ISSN 2278-3091

Volume 9, No.5, September - October 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse185952020.pdf

https://doi.org/10.30534/ijatcse/2020/185952020

Web Scraping: Applications and Scraping Tools

Priya Matta[1], Nikita Sharma[2], Devyani Sharma[3], Bhasker Pant[4]
Sachin Sharma[5]
[1]Graphic Era Deemed to be University, India, mattapriya21@gmail.com
[2]Graphic Era Deemed to be University, India, nikkz0524@gmail.com
[3]Graphic Era Deemed to be University, India, devyani51sharma@gmail.com
[4]Graphic Era Deemed to be University, India, pantbhaskar2@gmail.com
[5]Graphic Era Deemed to be University, India, sxsharma88@gmail.com

ABSTRACT

The world of Artificial Intelligence and machine learning has its common roots with data, which is primarily the most important entity on its own. Data has already impacted so many businesses worldwide and can never take a back seat when it comes to this technical world. To get access to data in its best form, web scraping was brought to use. Data provided on the internet is of so much use that the whole world is running after it. Web scraping was brought into practice long back and is still useful to date. This paper aims to make people aware of this technology to help them expand their knowledge. Tools and applications related to web scraping are also mentioned.

Key words: Artificial Intelligence, Machine Learning, Web scraping, Web scraping tools.

1. INTRODUCTION

In today's era, every emerging paradigm is giving birth to a generation of a huge amount of data. Whether the data is in text, image, audio, video, or in any other raw or furnished form, it always acts as an important resource, so in the case of e-commerce. 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.” You can view the data on your competitor’s website. Now, the question arises how will you download it in a usable format? Most people prefer to copy and paste it manually. However, it’s not a feasible way to do it when concerning large websites with hundreds of pages. So, this is where web scraping plays its role. It is a process of automating the uprooting of data adequately and actively, no matter what is the amount and size of the data, on your computer. Moreover, there are web sites that do not permit to copy and paste the data. here, Web scraping serves as the best technique to extract any kind of data needed. That's not enough. Let’s assume, you copy and paste some useful data but how will you turn it in a format of your preference? Web scraping works on that too. It helps to save the data in a particular format, mostly CSV, hence You would then be able to retrieve, examine, and utilize the data the style you desire. Although the most common format is CSV, many web scraping techniques and tools produce the data in an excel sheet too. Other than these two formats some advanced formats are also supported by modern web scrapers, namely JSON, which is more advantageous as it supports API also. So, web scraping clarifies the process of deriving data, promotes it up by automating it, and generates easy access to the scrapped data by providing it in the desired format. Some websites contain a large amount of data regarding stock prices, company contacts. The extraction of data if required is a very tedious process if done manually by either using the way the website allows or copy every piece of information. To make that task a bit easier, we use web scraping.

This paper comprises of six sections. After the introduction in section 1, we have discussed the various definitions of web scraping, proposed by various academicians and practitioners. In section3 we have presented the motivation behind the work. In section 4, we have discussed various applications of web scraping. In the next section, i.e. section 5, we have presented the available tools for web scraping. Finally, in the sixth section, we have concluded our paper.

2. DEFINITIONS

As the scope of the world wide web keeps on diversifying, there is a demand for different sets of approaches that will boost the entire network related to market, businesses, and even our day to day lives. Businesses need to expand themselves to survive in the market. The world takes the help of Data Extraction and Data Analysis to compete, web scraping being a part of it. The extraction of data if required is a very tedious process if done manually by either using the way the website allows or copy every piece of information. To make that task a bit easier, we use web scraping. According to Wikipedia, web scraping can be interchangeably termed as web extraction or harvesting, It defines web scraping as “a technique to extract data from the world wide web (www) and save it to a file system or database for later retrieval or analysis.” Web Scraping can be so helpful in this era where we need data retrieval. According to Diouf et al. [2], “The main objective of Web Scraping is to extract information from one or many websites and process it into simple structures such as spreadsheets, databases, or CSV files.” Web scraping is performed both physically as well as through software that provokes human web surfing to collect specified information from websites. Web scrapping has drawn a lot of controversies as some web-sites don’t allow certain kinds of data mining. Still, overall web scraping promises to become an approach followed worldwide for data extraction. As mentioned in Apress [3], “Sometimes it is necessary to gather information from web sites that intend for human readers, not software agents. This process is known as web scraping.”

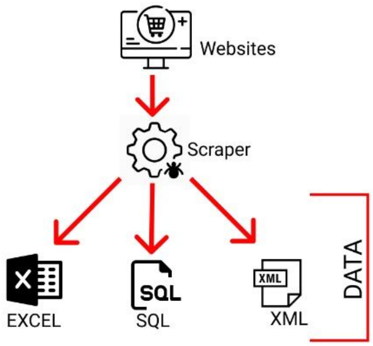


Figure. 1: The procedure of Web-Scraping.

Web scraping or web harvesting is a process of extraction of data from websites to get some useful information out of it. The data that is extracted is exported into a useful format ie. a spreadsheet. Web scraping can be done both manually and by software that provokes human web surfing to collect specified information from websites. Some researchers, Saurkar et al. [4] proposed their view about web scraping as,” scraping is a technique used to crop information from web pages based on script routines” according to them, those documents are either written in hypertext mark-up language (HTML) or XHTML. The way of representing these documents are generally hierarchical based document object model, or simply the dom tree. According to Chaulagain et al. [5], “web scraping is one of the major sources for the extraction of unstructured data from the internet, we have analyzed the scraping process when introduced to a bulk of data extraction.”

According to Techopedia,[6] “web scraping is essentially a form of data mining.” Other practitioners [7] define web scraping as, “web scraping, also known as web data extraction, is the process of retrieving or scraping data from a website. unlike the mundane, mind-numbing process of manually extracting data, web scraping uses intelligent automation to retrieve hundreds, millions, or even billions of data points from the internet’s seemingly endless frontier.” Vargui et al. [8] outline that, “web scraping is the set of techniques used to automatically get some information from a website instead of manually copying it.” according to them, the basic aim of available web scrapers in the market, is to choose and pick the required data and information and then to cumulate that into new web pages.

Mitchell [9] explains the web scraping as, “web scraping bridges the gap between human-understandable and machine-readable data and opens up a new world of data to researchers by automatically extracting structured data sets from human-readable content”. Boeing et al. [10] defined the web scraper as, “a tool that accesses web pages, finds specified data elements on the page, extracts them, transforms them if necessary, and finally saves these data as a structured data set.”

3. MOTIVATION

Information is the most important asset in the world, but for retrieving it we need data. Data being the second important asset is not accessible to all the people around. Everyone can’t get access to data which they require, for this purpose web scraping come up to the surface. Web Scraping has entirely shifted the way we used to see this world with less amount of data. Analysis and Retrieval have become so easy as of now. Our life wouldn’t have been the same without web scraping. Many businesses have been able to skyrocket because of Web scraping as the collection of leads was made possible with it. The process of gathering unstructured data on the web is an interesting area within many contexts whether it be for business, scientific or personal usage. According to Mazlin et al. [11], “selecting optimum significant features from high dimensional data may produce a challenge especially to an overfitted data that consequently resulting to data dimensionality issues.” The advertisement business relies on the directed advertisement which is distributed throughout several pages, for the service to understand its current context, web data extraction, and web wrappers can be used in conjunction with content analysis tools for contextual analyses of the current page. In science, data sets are shared and used by several researchers and often publicly publicized. In some cases, the data sets are provided through a structured API, but often data is only accessible through search forms and HTML documents which call for web wrapping methodologies to be used. Personal use has also grown as services have started to emerge which provides users with tools to mashup components from different web pages into own collection web pages [12]. In general Web, scraping offers so many benefits, some are listed below:

- Research
- Enhancement of businesses.
- E-commerce
- Educational Purpose
- Information Retrieval

There has been a wide usage of Web Scraping for E-Commerce websites as it helps to know the status of the competitor's page, leading to a growth in the marketplace.

4. APPLICATIONS

Every emerging paradigm gains its value with an increase in its applications. Similar is the case with web scraping. Many academicians [13,14,15,16,17] 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.

4.1 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.

4.2 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.

4.3 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. Here Sentiment Recognition Algorithms sense hints and detect patterns that also go beyond your tweet itself. It can conclude by using your location of the phone. This is one branch of ML that would be declared useless, and every research would discontinue if not for web-scraping. Those times are gone when tweets used to be grouped, and one would apply logistic regression based on the smileys detected in it, or the hashtags following the data statements. Now, it could even sense the variation between a passive and an active voice and make opinions about your personality through your activities on Facebook, Instagram, Twitter, etc. and this all made successful by web scraping.

5. 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; some are mentioned below:

5.1 SCRAPER API

Usually, the data that is extracted is in tabular form. The tools mentioned above, as well as the programs, interact and retrieve the webpages with the help of some API or commands, also known as application programming interfaces. Scraper API is a tool that lets the user build web scrapers by handling proxies and CAPTCHAs to get the raw HTML quickly. It is helpful in social media scraping, search engine scraping, and more.

5.2 FMINER

This tool is helpful in so many areas and has a variety of techniques by which we can quickly master our data mining game. It is an effective visual design tool that captivates each action; models a map that can communicate with the target site to get the desired information.

Characteristics/features of FMiner:

- Visual design tool
- Coding is not required.
- Boost of advanced features.

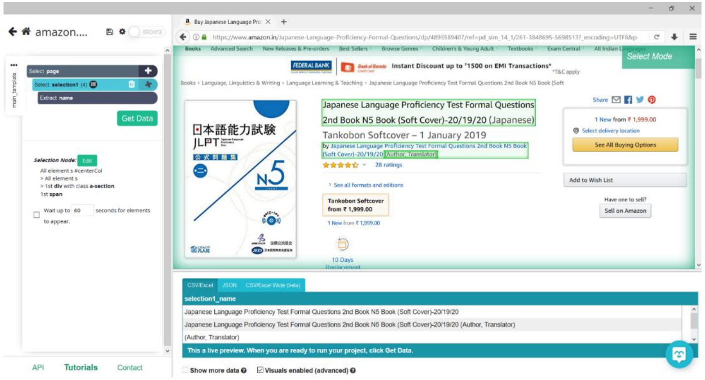
- Available with multiple path navigation options.
- Keyword Input List
- Nested Data Elements
- Multi-Threaded Crawl
- Export Formats
- CAPTCHA tests

5.3 OCTOPARSE

It's a tool that lets people who don't want to code use web scraping at its best use. It features a point and clicks web scraper making the user experience web scraping at its best form. It allows users to scrape all the types of structures, render JavaScript, and so much more. It is also available with a site parser and a hosted solution for cloud-based scraping. With all of that, It also has a feature that lets the user build up to 10 crawlers for free, making it a suitable fit for users who want to access the data efficiently.

5.4 PARSEHUB

ParseHub is an incredible tool that lets us scrape any interactive website. It is quite easy to use and offers a wide range of options or filters for our relevance. ParseHub has fulfilled a better answer for every data requirement. Darcy Byrne, CEO at Fruitbat said and as quoted "It's simple API has allowed us to integrate into our application.". during the origin of web-scraping, the user used to merely select the section whichever he/she wants to retrieve, then using the ParseHub tool it automatically chooses similar data elements from different sites. For picking other information from any targeted website, a 'relative' search option is available as a subset of information about the previously select item. similarly, the users derive all data from the web. At the time of the extraction of elements from any particular site, ParseHub provides a URL which is an optional field. Following prosperous web scraping data, collections are stored in a CVS format as given below:



Select page
Select selection1 (4)
Extract name
Get Data

Selection Node: Edit
All element s #centerCol
> All element s
> 1st div with class a-section
1st span

□ Wait up to 60 seconds for elements to appear.

https://www.amazon.in/Japanese-Language-Proficiency-Formal-Questions/dp/4893589407/ref=pd_sim_14_1/261-3848695-5698513?_encoding=UTF8&p

CSV/Excel | JSON | CSV/Excel Wide (beta)

selection1_name
Japanese Language Proficiency Test Formal Questions 2nd Book N5 Book (Soft Cover)-20/19/20
Japanese Language Proficiency Test Formal Questions 2nd Book N5 Book (Soft Cover)-20/19/20 (Author, Translator)
(Author, Translator)

This a live preview. When you are ready to run your project, click Get Data.

□ Show more data ☑ Visuals enabled (advanced)

amazon....
BROWSE
main_template
Buy Japanese Language Pro
Select Mode
Japanese Language Proficiency Test Formal Questions 2nd Book N5 Book (Soft Cover)-20/19/20 (Japanese)
Tankobon Softcover – 1 January 2019
See All Buying Options
Add to Wish List
Sell on Amazon
API
Tutorials
Contact

Figure. 2: Showing the functioning of ParseHub [13]

5.5 SCRAPY

Scrapy is a popular open-source tool and is free of cost. It is a python library that helps in web-scraping. Scrapy is so much in use these days. The procedure is as follows- "First, a user navigates to a page that he would like to scrape and generates a template for the content that he would like from that page. Next, the user selects a set of links that point to pages matching the content template defined by the user. Finally, the user selects an output data format, and Scrappy crawls the connections specified by the user and scrapes content corresponding to the user's template." [12].

5.6 WEB CONTENT EXTRACTOR

Another Simple and user-oriented tool for data scraping is Web Content Extractor (WCE), which is developed by Newprosoft. It has a useful wizard that guides the user to setup scraper. Users can scrape data from the website with few clicks. WCE is self-intelligent for putting data into different formats like Excel, text, HTML formats, Microsoft Access database, Structured Query Language (SQL) Script File, MySQL Script File, Extensible Markup Language (XML) file, HTTP submit the form and Open Database Connectivity (ODBC) Data source.

6. 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.

REFERENCES

1. Norulhidayah Isa, Nur Syuhada Mohd Yusof and Muhammad Atif Ramlan, "*The Implementation of Data Mining Techniques for Sales Analysis using Daily Sales Data*", International Journal of Advanced Trends in Computer Science and Engineering, Vol. 8, Issue. 1.5, pp. 78-80, (2019)
2. Diouf, Rabiyatou, Edouard Ngor Sarr, Ousmane Sall, Babiga Birregah, Mamadou Bousso, and Sény Ndiaye Mbaye. "Web Scraping: State-of-the-Art and Areas of Application." In *2019 IEEE International Conference on Big Data (Big Data)*, IEEE, (2019). pp. 6040-6042.

3. Watson, Mark. *Scripting Intelligence: Web 3.0 information gathering and processing*. Apress, (2009).

4. Saurkar, Anand V., 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* 4, no. 4 (2018): 363-367.

5. Chaulagain, Ram Sharan, Santosh Pandey, Sadhu Ram Basnet, and Subarna Shakya. "Cloud-based web scraping for big data applications." In *2017 IEEE International Conference on Smart Cloud (SmartCloud)*, pp. 138-143. IEEE, (2017).

6. Bradley, Alex, and Richard JE James. "Web scraping using R." *Advances in Methods and Practices in Psychological Science* 2, no. 3 (2019): 264-270.

7. Landers, Richard N., Robert C. Brusso, Katelyn J. Cavanaugh, and Andrew B. Collmus. "A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research." *Psychological methods* 21, no. 4 (2016): 475.

8. Vargiu, Eloisa, and Mirko Urru. "Exploiting web scraping in a collaborative filtering-based approach to web advertising." *Artif. Intell. Research* 2, no. 1 (2013): 44-54.

9. Mitchell, Ryan. "*Web scraping with Python: Collecting more data from the modern web*. " O'Reilly Media, Inc., 2018.

10. Tengku Mazlin, Tengku Ab Hamid, Roselina Sallehuddin, Zuriahati Mohd Yunos and Aida Al, "*Ensamble Based Multi Filters Algorithm for Tumor Classification in High Dimensional Microarray Dataset*", International Journal of Advanced Trends in Computer Science and Engineering, Vol. 8, Issue. 1.6, pp. 116-123, (2019)

11. Boeing, Geoff, and Paul Waddell. "New insights into rental housing markets across the United States: Web scraping and analyzing craigslist rental listings." *Journal of Planning Education and Research* 37, no. 4 (2017): 457-476.

12. Olofsson, Joacim. "Evaluation of web scraping tools for creating an embedded web wrapper." (2016)

13. 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.

14. 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 Conference (ICT-ISPC)*, pp. 1-5. IEEE, 2018.

15. 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.

16. 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.

17. 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.