraping has become indispensable in the data-centric digital landscape, bridging the gap between the vast sea of unstructured web content and the structured information needed for decision-making. The ability to automate data collection from online sources has empowered businesses, researchers, and individuals to access, analyze, and utilize web data efficiently. It has fostered innovation, allowing data-driven insights to guide strategies and discoveries across industries. In an age where data reigns supreme, web scraping offers a key to unlock the potential of the internet's vast information repositories. This paper aims to delve deeper into its evolution, applications, and the essential ethical and legal considerations that underpin its responsible use, underlining the transformative impact it can have on data-driven endeavors

III. WEB SCRAPING AND ITS APPLICATIONS

Web scraping is a technique used to extract unstructured data from websites and convert it into structured data which can be further stored and analyzed in a database. It involves accessing a website's HTML code, parsing it, and then collecting specific data elements, such as text, images, links, or other content, in a structured format. Web scraping is commonly used for a variety of purposes, including data analysis, data collection, competitive research, price monitoring, content aggregation, and more. It can be performed using various programming languages and tools, and it allows users to transform unstructured web data into structured, usable information for further analysis or applications. However, web scraping must be done in compliance with ethical guidelines and legal regulations to respect the rights of website owners and protect user privacy. The overall goal of the web scraping process is to extract information from webpages and transform it into an understandable structure like spreadsheets, database or comma-separated values(csv) or any other format shown in the figure below.

A. APPLICATIONS

- Online price comparison
- Contact scraping
- Weather data monitoring
- Research
- Extract offers and discounts (E-commerce)
- Scrap job posting information from job portals
- Collect property list and details
- Market Analysis
- Collecting government data
- Social Media

IV LITERATURE REVIEW

1. WEB SCRAPING OR WEB CRAWLING: STATE OF ART, TECHNIQUES, APPROACHES AND APPLICATIONS

This research paper delves into the realm of web scraping and its various applications, techniques in our data-rich age. It explores the intricacies, and different tools and techniques like HTTP programming, HTML Parsing, DOM Parsing, and challenges associated with web scraping, with a particular emphasis on its ethical and legal boundaries. Various web scraping software's are also

2. WEB SCRAPING WITH PYTHON AND SELENIUM:

In this paper the method for retrieving web information has been elaborated, using a block-based structure obtained by python script. The proposed work focuses on data extraction developed in python using HTML, parsing running on Anaconda Platform, Script is supported by Selenium library.

3.WEB SCRAPING: APPLICATIONS AND SCRAPING TOOLS:

This paper comprises of different web applications and tools discussed in detail, some of the tools like Scraper API, FMiner, Octoparse, Parsehub, Scrapy, Web content Extractor (WCE) etc. This paper also dives into the field of Machine Learning where tremendous amount of data is needed.

4.WEB SCRAPING TECHNIQUES AND APPLICATIONS: A LITERATURE REVIEW:

This paper consists of the uses of the web scraper in different fields like healthcare, social media, finance, marketing and research, the author has also described the comparison between different crawler and their types to choose depending on the situations to work on.

5. A SURVEY ON WEB SCRAPING AND ITS APPLICATIONS

This paper elaborates the use of web scraper in the field of finance, business and data science, the research depicts the different approaches for web scraper like Mimicry, Weight Measurement, Differential and Machine Learning approach.

V. WEB SCRAPING TECHNIQUES

Web scraping is a versatile process that involves multiple techniques and strategies for extracting data from websites. The choice of technique depends on the nature of the data, the structure of the website, and the specific requirements of the scraping task. Here are some key web scraping techniques:

Static Web Page Scraping: This is the most basic form of web scraping, where data is extracted from static HTML web pages. It involves sending an HTTP request to the web server, receiving the HTML content in response, and then parsing and extracting the desired data using tools like BeautifulSoup in Python

Dynamic Web Page Scraping: Many modern websites use JavaScript to load data dynamically. To scrape data from such websites, you can employ headless browsers like Selenium. These tools automate the interaction with the website, allowing access to dynamically generated content.

API Scraping: Some websites offer Application Programming Interfaces (APIs) that allow structured access to their data. API scraping involves making HTTP requests to the API endpoints and receiving data in a structured format, typically in JSON or XML.

XPath Scraping: XPath is a language for navigating XML documents and is widely used for scraping data from web pages. It provides a structured way to locate and extract specific elements within an HTML or XML document.

DOM Parsing: Browsers such as Mozilla browser or Internet Explorer can parse web pages into a DOM tree using the parts of pages retrieved by programs. The resulting DOM tree is parsed using languages like Xpath.

VI. WEB SCRAPING SOFTWARE:

Web scraping software, are tools or web scraping software programs that facilitate the process of extracting data from websites. These tools provide a range of features to make web scraping more efficient, user-friendly, and powerful. Here are some popular web scraping software options:

Visual Web ripper: - Visual Web Ripper is a proprietary web scraping software and data extraction tool developed by Sequentum. It is designed to simplify the process of extracting data from websites, especially for users who may not have extensive programming or coding skills. Visual Web Ripper provides a user-friendly and visual interface, making it easier to set up and manage web scraping tasks

Out wit hub: "OutWit Hub" is a web scraping and data extraction software tool that allows users to collect data from websites and web pages. It provides a range of features for web scraping and data harvesting, making it easier to gather information from the internet.

WebHarvy: WebHarvy is a visual web scraping software application developed by SysNucleus. It is designed to make the process of extracting data from websites easy and accessible, WebHarvy provides a point-and-click interface for creating web scraping agents, making it straightforward to collect data from websites.

Helium scraper: Helium Scraper is a web scraping and data extraction software tool designed to extract data from websites. It provides a range of features to make web scraping more accessible and user-friendly, making it easier for users to collect the data from the internet.

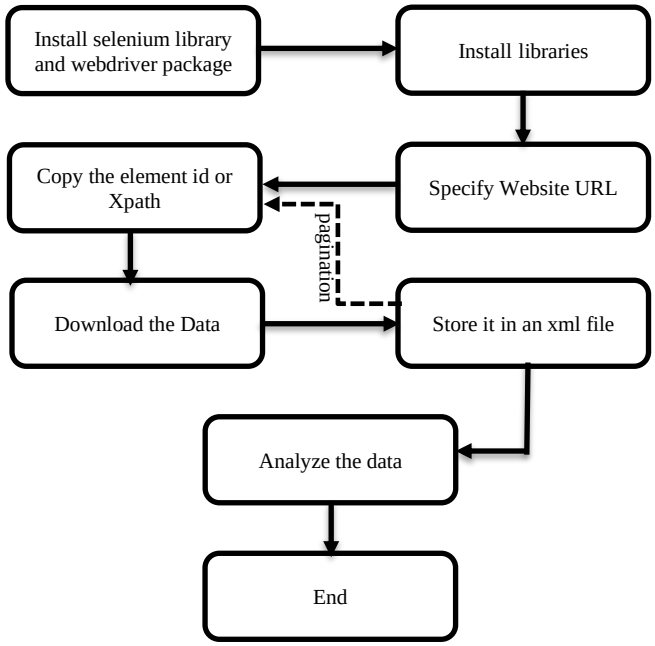
FMiner: Fminer is one of the best visual web scraping tool built in python. It has nice diagrammatic representation of scraping flow and actions.

VII. METHODOLOGY:

The work mainly focusses on analyzing a website and extracting required data which can be in the form of lists, tables or unstructured data in various available structured formats such as SCV, XML, SQL databases using respective python libraries

libraries used-

1. *Python*
2. *Selenium library:* web testing library used for extracting data from any webpage using elements id, Xpath, CSS Selectors etc.
3. *Requests library:* library that provides interaction with http webpages
4. *openpyxl library: library used to store data fetched from the website to xml file.*



WEB SCRAPING CODE OF CONDUCT:

- Not to distribute the downloaded materials without valid permission.
- Downloading copies of documents that are not public is not permitted.
- Protection and security of confidential user data information.
- Respect for websites terms of service.
- Data collected should not be used for other illegal purposes.

VIII. DISCUSSION:

When we start web scraping, we get to know the value of the data that we collect especially seeing the size, variety and amounts of data, we get different types of data from browser every day in the forms of data packets over internet which can be in the form of videos, images, files, documents, xml sheets etc. The internet serves as a goldmine for a lot of unstructured and structured data but unfortunately, the unstructured data is hard to extract and difficult to access. Modern browsers are brilliant at showcasing visuals, displaying motions, and arranging out websites in pleasing order and style but they do not offer a capacitance to export their data at least not in most of the situations. Hence web scraping, instead of seeing the webpage through the interface of your browser, gathers the data from the browser. Nowadays most of the websites provide an API with gives access to the data, but most of the website may or may not provide an API to interact with or doesn't expose an API required for the functionality.in these cases web scraper can be handy.

IX. REFERENCES:

[1] Sarah Fatima, Shaik Luqmaan, Nuha Abdul Rasheed (Web Scraping with python and Selenium) 2021.

[2] Nikita Sharma, Bhasker Pant,Sachin Sharma (Web Scraping: Applications and Scraping tools) 2020.

[3] Chaimaa Lotfi, Swetha Srinivasan, Myriam Ertz, Imen Latrous (Web scraping techniques and Applications: A Literature review:) – 2021

[4] Prof.Shivsagar Gondi (A Survey on Web Scraping and its Applications) [IJCRT] 2021

[5] Moaiad Ahmad Khder - (Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application) 2021

[6] A Comparative study on web Scraping de S Sirisuriya – 2015

[7] Qingli Niu, Irfan Ali Kandhro, Anil Kumar, Shahnawaz shah, Muhammad Hasan, HifzaMehfooz Ahmed, and Fei Liang – (Web Scraping Tool for Newspapers and Images Data Using Jsonify)