IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, June 2022
Impact Factor: 6.252

Review on Web Scraping

Alen Geo Alex[1], Nakul S Kumar[2], Prithwiraj Prakash[3]
Student, Santhigiri College of Computer Sciences, Vazhithala, India[1,2,3]

Abstract: *Web scraping (also known as Screen Scraping, Web Data Extraction, Web Harvesting, and so on) is a method of extracting publicly uploaded data, processing it, and constructing it into plain or processed data. The result may then be applied to a variety of domains such as Artificial Intelligence, Machine Learning, Market Evaluation, and so on.*

Keywords: Web, Wanderer, JSoup, Scraping, Browser, Data, Webpage, etc.

I. INTRODUCTION

Web Scraping is a technique for automatically extracting large or small amounts of data from websites. The majority of this information is unstructured HTML data that can be converted to structured data in a spreadsheet or database before being used in various applications. Web Scraping can be used in a variety of ways to collect data from websites.

There are several options, including using internet services, specific APIs, and even writing your web scraping code from scratch. Many huge websites, such as Google, Twitter, Facebook, Stack Overflow, and others, provide APIs that let you access their data in a structured fashion. This is the greatest option, but there are alternative sites that either doesn't allow users to access big volumes of data in a structured format or aren't technologically advanced enough. In that case, scraping the website for data with Web Scraping is the best option.

II. A BRIEF HISTORY OF WEB AND WEB SCRAPERS

Web scraping is currently prevalent throughout the business. It became one of the most popular and favored approaches on the web in barely thirty years. Although the use case is mainly connected with web-data extraction the developer's initial implementation was for a different purpose, today it is one of the most effective and straightforward ways of transferring massive amounts of raw data across the globe via the internet. These three decades of adaptation did give us the web scraper that we see today.Here is a small timeline that shows how it originated and how it is now (Figure 1).

Figure 1: History of Web Scrapers

A. Root: World Wide Web

The first forms of web scraping may be traced back to 1989 when British scientist Tim Berners-Lee invented the World Wide Web.

"In those days, there was different information on different computers, but you had to log on to different computers to get at it. Also, sometimes you had to learn a different program on each computer. Often it was just easier to go and ask people when they were having coffee…",

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, June 2022
Impact Factor: 6.252

Tim felt he saw a solution to this dilemma, one that may have far-reaching implications. Already, millions of computers were linked together via the rapidly expanding internet, and Berners-Lee realized they might share information by utilizing a new technology known as hypertext.

With the advent of the World Wide Web emerged three critical components for the use of web scraping.

- The URL, or Uniform Resource Locator, on which we locate and uniquely identify a webpage.
- The embedded hyperlink, allows scrapers to traverse across the webpage.
- The webpage itself, contain both raw and processed data such as audio-video text, etc.

B. Web Browsers

Tim Berners-Lee continued his work two years later, creating the first web browser, operating on a server from his NeXT computer, allowing users to view and interact with the World Wide Web. Web browsers are the platform that served as a link between the sophisticated web and the user. It is because of this element, as well as its simplicity, that the world has migrated to this new technology; without it, there may not even be a need to progress further.

C. The Wanderer and Wandex

The original crawling concept was born in 1993. More specifically, the Wanderer. The World Wide Web Wanderer, created by Matthew Gray at the Massachusetts Institute of Technology, was a first-of-its-kind Perl-based web crawler whose main aim was to calculate the size of the internet. Soon Wandex was developed to create the index of the available web pages.

D. JumpStation

Have you heard of search engines? JumpStation, the technology that paved the way for such search engines, was born in 1993. It was one of the earliest crawler-based search engines, indexing millions of online pages and offering material in response to user queries.

E. BeautifulSoup

Almost a decade later, in 2004, BeautifulSoup - HTML Parser was launched. Written in Python, it became one of the most popular web scraping libraries. BeautifulSoup introduced the idea of analyzing HTML layouts and modeling such structures into individual components. During the era when the web was at its peak, millions of files were posted to it, and manually processing them would be inconvenient. Things started changing within the BeautifulSoup idea!

F. Rise of Web Scrapers

Web scraping as we know it was born soon after. Stefan Andresen's visual web scraping software Web Integration Platform version 6.0 allows users to highlight the relevant information of a web page and arrange that data into a useable excel file or database, allowing non-programmers to join and simply extract data from the web.

III. THE WORKING OF WEB SCRAPERS

Web scrapers' operations can be classified as simple to advanced. Web sites are built and developed to make people's lives simpler. Companies invest millions of dollars in UI/UX designs to make websites more engaging to end-users. The complexity of online scrapers will gradually increase as new technologies and languages emerged.First and foremost, we must inform the web scraper of what is to be scrapped. This is commonly accomplished by supplying the URL, or Uniform Resource Locator. A URL is simply a string or a group of letters. A URL, in principle, must be unique and refer to specific resources over the internet. The resource might be an API endpoint from which data is requested, an HTML page, a video stream, and so on. This is an example of a normal URL (Figure 2).

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, June 2022
Impact Factor: 6.252

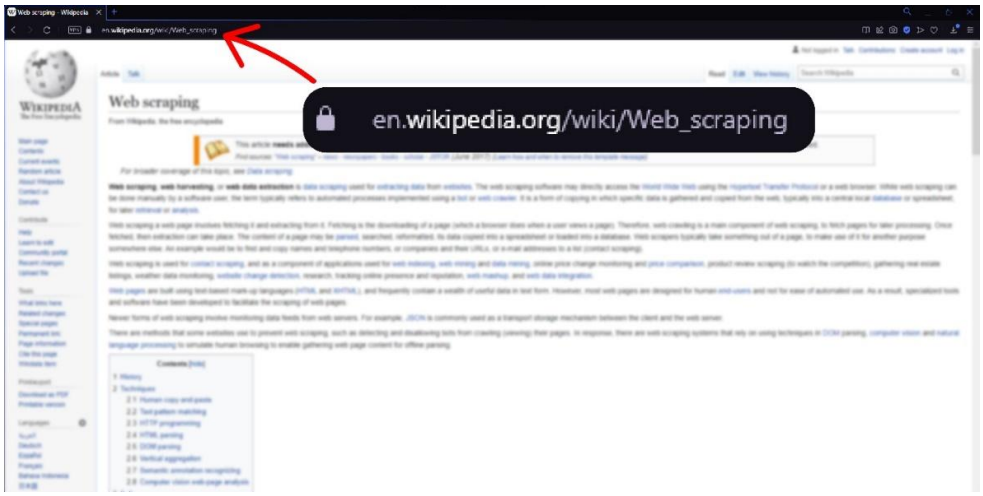

Figure 2: An Example of a Uniform Resource Locator or (URL)

By supplying the stated URL, we instruct our scraper to go through the internet, get the said web page, and prepare it for scraping. After that, the scraper would take the given data and model it into its format. This is dependent on the parser library handling the data. Once the data is in a format that the parser understands, it may begin to extract all of the data on the page or the one that the user requested. Ideally, one should have a reason for scraping the data, and the user should then indicate what they want to be scraped from those web pages. For example, most web pages contain headers and footers that provide information such as contact information, email addresses, and so on. If the user desires to extract data about a product stated on the web page, he or she will not require the above-described data.

The last section describes how to present the scraped data in a useful way. There are several common choices for this. Some of the most prevalent methods include Excel sheets, CSV documents, JSON, and so on. Ultimately, this aspect should be significantly dependent on the user. If a user needed to gather data, he or she would probably save the scraped data to a storage bucket anywhere in the world. If the intention for scraping out the is to be deployed as an API endpoint, delivering an Excel spreadsheet would be less convenient; instead, he or she would use structured formats such as JSON or XML.

IV. WHY USE WEB SCRAPERS?

In terms of technology, the twenty-first century is magnificent. Every project, regardless of size or scalability, requires some level of automation. It is not viable to upload files manually when an automated system can do it in a fraction of the time it takes to do it manually. The same holds true for data harvesting. The use of scraping technologies offers the benefits of scalability and flexibility.

In reality, one can spend money and effort manually copying and pasting contents or information from websites into an excel sheet or a document type. The manual procedure is time-consuming, tiresome, and error-prone. The fact that an automated web scraping technology can achieve the same thing for a fraction of the investment while producing high-quality and accurate data outweighs the need to use the manual human technique.

V. IS THERE A LEGAL ISSUE IN USING WEB SCRAPING

You can scrap public data, that is anyone with an internet connection can access those data. Web scraping is considered illegal when one tries to extract data that is private. The controversy regarding the legality of web scraping began when people started using automated bots for scraping webpages. The usage of automated bots started to be a menace. Over time companies started to deny one's ability to scrap the data out, things started to span out and matters were taken out to the public courts. Some instances of this are:

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, June 2022
Impact Factor: 6.252

A. HiQ Labs vs LinkedIn

HiQ Labs is a data analytics firm that focuses on workforce data and people analytics. Their analysis providesinsights for their clients about specific industries. HiQ Labs scraped data from public LinkedIn profiles as one of the ways they gathered data for their insights.LinkedIn responded by preventing HiQ Labs' tools from accessing this publicly available data and issuing a cease-and-desist letter. They claimed that HiQ Labs' actions were illegal under the Computer Fraud and Abuse Act (CFAA). HiQ fought back in 2017 by filing a lawsuit and obtaining a preliminary injunction. HiQ's claims that accessing publicly available data was not a violation of the CFAA were found to be "likely to succeed" by the district court.

B. Craigslist

Craigslist filed a lawsuit in 2017 against several startups (including Padmapper) that scraped Craigslist data to support their services. The defendants were concerned after the case was not dismissed by the trial court. As a result, the case was settled without the need for a trial. Because of the hiQ Labs vs LinkedIn case, cases like these are likely to become less common.

How can one determine if data on the internet is considered public?

- The user who posted the information chose to make it public.
- To access the data, a user does not need to create an account or log in.
- Web scrapers and spiders are not blocked by the website's robots.txt file.

VI. APPLICATIONS OF WEB SCRAPING

Profit in a company is determined by the selection made from the available options. In this day and age, most corporations make decisions based on the data at their disposal. As a result, web scraping or data harvesting would be a vital phase for such businesses. Regardless of industry, the internet as a whole delivers actionable insights and decision-making facts. This data, if used correctly, may create a competitive edge.These are some of the industries that are currently relying upon web scraping for such information:

A. Customized Web Scraping Tools

Many businesses are now offering customized web scraping tools to their customers, which collect data from all over the internet and organize it into useful and understandable information. It saves time and money by eliminating the need to manually visit each website and collect data.

B. E-commerce System

Web scraping is a useful tool for determining a product's price in an eCommerce store.The majority of businesses are formulating strategies based on data scraped for digital monitoring of a competitor's website. In the coming years, this trend will be amplified.

C. Data Analysis

You might want to gather and analyze data from multiple websites about a specific category. Real estate, automobiles, electronic devices, industrial equipment, business contacts, and marketing are examples of possible categories. The information on the various websites that belong to the specific category is presented in a variety of formats.Even with a single website, it's possible that you won't be able to see all of the information at once. Under various sections, the data may be spread across multiple pages (similar to Google search results, which is known as pagination or paginated lists). You can extract data from multiple websites into a single spreadsheet (or database) using a Web Scraper, making it easier to analyze (and even visualize) the data.

D. Real Estate

Web Scraping software can be used to extract property details from real estate websites such as Zillow, Realtor, and others. In addition to property details, web scraping can be used to scrape agent and owner contactinformation.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, June 2022
Impact Factor: 6.252

E. Price Comparison & Competition Monitoring

Companies that provide products or services must have detailed information about competitors' products and services, which appear on the market daily. To keep a constant eye on this data, web scraping software can be used.

F. News Monitoring

Web scraping news sites can provide a company with detailed reports on current events. This is especially important for businesses that are frequently in the news or rely on daily news for their day-to-day operations. After all, a single day's news can make or break a company.

G. Sentiment Recognition

Sentiment Analysis is a must for companies that want to understand the general sentiment for their products among their customers. Companies can use web scraping to gather information about general sentiment about their products from social media sites like Facebook and Twitter. This will assist them in developing products that people want and gaining a competitive advantage.

Scraping data from social media or other forums with comment sections is primarily used for this one.Recognition of Emotions Algorithms detect patterns and detect hints that go beyond your tweet. It can infer information about you based on your location or the phone you used to tweet. If it weren't for website scraping, this branch of machine learning would be rendered useless, and all research would come to a halt. Gone are the days when tweets were grouped and logistic regression was performed based on the smileys in them or the hashtags that followed them. Even the difference between a passive and active voice can be detected, and machines can infer information about your personality and nature from your Facebook or Twitter activity.

H. Improving Image Recognition Algorithms

SURF and SIFT were created in 2006 and 2010, respectively, and are still the most popular algorithms for detecting image similarities. The race, however, is far from over. The search is on for an algorithm that will not only look at pixels but will also have something to say based on previous experience (the data that it has already gone through). Images are readily available and frequently include tags, allowing you to quickly create a labeled dataset. So, whether you're trying to write your first algorithm to distinguish cats from dogs or running one to distinguish between satellite images with forest fires andthose without, you can easily get your data if you crawl it off the web. The internet is by far the most extensive and nearly limitless repository of images. In the case of images, the more you train, the closer your machine gets to detecting a pattern that no human brain can deduce.

VII. EXAMPLE USE CASE

We have seen many components of Web scraping throughout this article, but not a real testing situation.

Here are some examples of web scraping applications. The fundamental goal of these use cases is to demonstrate three key objectives.

- The simplicity of the use case
- Why should one use web scraping? What is the alternative solution left?
- Availability of different libraries.

We will examine the following objectives in addition to the stated use case. This goal demonstrates how popular and simple web scraping is for obtaining data from enormous amounts of raw data.

Scraping tabloid notification of a webpage

In this use case, we would like to demonstrate the use case of scraping individual text in a notification tabloid shown below on (Figure 3).

1. Objective

- Simplicity: - The entire project from start to end only took just around 30 lines of code. This includes the boilerplate codes that are included in the programing language.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, June 2022
Impact Factor: 6.252

- Why should one use web scraping: - In this case, there is no other option rather than scraping the data out, as the said webpage is not providing a valid API endpoint to request or collect data from.
- Library Used: - This snippet of code is based on JSoup 1.14 and Java 17

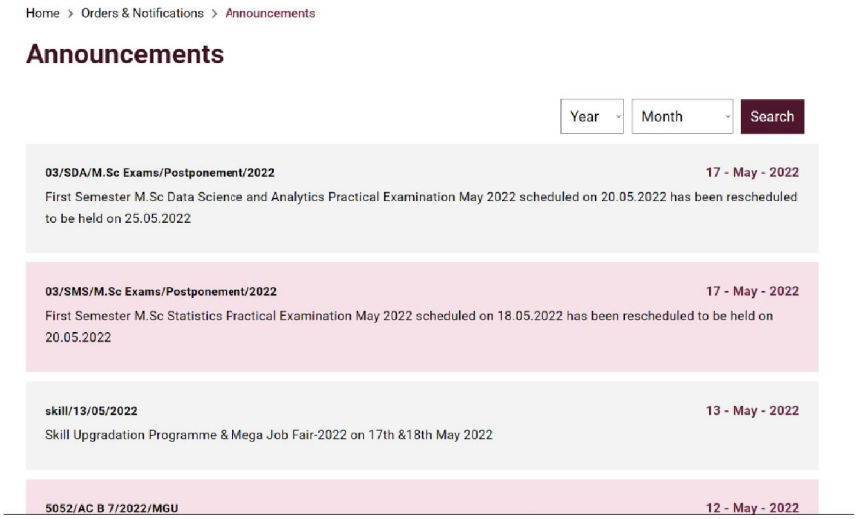

Figure 3: Notification Webpage

2. Code Snippet

Figure 4: Code Snippet

3. Working

- Initially, we add all the imports that are related to the project.
- We also provide the URL for the website to be scraped into a string variable called ANN_URL

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, June 2022
Impact Factor: 6.252

- Inside the main function, JSoup is instructed to download the web page from the provided url and convert it into a JSoup object (Document).
- By analyzing the code structure of the webpage, we came to know that all the tabloids are inside the HTML class order-listing. With this information known, we tell JSoup to get all the elements inside the said class.
- Once that is done, JSoup provides an iterative list of Elements. We iterate through them and print all the necessary information we needed out to the console.

4. Result

```
--------------------------------
5052/AC B 7/2022/MGU
12 - May - 2022
CAT Entrance Examination 2022 for admission to Inter School Centres & Departments-Time Table -Order issued
https://www.mgu.ac.in/uploads/2022/05/5052-AC_B_7-2022-.-.-cat.pdf?x91270
--------------------------------
4968/AC B 2/2022/M.G.U
11 - May - 2022
SNM college maliyankara- appointment of DDO- extension-order issued
https://www.mgu.ac.in/uploads/2022/05/4968-AC_B_2-2022-.-.-.pdf?x91270
--------------------------------
4860/AC A 1/2022/MGU
10 - May - 2022
Affiliated Arts and Science/ Teachers Training Colleges-Instructiors regarding formation of Admission committee-Order issued
https://www.mgu.ac.in/uploads/2022/05/4860-AC_A_1-2022-.-.-.pdf?x91270
--------------------------------
4874/AC A 1/2022/MGU
10 - May - 2022
Implementation of the Awards for Excellence - Sanctioned- Order issued.
https://www.mgu.ac.in/uploads/2022/05/4874-AC_A_1-2022-MGU.pdf?x91270
--------------------------------
```

Figure 5: Result Snippet

VIII. CONCLUSION

This paper from the beginning acknowledges Web Scraping and the different aspects surrounding it such as its history, its use, and the different applications it possesses. Web Scraping is one of the easiest ways of extracting a large amount of data from a website without actually having to do any of the tedious manual labor. This allows people to scrape data for either individual or commercial purposes, the growth of Web Scraping surrounding the past few times was tremendous as it holds many possibilities and its scope for the future is very prominent as well.

There are many different libraries over many programming languages that support Scrapers, some of them include Jsoup, BeautifulSoup, Scrappy, Selenium, etc. Data Scrapers allow data collection on different fields associated with computers such as Machine Learning, Robotics, A.I, etc. But if these data are grabbed manually then it's going to take a lot of time to extract, hence using Web Scrapers we can easily grab these data within a matter of time.

REFERENCES

[1] Al Sweigart, Automate the Boring Stuff with Python, 2nd Edition: Practical Programming for Total Beginners 2nd Edition.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, June 2022
Impact Factor: 6.252

[2] Michael Heydt, Python Web Scraping Cookbook: Over 90 proven recipes to get you scraping with Python, microservices, Docker and AWS.
[3] Ryan Mitchell, Web Scraping with Python: Collecting More Data from the Modern Web 2nd Edition.
[4] Bart Baesens, Practical Web Scraping for Data Science: Best Practices and Examples with Python 1st ed. Edition.
[5] Jaime Buelta, Python Automation Cookbook: 75 Python automation ideas for web scraping, data wrangling, and processing Excel, reports, emails, and more, 2nd Edition.
[6] Olgun Aydin, R Web Scraping Quick Start Guide: Techniques and tools to crawl and scrape data from websites.
[7] SeppevandenBroucke, Practical Web Scraping for Data Science: Best Practices and Examples with Python 1st ed. Edition.
[8] https://webscraper.io/blog/brief-history-of-web-scraping
[9] https://www.parsehub.com/blog/web-scraping-legal/
[10] https://hevodata.com/learn/8-best-web-scraping-tools/#topWST
[11] https://en.wikipedia.org/wiki/Web_scraping