International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 01 | Jan - 2022 ISSN: 2582-3930

# ANALYSIS OF WEB-SCRAPING AND TEXT-BASEDEXTRACTION

**Aniket Singh[1], Arpit Goel[1], Upmanyu Agrawal[1] and Mr. Ajay Tiwari[2]**

[1]Department of Computer Science & Engineering, Maharaja Agrasen Institute ofTechnology, New Delhi, India.

[2]Professor, Department of Computer Science & Engineering, Maharaja Agrasen Institute ofTechnology, New Delhi, India.

**Abstract:**

Web scraping is a common practice nowadays but it is a very tedious task for beginners. Web scraping is an automatic method to obtain large amounts of data from websites. It is required to write script for scraping but not everyone is familiar with script. There are many known methods for web scraping such as copy paste, selenium, puppeteer, beautiful-soup, semantic-based, text-based-extraction etc. These methods uses tedious and time consuming approaches in various different cases which will leads to a confusing environment for scrapers. So this research focus on finding suitable method to be used in a particular condition so that scraping become easy for everyone .

## 1. INTRODUCTION

Web scraping is a fundamental task for investigating all levels of government and industry. However, almost all web scrapers are implemented as custom tools tied to a specific website and use case. This leads to unnecessary duplication of work.

Many sites include session identifiers and token systems which need to be extracted to successfully perform HTTP requests. The increasing use of client-side JavaScript in search, forms complicate the use of scrapers that operate using direct server requests. Web pages often change their style, layout, infrastructure, or query mechanism. The bulk of data extraction techniques , required for obtaining session identifiers, tokens, and data, use HTML Document Object Model (DOM) path techniques.

The Project will try to reproduce human activity to communicate with web pages. To make information extraction easier , An Interface will be created where the user will be able to extract information. By giving the input URL(link of the site to be scraped), and wanted-list(list of texts which is used to fetch similar data ) make a call to the backend to get the desired results. The idea also proposed to bring an important feature of pagination i.e. to extract data from multiple pages.

Data like item pricing, stock pricing, different reports, market pricing and product details, can be gathered through web scraping. Extracting targeted information from websites assists you to take effective decisions in your business.[1]

## 2. RESEARCH PROBLEM:

Web scraping has become a hot topic among people with the rising demand for big data. More and more people hunger for extracting data from multiple websites to help with their business development.

Therefore, analyzing different scraping techniques, and the efficient one suited for different situations. Also providing a model for text-based extraction method i.e.

a. Easier to use.
b. Provide accurate data.
c. Can be used by non-coders.

## 3. LITERATURE SURVEY:

Sameer Padghan et. al., [2] projected an approach would enable the data to be scrapped from numerous websites that will minimize human intervention, save time and also enhance the quality of data relevance. It will also support the user in gathering data from the site and to save the data to their intent and use it as the individual wishes. The scraping used would increase significantly and will often encroach on the framework to obtain the details.

Anand Saurkar et. al., [3] stated Web scraping is a quite important methodology used to produce structured data based on the unstructured data available on the internet.

This research focuses on a summary of the data extraction process of web scraping, various web scraping strategies and most of the latest tools utilized to scrap web. They concentrated on the Web scraping techniques.

Ingolf Boettcher [4] discovered that that Web scraping innovation provides a range of choices and can satisfy various purposes: A web crawler's basic requirement is to automate the normally physical work of gathering price estimates and website article details. A web crawler's ultimate requirement will be to discover previously inaccessible pricing data outlets and include a census of all web-available price information. The actions to build web scraping for price analytic include significant analytical and administrative consequences.

## 4. Analysis of various Web Scraping Techniques

### 4.1 Selenium, Puppeteer, Beautiful-Soup

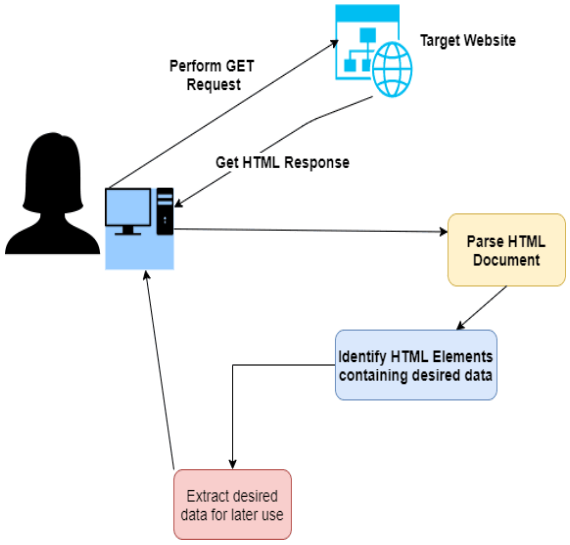
These are the open-source tools that automate web browsers. They provide interfaces that let you write test scripts in various programming languages. To scrape data we must create a web-driver instance, Navigate to a webpage. Locate a web element on the webpage via locators. Perform one or more user actions on the element and record results. These are very efficient techniques for complex websites which may require user input(Authentication etc.). Although very tedious and time-consuming work which requires good knowledge of running scripts.

### 4.2 Text-based-Extraction

Typically this is done by parsing the HTML DOM and extracting XPath or CSS path tags, each corresponding to some component of the page. Information contained in these elements is then typically used to fetch additional pages or data itself. This Technique builds a list of all tags on a page and associates them with their text. When given a user-defined ruleset for links/buttons to click and forms to seek, It can perform a tree search of a site, interacting with elements and downloading documents/pages as the site is traversed. This enables the extraction of data without any prerequisites. This is just like automated manual copy-paste, with pagination enabled.

Although not very useful when user input is required to move forward to scrap.

## 5. METHODOLOGY:




Architecture

### 5.1 Getting data using wanted-list:

On entering the URL from which data is to scrape, the wanted list(i.e. similar data that needs to be extracted) is provided by the user. then on submission, the data is scraped by generating a set of rules. Then these rules are used to extract the data and shown on UI in the form of a table.

### 5.2 Get Trained data:

Often what happens on a site is that the same text is repeated at different places making it ambiguous which text should be extracted, so redundant data is extracted. Here it allows the user to provide multiple keys to train the program to get the desired result.

### 5.3 Get Paginated data:

If the user wants to scrape the data on a site having multiple pages, it stores a set of rules represented as a list based on the wanted-list, which can be reused for scraping similar elements. So the user just needs to provide the page URLs needed to be scrapped.

## 6. RESULT:

We considered various Scraping Techniques (Selenium, Beautiful-Soup, Text-Based, etc.) and found out the efficiency of these techniques basedon

1. Complexities, regular changing structures,and input requirements of Websites.
2. Knowledge of the user i.e history inwriting scripts.

We found that using any of these script-based techniques is more efficient on Websites of highercomplexities and where user input is required to move forward. And Text-based-technique is moreefficient in cases where

1. Websites change their structures frequently and are less complex.
2. The user has no prior knowledge of writing scripts in any programminglanguage.

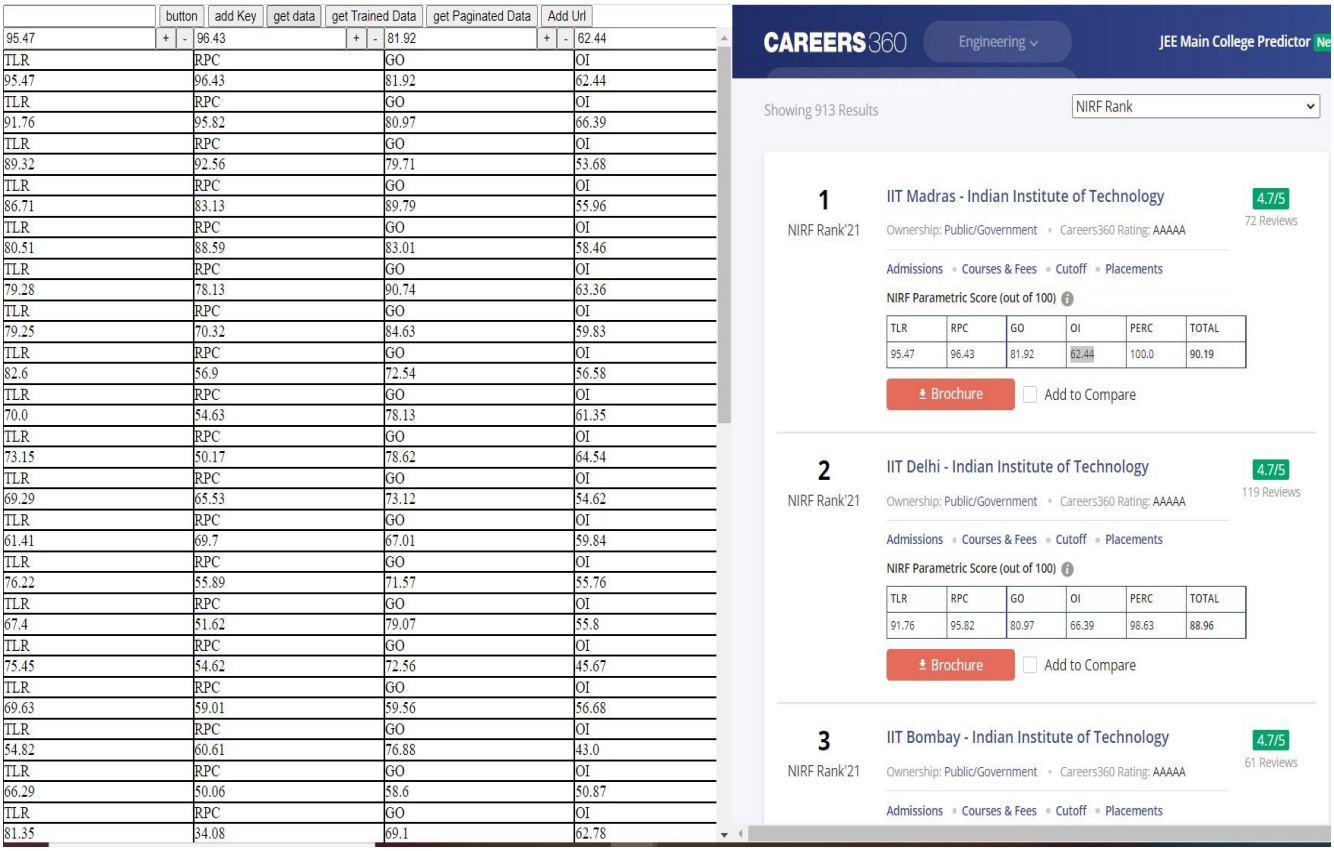



Figure 1 : Getting data using wanted-list

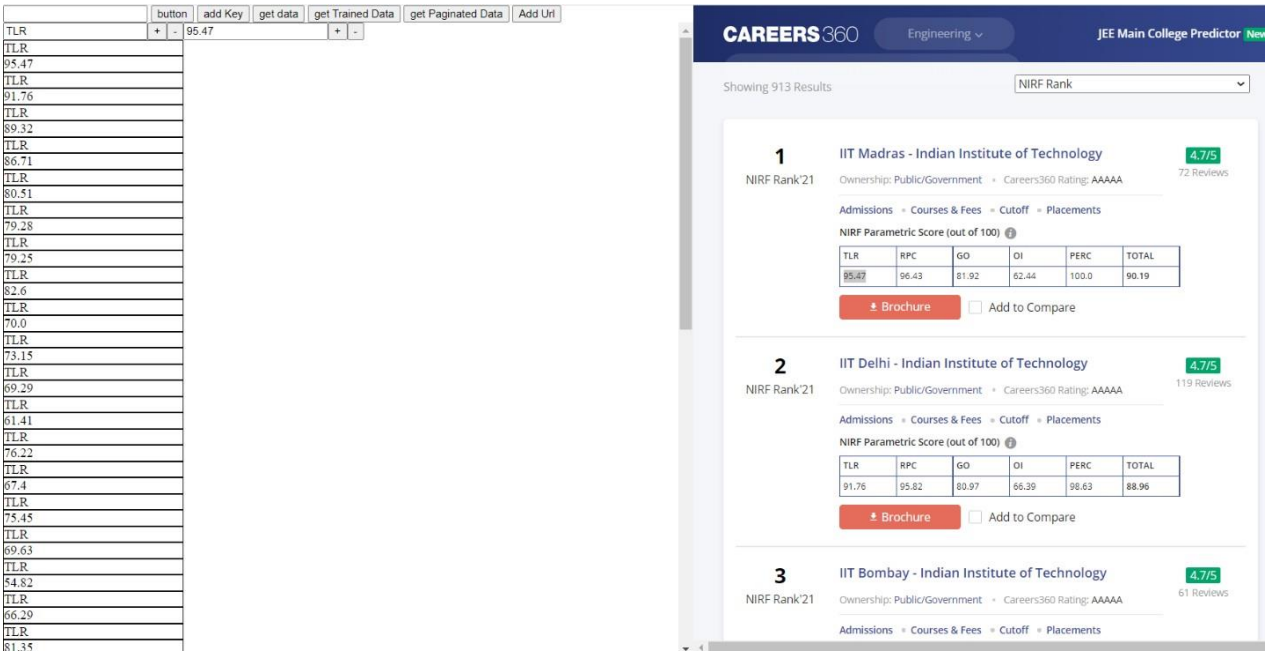



Figure 2 : Get Trained data

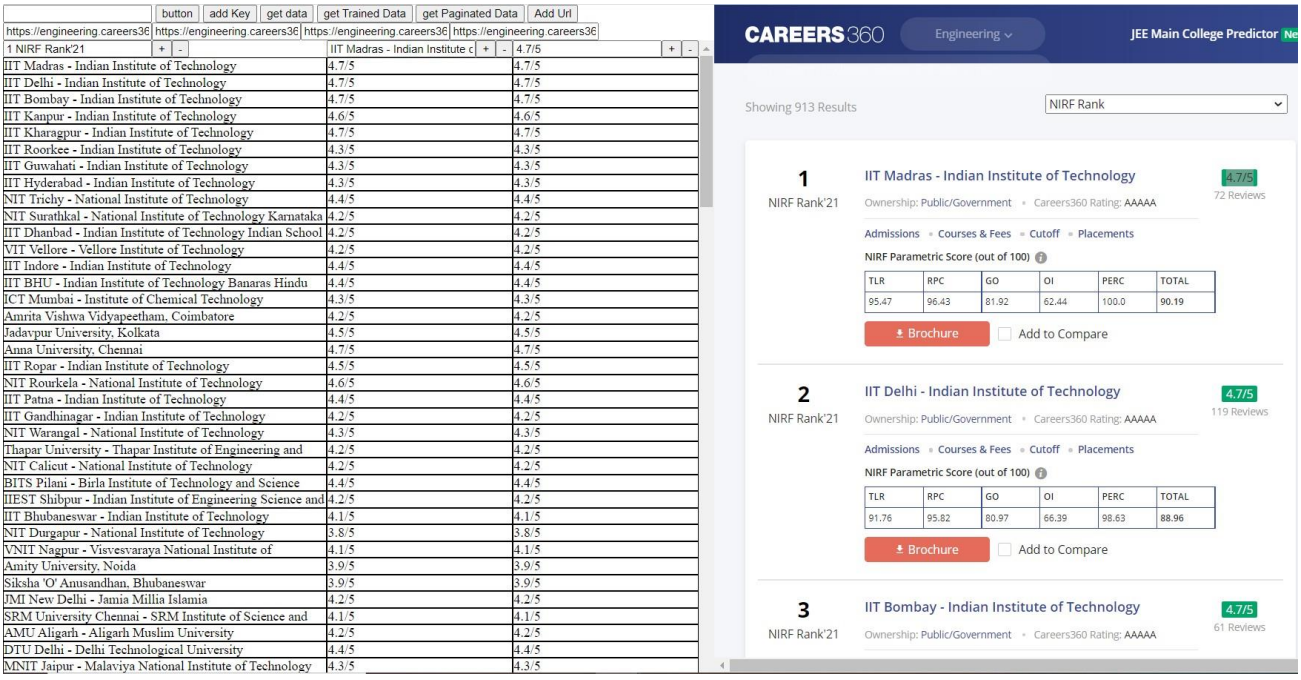



Figure 3 : Get Paginated data:

### 5.1 Hardware Requirements:

- 4 GB Ram.And Dual core processor for smooth functioning of the application.

### 5.2 Software Requirements:

- Windows OS, Mac OS
- Google Chrome, Safari, Windows Edge, Firefox

## 6. CONCLUSION:

As this research paper discussed different methods for web scraping using selenium, puppeteer, beautiful soup, text-based extraction and more. It allow us to understand that for simple sites and for beginners , classical copy paste is good for small data but for regular changing-sites text-based extraction, while for complex sites, we have to rely on writing manual scripts requiring skilled personal selenium, puppeteer are highly recommended for coders. So, for non-coders to ease the scraping we proposed faster classic approach.

## REFERENCES:

[1] S.C.M. de S Sirisuriya,2015, A Comparative Study on Web Scraping .Proceedings of 8th International Research Conference, KDU.

[2] Sameer Padghan, Satish Chigle and Rahul Handoo, "Web Scraping-Data Extraction Using Java Application and Visual Basics Macros," Journal of Advances and Scholarly Researches in Allied Education, pp. 691-695, Vol.15, 2018

[3] Anand V. Saurkar, Kedar G. Pathare and Shweta A. Gode, "An Overview On Web Scraping Techniques And Tools," International Journal on Future Revolution in Computer Science & Communication Engineering, pp. 363-367, Vol. 4, 2018.

[4] Ingolf Boettcher, "Automatic data collection on the Internet," pp. 1-9, 2015.7.