Web Scraping

Introduction:

- Some websites can contain a very large amount of valuable data like Stock prices, product details, sports stats, company contacts, and many more.
- If you wanted to access this information, you would have to copy-paste the information manually into a new document. Here's where web scraping can help.
- Web scraping refers to the extraction of data from a website. This information is collected and then exported into a format that is more useful for the user. Be it a spreadsheet or an API.
- In web scraping we make use of automated tools as it is cheap and faster as compared to work done manually.
- Web scraping in static websites are easier as compared to scraping in dynamic web site. But Scraping can be done in dynamic websites by using automation libraries like selenium.

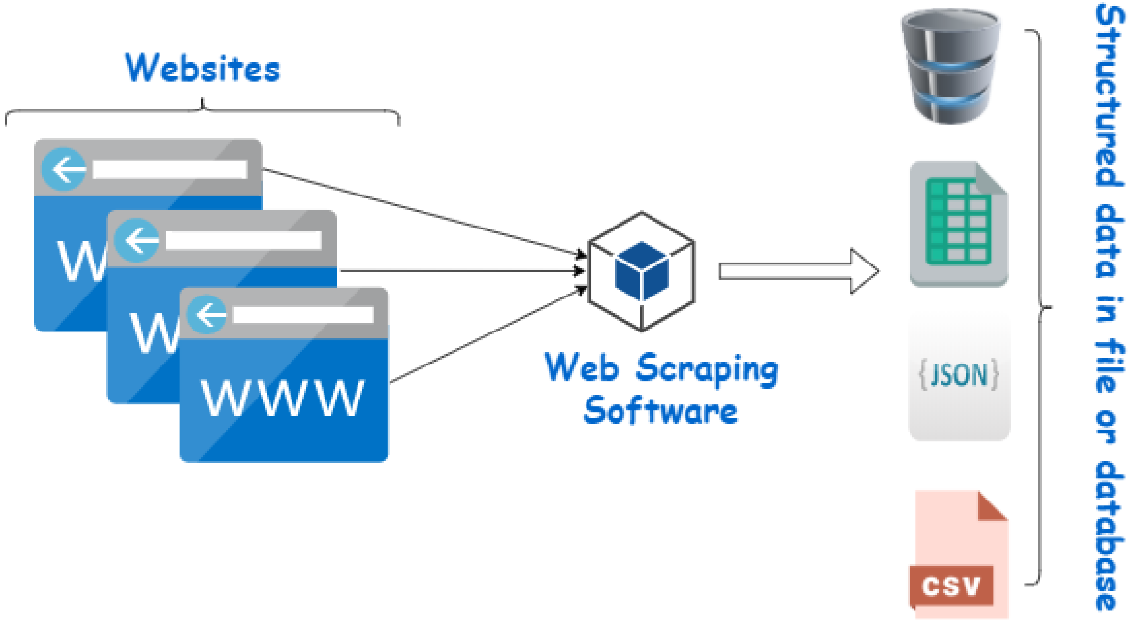
Applications:-

- Scraping stock prices into an app API
- Scraping data from a store locator to create a list of business locations
- Scraping product data from sites like Amazon or eBay for competitor analysis

How Web Scarping Works?

- Automated web scrapers work in a rather simple but also complex way.

- First, the web scraper will be given one or more URLs to load before scraping.
- The scraper then loads the entire HTML code for the page in question. More advanced scrapers will render the entire website, including CSS and JavaScript elements.
- Then the scraper will either extract all the data on the page or specific data selected by the user.
- Ideally, the user will go through the process of selecting the specific data they want from the page.
- After the extraction of data the scraper gives output in form of csv format or spreadsheet format.



Platform for doing Web scraping.

We can use multiple platforms for doing web scraping some of which are listed below.

- Scraping-Bot.io :-
  - It is an efficient tool to scrape data from a URL. It provides APIs adapted to your scraping needs: a generic API to retrieve the Raw HTML of a page, an API specialized in retail websites scraping, and an API to scrape property listings from real estate websites.
- Scrapingbee :-
  - It is a web scraping API that handles browsers and proxy management. It can execute Javascript on the pages and rotate proxies for each request so that you get the raw HTML page without getting blocked.
  - They also have a dedicated API for Google search scraping
- xtract.io :-
  - It is a scalable data extraction platform that can be customized to scrape and structure web data, social media posts, PDFs, text documents, historical data, even emails into a consumable business-ready format.

- Also if the person is from technological background then he/she can create their own web scraping applications by using some of the programming languages and their respective libraries which are listed below.
- Python:

  - Python is beneficial tool for web scraping because it includes two impactful frameworks which matters while conducting this process, Scrapy, and Beautiful Soup.
  - The use of 'Beautiful Soup' application in python is intended for quick and efficient data extraction practices.
  - It contains advanced web scraping libraries which makes Python a better hit when compared to the remaining web scraping languages.
  - It contains a variety of the finest data visualization libraries for users like you to function better with.
- NodeJS
  - NodeJS is Beneficial for streaming activities
  - Can conduct API's as well as socket-based activities
  - Has a built-in library
  - Can conduct basic web scraping data extraction activities
  - Has a basic stable communication
- Ruby:
  - It is a simple web scraping languages
  - It is more on the productive process
  - No signs of code repetition take place
  - You require less writing for such a language
  - This language is supported by a community of users
  - Supports multithreading

Problems while doing Web scraping

The main problems that occur while web scraping are:

- Being detected by the website you're scraping and banned;
- Geo-restrictions;
- Website Structure Changes;
- Anti-Scraping Technologies;
- Quality of data.
- While some of these problems can be solved, others you have to accept and move on with your work.
- Being detected by the website is a pretty common problem because with nowadays technologies it's not so difficult to detect non-human activity online.
- When you're scraping, you send thousands of requests over and over again and it's pretty obvious that normal humans wouldn't be able to do that so web scraping tool is involved.
- To avoid that, you need n numbers of IP that would hop around all the time and would imitate human activity and hide your tool.
- Another pretty common problem is geo-restrictions. When you're scraping you want to gather as much data as possible and sometimes that data isn't available in your country/region.
- That means that you lose some of the important info that could benefit you a lot.
- While lot's of problems most of the time depends on the quality of your web scraping and its technical abilities, some others depend on the content itself and you can't avoid that. This is why it is important to choose a high-

quality web scraping tool with lots of features that you can use for your own benefit and to be clear on what data you want to scrape and where you can find it.

On which websites can I do web scraping legally?

- The thing is that web scraping itself isn't illegal and you can scrape websites that provide you various public data. The more concerned question arises when we talk about scraping private data and use it for own purposes.
- You can avoid all the misunderstandings by simply reading the **robot.txt file** where you can find all the necessary info about everything that's allowed and what's beyond scraping.

Difference between web scraping and web crawling.

- A web crawler sometimes called a "spider," is a standalone bot that systematically scans the Internet for indexing and searching for content, following internal links on web pages.
- In general, the term "crawler" means the ability of a program to navigate web pages on its own, possibly even without a clearly defined end goal or goal, endlessly exploring what a site or network can offer.
- Web crawlers are actively used by search engines such as Google, Bing and others to extract content for a URL, check this page for other links, get URLs for these links and so on.

- On the other hand, web scraper is a process of extracting specific data. Unlike web crawling, a web scraper searches for specific information on specific websites or pages.
- Basically, web crawling creates a copy of what's there and web scraping extracts specific data for analysis, or to create something new.
- However, in order to conduct web scraping you would first have to do some sort of web crawling to find the information you need.
- Data crawling involves certain degree of scraping, like saving all the keywords, the images and the URLs of the web page.
- Web crawling would be generally what Google, Yahoo, Bing etc. do, searching for any kind of information. Web scraping is essentially targeted at specific websites for specific data, e.g. for stock market data, business leads, supplier product scraping.

Can I scrape data behind a login page??

- Some websites might hide their content and data behind login screens. This practice actually stops most web scrapers as they cannot log in to access the data the user has requested.
- However, there is a way to simply get pass a login screen and scrape data while using a free web scraper.
- Example of free web scrapper is ParseHub

Can you Crawl Facebook/ LinkedIn??

- LinkedIn and Facebook are two of the leading social media platforms with huge user bases and unmatched reach worldwide.
- It is only natural that many business owners who venture into web scraping and data acquisition want to crawl data
  from LinkedIn and Facebook.
- They are typically inclined towards scraping these sites and usually overlook the other options out there.
- We agree that Facebook and LinkedIn have their monopoly in the social media space which makes them the go-to sources for anyone looking to extract social media data.
- However, there are certain issues which render LinkedIn and Facebook scraping not feasible.
- They disallow bots in their robots.txt file
- Both LinkedIn and Facebook have a massive amount of user-generated content. And they are not happy with sharing this data with anonymous businesses who might be looking to improve their operations using the same. Robots.txt is a file used by websites to communicate with web crawling bots about how they (bots) can access the data available on the website.
- Unfortunately, LinkedIn and Facebook deny access to bots in their robots file which means, you cannot crawl data from them by any automated means.
- Also there are legal complications.
- When a website blocks access to crawlers, the ethical thing to do is leave that site and look for alternative sources.

- However, if you proceed with scraping LinkedIn/Facebook ignoring the robots file rules, be warned that they have been quite aggressive towards illegitimate scraping in the past.
- LinkedIn's legal battle with HiQ is popular and you probably don't want to get into something like that when there are alternate sources for the same kind of data.