# Web Scraping

[1]Aniket Kolhe, [2]Dhiraj Khachane, [3]Suresh Kandelkar, [4]Dipanjali Shipne [1,2,3] Student Of B.E Information Technology, Department Of Information Technology, Paramhansa Ramkrishna Maunibaba Shikshan Santha's Anuradha Engineering College,Chikhli, Maharashtra, India
[4]Assistant Professor Department Of Information Technology, Paramhansa Ramkrishna Maunibaba Shikshan Santha's Anuradha Engineering College, Chikhli, Maharashtra, India
[1]aniketkolhe1041@gmail.com, [2]dhirajkhachane999@gmail.com,
[3]sureshkandelkar505@gmail.com, [4]d.shipne@gmail.com

**ABSTRACT:** Web scraping is a powerful technique used to extract data from websites. This process involves the use of software to access web pages, parse their HTML or XML content, and retrieve specific information for various applications. Web scraping has become increasingly important in fields such as data science, business intelligence, and market research, where access to large datasets is crucial for analysis and decision making. This study explores the methodologies and tools used in web scraping, including popular libraries such as BeautifulSoup, Scrapy, and Selenium. It also addresses the ethical and legal considerations associated with web scraping, such as compliance with website terms of service and data privacy regulations. Furthermore, the abstract highlights the challenges faced in web scraping, such as handling dynamic content, CAPTCHA systems, and anti-scraping measures implemented by websites.

**KEYWORDS:** Web scraping, Web scraping tools.

## INTRODUCTION:

Web Scraping is the automation of the data extraction process from websites. One way is to copy-paste the data, which is both tedious and time-consuming manually , So This event is done with the help of web scraping software known as web scrapers. They automatically load and extract data from the websites based on user requirements. These can be custom-built to work for one site or can be configured to work with any website. Assuming the data as the most important resource for your e-Commerce business. According to Isa et al. [1], “It is important to every business to know their level of market competition, for example customers demand, customers’ pattern of buying and their sales performance.”

Scraping browsers are popular these days for data scraping projects owing to their efficiency. One such browser is Bright Data Scraping Browser. It is an automated browser designed specifically for data scraping purposes. Its efficient website unblocking capabilities, compatibility with Puppeteer and Playwright, scalability, and AI technology make this tool a market hit. Along with saving you time and resources
while fulfilling your data scraping needs, it is also great for automating any number of other browser-based activities. It can bypass the toughest website blocks and has a way around bot-detection systems. The data displayed by most of the sites are viewed using web browsers. The web browser

does not offer to save the data in a user-friendly format. The data can be saved only as a web page, and most web pages only give one option to the user- to manually copy and paste the data. Web Scraping is a smart technique that can be utilized to extract vast amounts of information from the target websites. The extracted data then can be saved to a local file on your system or as a spreadsheet format. Web scraping advantages and processes can automate the processes of extracting data from the website using scripts.

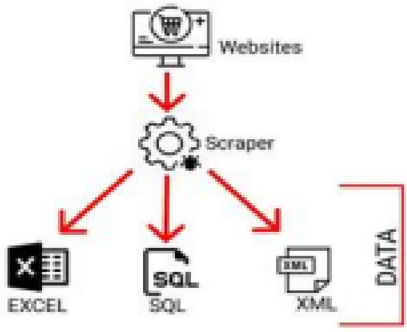


**Fig 1 :** Working Of Web Scraping.

## WEB SCRAPING USES.

The Internet is a data store of world's information - be it text, media or data in any other format. Every web page display data in one form or the other. Access to this data is crucial for the success of most businesses in the modern world. Unfortunately, most of this data is not open. Most websites do not provide the option to save the data which they display to your local storage, or to your own website. This is where a Web Scraping software can help you. The uses of Web Scraping for business as well as personal requirements are endless. Each business or individual has their own specific need for gathering data.

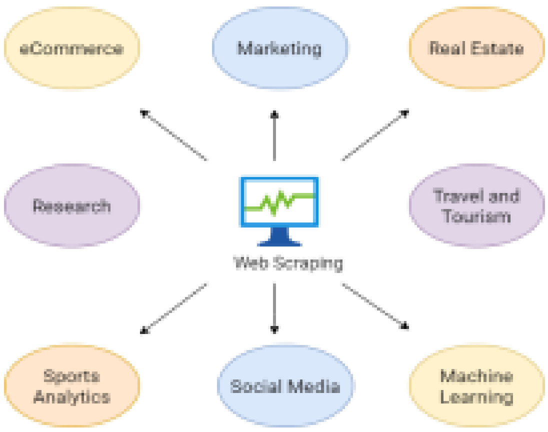


**Fig 2 :** Web Scraping Uses.

## WEB SCRAPING TOOLS AND TECHNIQUES.

• **Web Scraping Tools:-**

The Software that is used for web scraping helps to ease the manual task that is very tedious. The software automatically extracts all the data that is required or mentioned by the user, like if a user wants to keep track of a product, he just has to enter the link. The extracted data will be provided in a tabular form. So, web scraping can automate the manual work programmatically by visiting each page and extract data from pages and parsing the HTML pages. There are many software or tools that are available in the market.

**1. Beautiful Soup :**

BeautifulSoup is a Python parsing library that extracts data from HTML and XML files. It is easy to use and can navigate through HTML and XML documents. This tool extracts data from web pages such as images, texts, links, etc. BeautifulSoup mainly works by parsing HTML or XML files and generating a parse tree that it can traverse to locate specific elements. It also includes various functions for searching and filtering the parse tree.

**2. Scrapy :**

Scrapy is a Python-based framework aimed at extracting data from internet resources. It contains a rich set of web crawling, data extraction, and processing functionality, making it a powerful and flexible tool for web scrapers. One of the major advantages of Scrapy is its performance. It is fast, efficient, and able to extract a large amount of data in a short amount of time. It can also format the results in several formats, including CSV, JSON, XML, and others.

**3. Selenium :**

Selenium focuses on browsers, which means it can be slower and requires a browser, unlike other web scraping tools. However, Selenium's flexibility and versatility make it an effective and powerful tool for scraping dynamic pages. Selenium is compatible with popular programming languages like Python, Java, and C#. It can access the HTML of the web page and extract data. It also contains built-in methods for accessing specific elements from the web page using element IDs and classes.

**4. Octoparse :**

Octoparse is a web scraping tool perfect for anyone who needs to extract data from websites but wants to save time learning to code. With Octoparse, you can scrape data using a simple, visual point-and-click interface, which means you don't need any programming knowledge to get started. Octoparse can organize the result in various formats, such as Excel, CSV, JSON, and more. It also provides a cloud-based scraping
solution that enables users to run it on their server. Octoparse has a built-in data extraction engine that can automatically identify and extract the data you need from web pages, so you don't have to waste time manually selecting data fields.

• **Web Scraping Techniques :-**

**1. DOM Parsing :**

DOM parsing is a web scraping technique that involves analyzing the HTML structure of a web page to extract specific data elements like headings, paragraphs, images, links, etc. The Document Object Model (DOM) is a tree-like structure that represents the HTML structure of a web page. Think of it as a treasure hunt for data from numerous sources.

**2. Regular Expression :**

Regular expressions (or regex for short) are a powerful technique used in web scraping to identify and extract specific patterns in the text of a web page. It's like having a superpower that allows you to find and extract information based on a specific set of rules.

**3. XPath :**

XPath is a language that lets you navigate through the elements of an HTML document and select specific elements or attributes. You can use it with libraries like lxml or Scrapy to scrape data from websites.

## ADVANTAGES.

1. **Cost and time savings :**Web scraping can save time and money by automating data collection and reducing the need for human labor.
2. **Accuracy :**Web scraping can produce fast, reliable results that are difficult for humans to achieve.
3. **Real-time data :**Web scraping can provide access to up-to-the-minute information from websites.
4. **Competitive advantage :**Web scraping can help businesses gain insights into their competitors' strategies and stay ahead of the competition.
5. **Low maintenance :**Web scraping technologies require little to no maintenance over time.
6. **Customizable :**Web scraping APIs can be customized to collect and analyze data in a way that suits your needs.
7. **Multiple uses :**Web scraping can be used for a variety of purposes, including news monitoring, price comparison, market research, and email marketing.

## DISADVANTAGES.

1. **Time-consuming:** Web scraping can be time-consuming, especially when it comes to analyzing the data that's retrieved.
2. **Technical expertise:** Setting up a web scraper can be difficult and require technical expertise.
3. **Anti-scraping techniques:** Websites often use anti-scraping techniques like CAPTCHAs and IP blocking to prevent web scraping.
4. **Legal disputes:** Web scraping can be in a legal gray area, and scraping private, copyrighted, or restricted databases without authorization could lead to lawsuits.
5. **Overloading web servers:** Scraping a website extensively can cause heavy traffic and overload the web server, which could cause economic loss for the site owner.

## APPLICATIONS.

Every emerging paradigm gains its value with an increase in its applications. Similar is the case with web scraping. Many academicians [2,3,4,5,6] discussed about various applications of web scraping. Without the help of web crawling, one would never have Google providing them with search results in an increasingly efficient manner. Google crawls around 25 billion or more web pages every day to provide you with search results. Web-crawlers arrange and sort the page result as well as assess the quality of content extracted and play many other functionalities to carry out the indexing as a result. Hence, web crawlers play a vital in generating accurate results. Therefore, web scraping is integral to the functioning of search engines.

**1. PRICE MONITOR :**

In this period of e-commerce, cost plays a fundamental purpose. One needs to keep a record of the competitor's pricing maneuverings. However, manually tracking the prices is not a viable alternative where prices face a lot of ups and downs now and then. Here web scraping shows its effect. It automates the manner of deriving the estimates as well as updates you all about the new pricing approaches placed by your opponents. One has to bring changes to day to day trade with scraped product data and, unfortunately, increase your company's competitive environment. Scaling from automatic pricing solutions to profitable investment insights, this data moves mountains. Hence, web scraping helps in smart Investment Decision Making.

**2. MARKET RESEARCH** :

Market research is critical; therefore, it requires the most accurate data for a better decision-making process. One needs High quality, and profoundly insightful web scraped data fulfilling the requirements of market analysis as well as business intelligence worldwide. Hence web scarping is a better choice for Market Trend Analysis, Market Pricing, Optimizing Point of Entry, Research & Development, Competitor Monitoring, etc.

**3. SENTIMENT ANALYSIS:**

This is one of the popular applications of web-scraping data from many social media panels along with remark sectors. Nowadays, a computer can state with excellent efficiency, whether the photograph that is posted, is a pen or a pencil. Now the level is raised to that level that even for an Election, a computer would tell with fair
accuracy, the name of the winning candidate, by analyzing their tweets where it does not even have to be a direct name of the candidate itself.

## CONCLUSION:

In today's era, one can find the emergence of a new paradigm in every couple of years, so is the emergence of web scraping. This paradigm has its roots in the requirement of analysis of structured as well as unstructured data. There are various aspects related to web scraping. Some of them have been discussed in this paper. Initially we have discussed various applications of web scraping. This paper is majorly focused on tools of web scraping. The availability of these tools has made helped a lot of entrepreneurs to expand their business. As manual data extraction can be tedious tools offer a variety of services that help the user to meet his/her needs. The tools discussed above have some of their pros and cons which help us choose a better fit for the work to be done. Later we hope to expand our knowledge on these tools and come up with a prototype.

## REFERENCE:


**1.** Makino, Yuma, and Vitaly Klyuev. "Evaluation of web vulnerability scanners." In 2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), vol. 1, pp. 399-402. IEEE, 2015.

**2.** Makino, Yuma, and Vitaly Klyuev. "Evaluation of web vulnerability scanners." In 2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), vol. 1, pp. 399-402. IEEE, 2015.

**3.** Junjoewong, Lalita, Supatsara Sangnapachai, and Thanwadee Sunetnanta. "ProCircle: A promotion platform using crowdsourcing and web data scraping technique." In 2018 Seventh ICT International Student Project (ICT-ISPC), pp. 1-5. IEEE, 2018.

**4.** Indra, Evta, and Tamarai Dinesh. "Designing Android Gaming News & Information Application Using Java-based Web Scraping Technique." In Journal of Physics: Conference Series, vol. 1230, no. 1, p. 012069. IOP Publishing, 2019.

**5.** Le, Quang Thai, and Davar Pishva. "Application of Web Scraping and Google API service to optimize convenience stores' distribution." In 2015 17th International Conference on Advanced Communication Technology (ICACT), pp. 478-482. IEEE, 2015.

**6.** Barcaroli, Giulio, Alessandra Nurra, Marco Scarnò, and Donato Summa. "Use of web scraping and text mining techniques in the istat survey on information and communication technology in enterprises." In Proceedings of quality conference, pp. 33-38. 2014.