



## Covid Information Tracker (Web Scrapping / Web Automation)

*Hemant Nasa*

Student, Information Technology Dept., Maharaja Agrasen Institute of Technology Rohini, New Delhi

ABSTRACT –

Web scraping, also known as web extraction or harvesting, is a technique to excerpt data from the World Wide Web (WWW) and save it to a file system or database for future retrieval or analysis. Commonly, web data is scrapped utilizing Hypertext Transfer Protocol (HTTP) or through a web browser commonly chrome/firefox. This is proficiented either manually by a user or automatically by a bot or web crawler. Due to the fact that an immeasurable amount of different data is constantly generated on the WWW, web scraping is widely recognized as an efficient, effective and most powerful technique for collecting big data. Searches for grey literature can demand substantial resources to tackle but their inclusion is important for research activities such as systematic reviews.



Web scraping, the excerpt of patterned data from web pages on the internet services like google, has been developed in the private sector for business purposes, but it offers considerable advantages to those searching for grey literature. By building, developing and sharing protocols that excerpt search results and other data from web pages, those looking for grey literature can drastically and immensely increase their transparency and resource efficiency. Various options and views exist in terms of web-automation software/tool and they are introduced herein the report.

### INTRODUCTION

#### SCOPE

With the inclusion of more and more data in the world of our virtual world i.e. internet, the importance of web scraping is expanding drastically. Many companies are now offering customized web scraping tools and its software to their clients in which they collect, analyse and use data from all over the world of the internet and arrange them into convenient and quickly understandable data. It decreases the priceless and valuable man-power to manually visit each website and gather the data. Web Scrapers are designed and code for each and individual website and crawlers available for us do broad scraping. If the website has a complicated structure, then it is expected to have more coding to scrap its data as compared to a simple one. The Future of web scraping is indeed shiny and irreplaceable and it will become more and more important for every business as the time pass.

#### ABOUT PROJECT

The project basically revolves around the use of web scraping and automation tool or library. It basically build on the model called request-response model where the user request the required information and the server return the HTML page. The automation and web scrapping helps to crawl over the page and helps to find the required data as searched by the user. This technique will be used for tracking the covid cases or any information regarding

the covid in India. Thus automating the process of searching and finding the required data of the given state and then arranging and analysing it for further process,

### MAIN AIM

So the main aim of this project is to bring out the benefits and boons of utilising the web scrapping technology which helps in automating the task and ease down the manual process. The software is so designed that it helps to find the covid information of the required input state or country using scrapping and automation. It crawls over the HTML page provided by the server and search and execute the code to extract and gather information for future use and analysis.

## TECHNOLOGIES USED

### JAVASCRIPT :

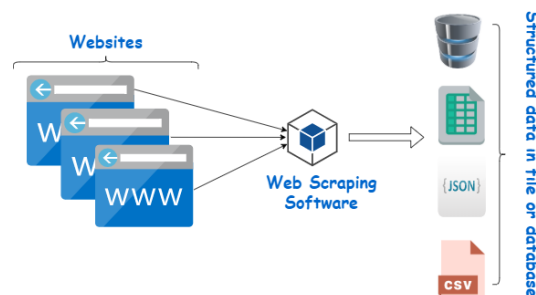
JavaScript is a cross-platform, object-oriented scripting language used to make webpages interactive (e.g., having complex clickable buttons, animations, popup menus, etc.). There are also more advanced server side versions of JavaScript such as Node.js, which helps us to allow you to add more functionality to a website than downloading files such as realtime collaboration between multiple computers. Inside a host environment (for example, a web browser), JavaScript can be connected to the objects of its environment to provide programmatic control over them.

### PUPPETEER :

Puppeteer is a Node library that helps and provides a high-level API to dominate headless Chrome or Chromium over the Devtool Protocols. It can also be configured to use full, which can also be said as non-headless, Chrome or Chromium. Testing and debugging has become more and more complicated. Puppeteer has been developed as a node library which is used to enable Chrome browser testing.

### SELENIUM :

Selenium is a free (open-source) automated testing framework which is used to validate web applications across different platforms and browsers ie. Chrome, firfox etc. We can use multiple programming languages like Java, C#, Python etc to create Selenium Test Scripts. Testing the software done using the Selenium testing tool is usually referred to as Selenium Testing. Selenium Software is not just a single tool but a suite of software, each piece catering to different Selenium QA testing needs of an organization.



### FS MODULE :

The Node.js file system module allows you to work with the file system on your computer. To handle file operations like creating, reading, deleting, etc., Node.js provides an inbuilt module called FS (File System). Node.js gives the functionality of file I/O by providing wrappers around the standard POSIX functions. All file system operations can have synchronous and asynchronous forms depending upon user requirements.

### OBJECTIVE OF PROJECT

Promoting Automation instead of Manual working:

Project revolves around automating the woking of data extraction which was usually done manually. This helps to reduce man labour work and promote machine working that not only helps in reducing the labour work but saving the other precious resources.

Saving Time:

As time is the one of the precios resources that need to saved as much as we can. Since, this project mainly focuses on automating the data extraction through automation tool and scraping, it saves much of the time which gets wasted on searching, selecting, copy pasting, and analysing the data.

Multiple Queries at single time:

It isseen that most of the site while finding the required result works on single query for ex. If you want to search details for the covid for 3-4 cities simultaneously, you will be required to hit multiple Enter and search cities Individually. But this project will help to get the desired result for multiple cities in a single call.

Improving Data analysis process:

The result that is achieved through scrapping is much more curated, readable and processed form in excel or word file that it can be easily used in the data analytics field for further process.

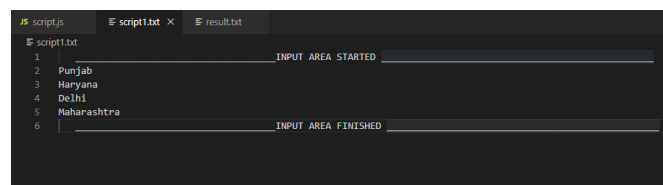
## PROPOSED APPROACH

The process starts with acquiring the required js library ie. Puppeteer and fs node module. It can be seen in the fig. 1 below.

```
const pup = require('puppeteer');
const fs = require("fs");
```

Fig. 1

Now creating a separate file for the input one can enter the required state or country name in that file which is further read by using fs node module for its use. The implemented .txt file can be viewed in the given fig. 2 below.



```
script1.txt
1
2 Punjab
3 Haryana
4 Delhi
5 Maharashtra
6
INPUT AREA FINISHED
```

Fig. 2

After the inputting the required state for which the information is to be gathered, we now need to execute the code/software. The puppeteer will automate the process of opening the chrome browser and will extract the information of the needed states by going through the site mentioned in the code. The opening of the chrome browser through puppeteer will be indicated by the small text ie. Given below in the fig. 3.

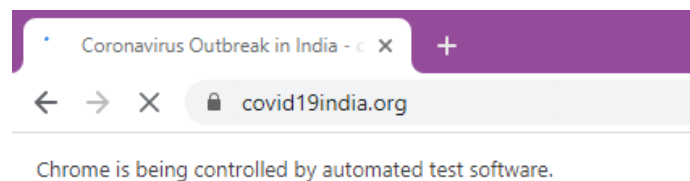


Fig. 3

Now the process of extracting the data will be completed soon within a second depending upon the specs and input data.

## RESULT

The result that will be gathered by the tool can be displayed in the desired format like .pdf, .doc, .txt, .xlsx etc. Here the result is displayed in .txt format for simplicity in fig. 4



```
result - Notepad
File Edit Format View Help
Rank -> 1
{"State/UT": "Maharashtra", "Confirmed": "66,11,078", "Active": "16,658", "Recovered": "64,50,585", "Deceased": "1,40,216", "Tested": "6.3Cr", "Vaccine Doses Administered": "9.8Cr"}
Rank -> 18
{"State/UT": "Punjab", "Confirmed": "6,02,401", "Active": "251", "Recovered": "5,85,591", "Deceased": "16,559", "Tested": "1.5Cr", "Vaccine Doses Administered": "2.2Cr"}
Rank -> 8
{"State/UT": "Delhi", "Confirmed": "14,39,870", "Active": "348", "Recovered": "14,14,431", "Deceased": "25,091", "Tested": "2.9Cr", "Vaccine Doses Administered": "2Cr"}
Rank -> 14
{"State/UT": "Haryana", "Confirmed": "7,71,252", "Active": "135", "Recovered": "7,61,068", "Deceased": "10,049", "Tested": "1.3Cr", "Vaccine Doses Administered": "2.6Cr"}
```

Fig. 4

---

## CONCLUSION

Thus, in the time where everyone is busy in saving their own time. Automation plays a key role in saving time especially in the field of data mining, data extraction, data analytics or data science. Web scraping services are considered as one of the most practiced activities done by most of the IT companies and Ecommerce Stores that operate across the globe.

## REFERENCES

---

<https://www.geeksforgeeks.org/introduction-to-web-scraping/>

<https://towardsdatascience.com/an-introduction-to-web-scraping-with-python-a2601e8619e5>

<https://librarycarpentry.org/lc-webscraping/>

<https://www.researchgate.net/publication/282658358> The Use of Web-scraping Software in Searching for Grey Literature

<https://www.researchgate.net/publication/317177787> Web Scraping