Web Scraping Tool For Newspapers And Images Data Using Jsonify

Qingli Niu[1], Irfan Ali Kandhro[2], Anil Kumar[2], Shahnawaz shah[3], Muhammad Hasan[2], Hifza Mehfooz Ahmed[2], and Fei Liang[1*]

[1] College of Information Engineering, Zhengzhou University of Science & Technology, Zhengzhou 450064, China
[2] Department of Computer Science, Sindh Madressatul Islam University, Karachi, Pakistan
[3] Department of telecommunication engineering, University of Sindh Jamshoro, Pakistan
[*] Corresponding author. E-mail: liangfei1989@163.com

Received: Feb. 28, 2022; Accepted: May 08, 2022

Web scraping is the process of extracting data from a website in an efficient and fast way. In such a scenario, python programming can offer useful set of methods that help web editors to improve the quality of the provided service. This scraper contains three steps 1) to understand the structure of web page, 2) design regular expression pattern and finally use that pattern to get certain data. In this paper, we also used Flask, Request, JSONify library to get the data, after processing, the data is transformed into the JSON form and ready for CSV with help of API. After generated all required regex patterns, the system uses these patterns as a set of rules, and with this, designed scraper tool works efficiently, and achieved outstanding results with help of support libraries to storing and extracting the news and web-based information. The proposed Web scraping tool eliminates the time and effort of manually collecting or copying data by automating the process. It is found that this designed scraper is easy and direct approach to extract the newspapers, websites, blogs, and images data.

Keywords: web scraping, extracting, retrieving, Python framework, API, manually collecting data.

http://dx.doi.org/10.6180/jase.202304_26(4).0002

1. Introduction

This research is based on a scraping tool. Through a scraper, we extract our data and save it in CSV format. Web scraping is the process of extracting data from websites through scraping software. Many people and organizations use scrapers, and many companies have now developed their own scraper tools to extract data. Many reasons to use scarper for business marketing and pricing research of their competitors, to monitor all the changes and analyze the data easily [1]. With the help of this tool, extraction of data is the easiest process. Human effort is reduced, and time is also saved. It is an efficient and powerful technique for collect huge amounts of data. Current online scraping techniques have evolved to adapt to several settings, ranging from ad hoc, human-assisted operations to the use of completely automated systems capable of converting entire webpages into well-organized data sets. On the internet, day by day, data increases, so the importance of web scraping is increasing. Many organizations have their own web scrapers that allow them to easily facilitate this [2].

There are so many scrapers and their unique methods, but as compared to web scraping tools, they analyze your data, create API and Data save it in a JSON format. In the JSON format, we have all the information about the targeted Website. Web scraping future scope includes forecasting user needs to improve usability, scalability, and user retention, as well as providing an efficient framework for Web personalization through the effective use of Web Log files. The semantic web is a futuristic vision in which web material can be analyzed and synthesized by automated processes. The majority of content on the internet is

human readable. The browser has a hard time understanding the text because it can only visualize HTML mark-up. Content interpretation, selection, and management are the three most difficult activities on the internet. These three tasks are currently managed by humans. The semantic web will restore the balance between machine and human intelligence by automating three tough activities.

It works like this: first select a website whose data you want to be gathered or extracted, put the URL in the scraper tool, and load the website. Hyper Text Markup Language (HTML)-based web pages, JSON-formatted data feeds, and multimedia such as audio, video, or image extraction process now parse HTML data. All the data is inserted into an HTML tag whose data is wanted through a website using just tags. Data would be gathered in an untidy or unstructured form and then transformed into the required format. It totally depends on your convenience which format you choose, like a document, PDF, or spreadsheet, where you want to store the data for analysis [3].

So many people assume web scraping and crawling are the same thing, but there are some little differences. Web scraping is the act of getting data from a website and putting it into our own database so we can do some analysis. Websites are typically designed to be read by humans to generate html parsing data from a website [4]. For example, we get the price of an item from Amazon and put it inside our database in an automated way, then use a web scraper to scrape the price of that specific item every hour. Maybe the price will change. Build up to running scarpers in an automated manner every hour to monitor the pricing.

Due to the expansion of the internet, it can be difficult to know the address of a web page. Then a search engine comes in, and this process is called a web crawler. Their algorithm works like this: first crawl the entire website, then fetch the website, parse the extracted link, and then repeat the same previous steps [5]. For example, when we look for gold, a search engine (Google) finds or crawls the entire website that belongs to the gold and all URL content is shown on the entire page. While web scrapers show the price, gold chart and diversity of websites. Web scraping is the process of extracting data from websites through scraping software. Many people and organizations use scrapers, and many companies have now developed their own scraper tools to extract data. There are many motives to use scarper for business selling and pricing inquiries of their opponents, weather data monitoring, and research, contact scraping, and monitoring all the changes and analyzed data easily [6]. If you don't use a proper web scraping tool, you may face some challenges in completing your task. Website may crash or go down sometimes because of maintenance. These types of problems can be occurred during web scraping sessions.

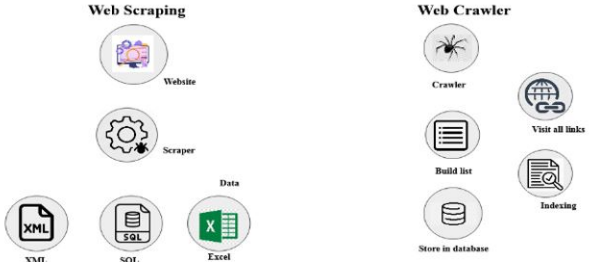

Fig. 1. Web scraping VS Web Crawler

HTML structure will be our initial priority for site scraping. When scraping material, the HTML structure must be in order. Scraping will be a disaster if the HTML code is not properly structured, since it will take a lot of time and pose a risk. It's a fantastic approach to collecting data if the content is well-structured [7]. HTML tags in the content must be properly formatted. It requires an ID or a class. It would be a disaster if the content HTML just had inline HTML. To get the data, it needs to be identified. The correct approach to inserting an ID or a class name for us to use Scraping is an excellent option if the content Html has this capability. To extract data from a web page, a web-scraping tool must first visit it. It takes time to download a web page and loading and extracting data from millions of online pages could take weeks or months. Because there are so many web pages on Amazon.com, extracting all the product data is almost impossible [8]. Fig. 1 shows the difference between the Web Scraping and Web Crawler.

A large amount of data is available on all the websites. Extraction of data from the website is difficult and manually extracting it is possible, but time-consuming and so much effort is required [9]. With the help of scraper tools, the extraction process is quick and easy, and we also monitor our data and keep it saved. Web scraping has so many purposes in today's world.

Companies use web scrapers to market, collect consumer data, and monitor their competitors. Indexes are updated using targeted data collected from e-commerce and advertising servers [10]. Which is based on pricing that varies regularly? Automated Web scraping indexes can provide more regular updating periods. Producers of blockbusters gather information about their current releases. User feedback, whether it is provided in a review on a movie site, is one example of such data. For government agencies and law enforcement agencies, monitoring illegal actions on social media platforms and specific forums is a valuable source of information. There are no obvious

sources for this type of behavior. We can presume they exist based on patents filed by the US government and publicly available information [11].

2. Literature review

Table 1 summarized the previous work of web scraping tools and technologies. To begin, there's 'Scrapy,' a web scraping extension that can be found in the Chrome Web Store. Scrapy is a basic but limited data mining add on that allows researchers to extract data online in the form of spread sheets. We can't gather proper data in a spreadsheet because the data is in a constrained format [12]. The extraction of data from selected websites is completely under the control of the users. It functions similarly to a hierarchical data base selection. The user merely picks the field that he or she wants to extract at the start of scraping, and Parse Hub automatically guesses similar data elements from a website [13]. When a user picks a piece of related data to extract, all similar components is extracted as well. A 'relative' search option is available for selecting other data elements from a particular website, which is subset information about the previously selected element [14]. Similarly, the user extracts all the data from the webpage. Parse Hub also provides a URL when extracting an element from a website. This URL is a field that can be left blank. Data sets are saved in CVS format after successful web scraping [12]. The World Wide Web is an interconnected network of information that people can access through websites. The way we share, acquire, and distribute data has changed dramatically because to the Internet. The amount of information available is always increasing [15].

According to IDC, the global data sphere will reach 163 zetta bytes by 2025. (That is a trillion gigabytes). That's ten times more data than the 16.1ZB created in 2016.[33] All this data will open up new commercial options and unique user-experiences (According to the International Data Corporation (IDC), 2017) [16].

Renita Crystal Pereira et. al., gave a summary of online scraping strategies and tools, which confront several challenges because data extraction isn't simple [22]. Because there is a great volume of data to handle and maintain, these tactics ensure that the data collected is correct [23], consistent, and has superior integrity. Although there are a few issues with functional approaches, such as the increased volume of web scraping, they can do serious harm to websites [24]. The web scraper's measurement level will differ from the original source file's measurement units, making it impossible to comprehend the data [25].

The use of social networking sites and the internet is growing by the day, such as Facebook, Twitter, LinkedIn, and others; user knowledge is also growing on the internet, which is accessible from anywhere. This also gives hackers an advantage when it comes to stealing information [19]. From a commercial standpoint, social networking is critical in the development of the concept of growing revenue [21]. It will assist consumers in achieving rapid shopping and saving time, like online shopping. Supporting the company and earning from it, on the other hand, has advantages [26].

Kaushal Parikh et. al., proposed a machine learning-based web scraping detection system. It is beneficial to companies that rely on research [12]. Web scraping has always been a tough assault to defend against [27]. When a firm posts information on the internet, it is possible that it will be copied and pasted and then used in a different context without the company's knowledge. Many safeguards have already been put in place, yet some of them are still being disregarded. As a result, the significance of machine learning emerges. Pattern detection is a skill that machine learning excels at. As a result, if we can teach the system to recognize an intruder's cadence, it will be able to prevent such dangers from occurring [28]. Web scraping solutions' main purpose is to convert complex data obtained over networks into structured data that can be recorded and analyzed in a central database. As a result, web scraping technologies have a significant influence on the outcome of the cause [29]. Sameer Padghan et. al., envisioned a method for extracting data from web pages to make web scraping easier. This technology would allow data to be scraped from a variety of websites, reducing human interaction, saving time, and improving the quality of data relevancy [30]. It will also assist the user in obtaining data from the site, saving it to their intent, and allowing them to use it as they desire.

The scraped data can be utilized for database creation, research, and other similar operations [31]. Scraping would become much more common, and it would frequently trespass on the structure to access the information. Scraping may be halted, however, by employing effective and secure online scraping techniques [22]. Anand Saurkar et. al., Web Scraping is a new approach that has been discovered. Web scraping is an important methodology for creating organized data [32] from unstructured data available on the internet. Scraping created structured data, which was then collected and analyzed in a central database's spreadsheets [33]. This study focuses on a summary of the web scraping data extraction process, numerous web scraping methodologies [34], and most of the most recent web scraping technologies. This methodology's main goal has been to collect web-based data and incorporate it into a specific repository [11]. In this paper, the authors covered the fun-

Table 1. Comparison of scraping tool previous work

Ref	Techniques	Dataset	Size	Limitation	Results
[17]	Web Scraping In A Knowledge Environment To Build Ontologies Using Python And Scrapy	website	2500	It works only on python library.	Salience and Beautiful Soup techniques used for accuracy in which Salience has better accuracy.
[1]	using python and MongoDB	Teachers' reviews.	850	Less effective in negative reviews than positive reviews.	It has 89% Accuracy in SVM.
[18]	web-scraping software in searching for grey literature	Reviews on Iot.	1500	It does not detect spam reviews.	It has good accuracy on non-spam reviews or positive reviews.
[19]	Exploratorily generate new hypotheses, test existing ones, or extend existing research	General Reviews from peoples.	800	Work good in Noun only.	SVM has high accuracy on comparison with ME and Naïve Bayes.
[20]	Web Data Extraction System.	Analysis of Linguistic	1100	Less effective on Verbs.	Support Vector Machine technique give maximum accuracy.
[7]	emplace the recharge material on the beach	Student Comments	3000	Less effective when dataset is not clean	85% Accuracy
[21]	Exploratorily generate new hypotheses, test existing ones, or extend existing research	Student Reviews of Loei University.	1148	The small dataset of students was utilized to teacher evaluation	The student comments corpus result should be 0.67

damentals of Web processing. They worked on scraping strategies for the web. The last section of the paper summarizes the various technological tools available for effective web scraping in the industry [35].

Federico Polidoro et. al., centered on the results of online scraping evaluation methodologies with a special focus on user electronics services and items across the commodity price studies sector Despite the fact that the research was completed in a short period of time [36], as evidenced by the following, it permitted the achievement of significant, but not conclusive, innovative efficiency outcomes [37]. Web scraping tactics utilized in the growth study will expose the researcher to a larger volume of data than is currently available in the data set, perhaps increasing the growth estimate. This topic was briefly discussed in the sections devoted to both examined items but dealing with this point of view necessitates a concern about the current survey architecture, which does not require or only selectively permits the use of big data approaches within existing sampling frameworks [25].

Table 1 compares past research on some of the relevant models, the approaches working in those models, and the type of data set used in those models. It also discusses the model's limits based on the accuracy of its results. The table's main feature is the limitations and results of previous models, which allows us to quickly determine how much work has been done on previous models and how accurate their results are, allowing us to create a model that produces more accurate results while reducing the limitations of previous models.

3. Methodology

The proposed system works based on web crawling technology and it works like bridge between the wedges between mysterious data. This technique transforms web links into visual blocks. A visual block is essentially a web page part. The framework is built from the ground up to detect the structure of web content [38]. It works like this: first select a website whose data you want to be gathered or extracted, put a URL in the scraper tool, and load the website [39]. Webpages created with Hyper Text Markup Language (HTML), data feeds in JSON format, and multimedia such as audio, video, or image extraction process, now parse HTML data. All the data is inserted into an html

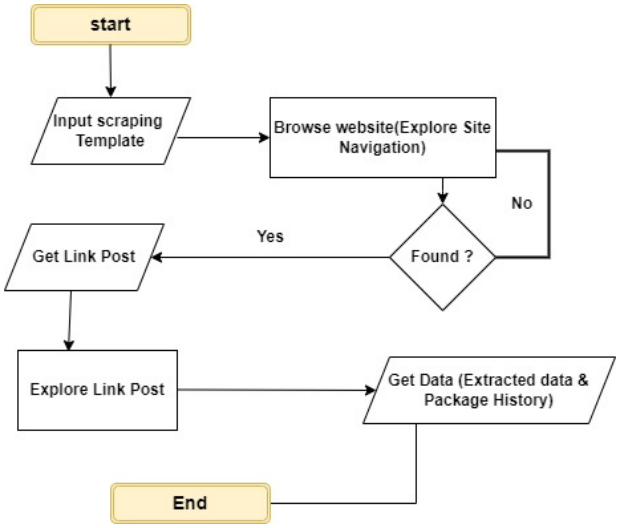

Fig. 2. Block diagram of proposed scraper

tag whose data is wanted through a website using just tags [37]. Data would be gathered in an untidy or unstructured form and then transformed into the required format [40]. It totally depends on your convenience which format you choose, like a document, PDF, or spreadsheet [41], where you want to store the data for analysis. Fig. 2 shows the proposed model of the scraper.

In the development approach, each part was completed before going to the next one. We build a web application using Flask. Flask was built using an html layout and index file web pages. The layout page is a page that never loads by itself, as it provides the layout of the elements of the webpage that do not change between pages [7]. The menu, or NAVBAR, at the top of each page is an example of this, as it remains consistent regardless of which page the application opens. Fig. 3 depicted the interface of scraping tool, the usage of tool is easy and user friendly, it works like Google search, the users are typed simple URL of website or blog. The spider (URL) method was introduced after that to allow the spider to be activated from a URL within the web application. Through Python libraries, scrapy extracts data and then stores it in JSON format [20]. Figs. 4 and 5 represented the command line and folder hierarchy of files, which takes Input to scrape, can be links, files, or a combination of the two [42], allowing you to create new files constructed from both existing and newly scraped content.

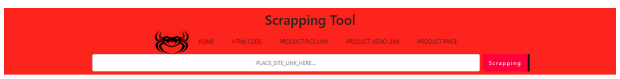

Fig. 3. Interface of scrapping tool

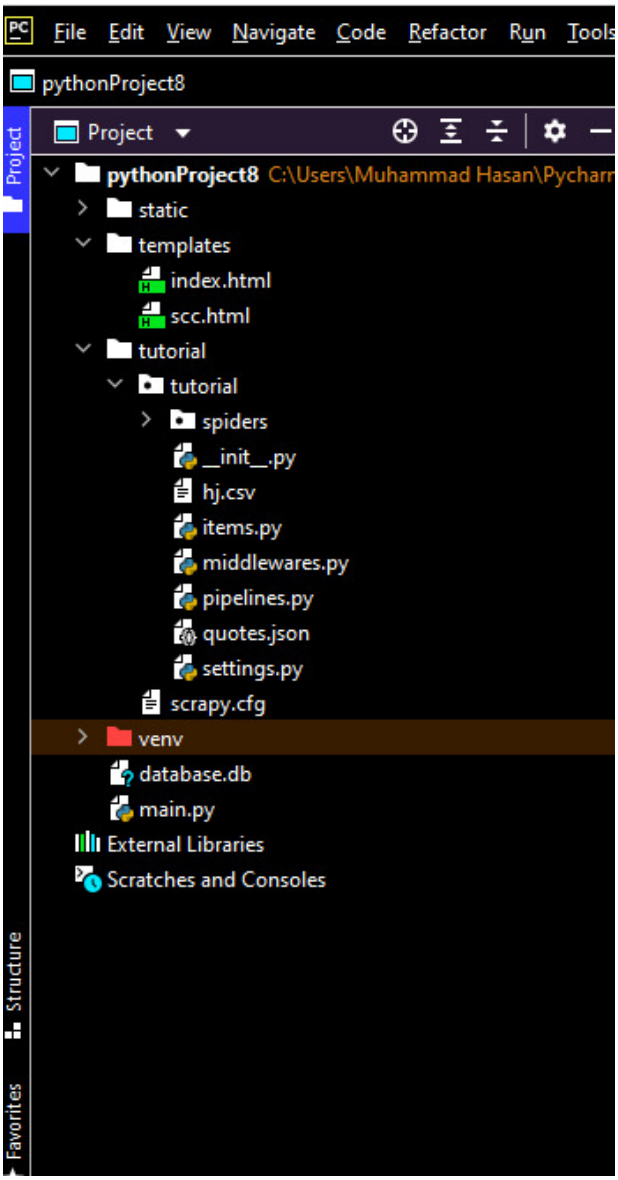

Fig. 4. Hierarchy and connection of files

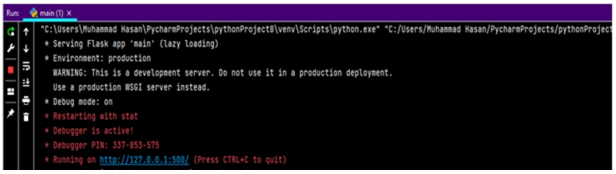

Fig. 5. Execution process

The following were the tools and languages utilized to create this system:

- Python and its libraries, Scrapy and Flask, for the back-end
- In the frontend, HTML, CSS, and jQuery

For two reasons, these technologies were chosen. The first is that, except for the Scrapy library, I already had expertise with these languages and libraries for developing web apps, and the second is that Scrapy is a Python library [43], so the application had to be created in Python as well. Because Scrapy was the only unknown tool, time was invested at the start of development in learning how to utilize the framework before moving on to the other, more familiar components [15].

We designed our project using **Flask**. And this project is based on web-based Flask is a Python Framework in which we use some libraries like "**Scrapy**". Scrapy will scrape the desired data from the internet for you, such as [44] From Amazon, ALIBABA, and many more sites; the **Request** library is used to get the data. Another thing is using the JSONify library, in which all your data will be converted to the **JSON** form. We chose the **JSON** format because we use the **JSON** data as an **API** for software [45]. And Fig. 6 shows the output of the working of the scraping tool.

3.1. Pseudo code

Being
Initialize user agent and visit website URL
Convert DOM data into Jsonify
Find all tags like<a><Img> and Add to a List
Forβ each Element in A list
If (Tag element .text.match(^ [a-z]$ ^ [A-Z]$))
Then:
Target_URL .add(tagElement.href)
End If
Endfor
Return(target_URLs)
end

4. Result discussions

Web scraping can be done with a variety of software tools and frameworks. This software tries to recognize a web page's data structure or provides an interface that eliminates the need to manually write web scraping code. Data can also be extracted straight from an API by some software. Parse hub is a web-based scraping program that allows you to scrape multiple and single URLs. It can retrieve complex data from websites that use AJAX, JavaScript, redirects, and cookies. The software will analyze, capture, and convert data from various sites into useful information. Scrapy is a free web scraping application for Python that may be used to develop web crawlers. It provides all the tools needed to gather, process, and save data from websites in the format and structure chosen.

The Web Scraper is a simple and free tool, as well as a Chrome plugin, for data extraction. It can scrape numerous pages at once and even retrieve dynamic data. The Web scraper is more powerful because it can handle pages with JavaScript and Ajax. The program saves data to a CSV or CouchDB file. In comparison to building your own web scraping system, ready-made web scraping software tools are significantly easier to use. For some applications, commercial software is not the best option. However, when a developer is hired to construct a custom web scraping solution, there is no need to start from scratch. Frameworks enable developers to save time by reusing generic modules, allowing them to concentrate on areas where an off-the-shelf solution isn't appropriate. The result is that our project converts the data into JSON format because several scrappers are available to extract out the information about the targeted website, but our project provides the facilities to convert the data into JSON format because this JSON data can be used as an API in software. In the JSON format, we have all the information about the targeted page. Table 2 showed the comparison of five scraping tools with their basic functionalities.

Fig. 8 shows the usage of scraping tool domain wise, it is shown that the fields where Web Scraping is used. The following statistics are based on data gathered from LinkedIn [46]. The following are the top ten industries with the highest need for web scraping skills: Insurance, banking, computer and network security, information technology and services, financial services, marketing and advertising, management consulting, internet and online media are all examples of industries that use computer software.

5. Conclusions

The web scraping tool has been developed for extracting the content from web such as news, comments, post, and images. The focus of automated tool is to improve the extraction process of data by reducing the Human effort and time. Due to the rapidity in the information broadcasting and the volume of available data, the traditional methods are not feasible to extract the complete and accurate information from web. The designed automated scraping tool is collecting, merging, categorizing and well managing the information and stored in JSON format. Data would be gathered in an untidy or unstructured form, and

Table 2. Frameworks for five programming languages the availability of three basic functionalities is represented.

	Dynamic Content	DOM navigation	DB integration	Programming Language
Scrapy	YES	YES	YES	Python
Web Scraper	NO	YES	YES	Perl
Goutte	NO	YES	YES	PHP
Capybara	YES	YES	YES	Ruby
jaunt	NO	YES	YES	Java

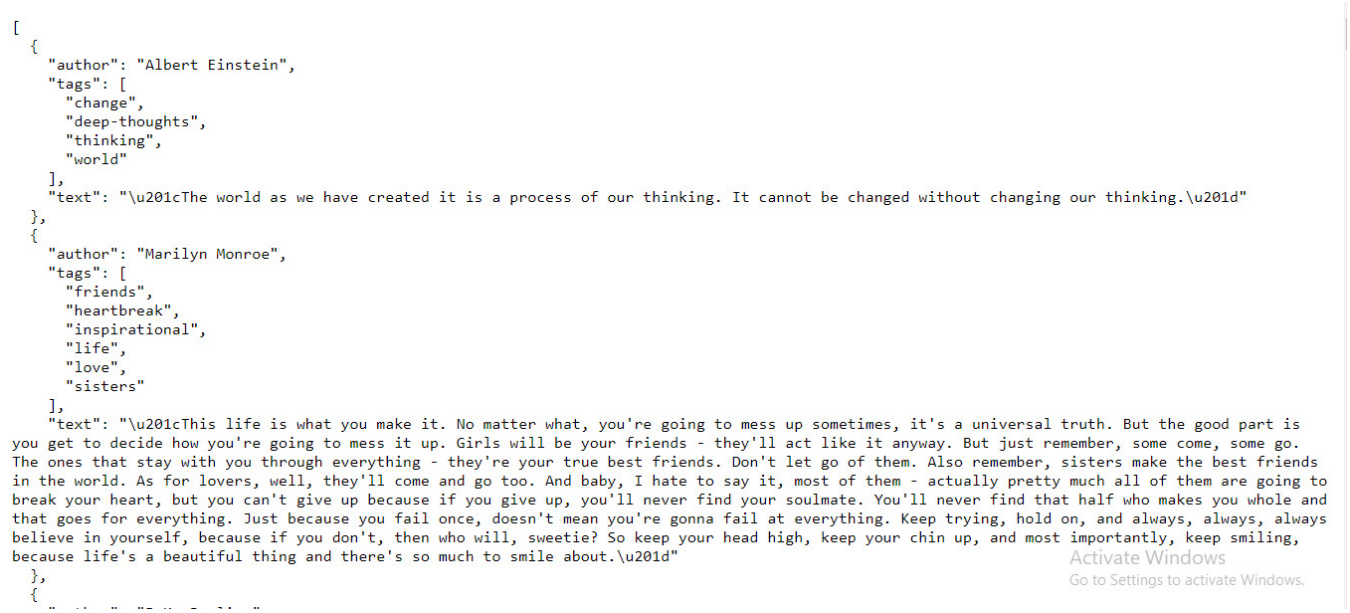

Fig. 6. JSON response

then transformed into the required format. This tool stored information in multiple formats such as CSV, PDF, and Spreadsheet. The scraping gathered a data as per design and structure of website, sometimes you revisit the site and layout has been updated. Even minor change creates big mess in data, as scrapers are designed as per old structure. This is the main purpose of software for marketing and pricing research and data mining of their competitor's product. It allows them to monitor all the changes and analyze the data easily. The proposed system frees the human endless copy and paste approach from the messy layouts. In future this work shall be extended, and data will be extracting from the more complex and tricky layouts. The development of autonomous agents that study the discovered rules and propose relevant courses of action or suggestions to users will be the focus of future effort. Web scraping future scope includes forecasting user needs to improve usability, scalability, and user retention, as well as providing an efficient framework for Web personalization through the effective use of Web Log files.

The semantic web is a futuristic vision in which web material can be analyzed and synthesized by automated processes. Most of the content on the internet is human readable. The browser has a hard time understanding the text because it can only visualize HTML mark-up. Content interpretation, selection, and management are the three most difficult activities on the internet. These three tasks are currently managed by humans. The semantic web will restore the balance between machine and human intelligence by automating three tough activities.

References

[1] V. Draxl. *"Web Scraping Data Extraction from websites"*. University of Applied Sciences Technikum Wien, 2018.

[2] B. Manjushree and G. Sharvani, (2020) *"Survey on Web scraping technology"* **Wutan Huatan Jisuan Jishu *16***: 1–8.

[3] L. Junjoewong, S. Sangnapachai, and T. Sunetnanta. "ProCircle: A promotion platform using crowdsourcing and web data scraping technique". In: *2018 Seventh ICT International Student Project Conference (ICT-ISPC)*. IEEE. 2018, 1–5.

Fig. 7. The steps of scraper tool

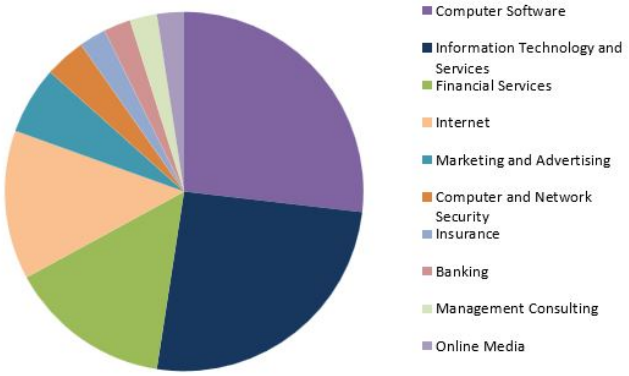

Fig. 8. Domain wise web scraping

[4] A. V. Saurkar, K. G. Pathare, and S. A. Gode, (2018) *"An overview on web scraping techniques and tools"* **International Journal on Future Revolution in Computer Science & Communication Engineering** *4*(4): 363–367.

[5] M. El Asikri, S. Krit, and H. Chaib, (2020) *"Using Web Scraping In A Knowledge Environment To Build Ontologies Using Python And Scrapy"* **European Journal of Molecular & Clinical Medicine** *7*(03): 2020.

[6] E. Gallagher, (2018) *"Scraping Websites for Law Enforcement"* **School of Computing, Engineering & Intelligent Systems, Computer Science**:

[7] A. Tarango. *"307 ChopShop Senior Design Web Scraper"*. (phdthesis). University of Wyoming.

[8] K. Parikh, D. Singh, D. Yadav, and M. Rathod, (2018) *"Detection of web scraping using machine learning"* **Open access international journal of Science and Engineering**: 114–118.

[9] R. Landers, R. Brusso, K. Cavanaugh, and A. Collmus, (2016) *"A primer on theory-driven web scraping: Automatic extraction of big data from the internet for use in psychological research"* **Psychological Methods** *21*(4): 475–492. DOI: 10.1037/met0000081.

[10] B. Ujwal, B. Gaind, A. Kundu, A. Holla, and M. Rungta. "Classification-based adaptive web scraper". In: ***2017-December***. cited By 10. 2017, 125–132. DOI: 10.1109/ICMLA.2017.0-168.

[11] B. G. Dastidar, D. Banerjee, and S. Sengupta, (2016) *"An intelligent survey of personalized information retrieval using web scraper"* **International Journal of Education and Management Engineering** *6*(5): 24–31.

[12] O. Lloyd and C. Nilsson. *How to Build a Web Scraper for Social Media*. 2019.

[13] S. Han and C. Anderson, (2021) *"Web Scraping for Hospitality Research: Overview, Opportunities, and Implications"* **Cornell Hospitality Quarterly** *62*(1): 89–104. DOI: 10.1177/1938965520973587.

[14] W. Liu, X. Meng, and W. Meng, (2010) *"ViDE: A vision-based approach for deep web data extraction"* **IEEE Transactions on Knowledge and Data Engineering** *22*(3): 447–460. DOI: 10.1109/TKDE.2009.109.

[15] K. Clark and A. Evert, (2019) *"Building an Alternative Web Scraper for Big Data Analytics"*:

[16] International Data Corporation.

[17] S. Chasins, M. Mueller, and R. Bodik. "Rousillon: Scraping distributed hierarchical web data". In: cited By 35. 2018, 963–975. DOI: 10.1145/3242587.3242661.

[18] R. Chaulagain, S. Pandey, S. Basnet, and S. Shakya. "Cloud Based Web Scraping for Big Data Applications". In: cited By 23. 2017, 138–143. DOI: 10.1109/SmartCloud.2017.28.

[19] N. R. Haddaway et al., (2015) *"The use of web-scraping software in searching for grey literature"* **Grey J** *11*(3): 186–90.

[20] V. Singrodia, A. Mitra, and S. Paul. "A Review on Web Scrapping and its Applications". In: cited By 11. 2019. DOI: 10.1109/ICCCI.2019.8821809.

[21] K. Weedman, (2002) *"On the spur of the moment: Effects of age and experience on hafted stone scraper morphology"* **American Antiquity *67***(4): 731–744. DOI: 10.2307/1593801.

[22] E. Vargiu and M. Urru, (2013) *"Exploiting web scraping in a collaborative filtering-based approach to web advertising."* **Artif. Intell. Res. *2***(1): 44–54.

[23] A. Sundas, S. Badotra, Y. Alotaibi, S. Alghamdi, and O. Khalaf, (2022) *"Modified bat algorithm for optimal VM s in cloud computing"* **Computers, Materials and Continua *72***(2): 2877–2894. DOI: 10.32604/cmc.2022.025658.

[24] B. Zhao, (2017) *"Web scraping"* **Encyclopedia of big data**: 1–3.

[25] D. S. Sirisuriya et al., (2015) *"A comparative study on web scraping"*:

[26] B. Audeh, M. Beigbeder, A. Zimmermann, P. Jaillon, and C. Bousquet, (2017) *"Vigi4Med scraper: A framework for web forum structured data extraction and semantic representation"* **PLoS ONE *12***(1): DOI: 10.1371/journal.pone.0169658.

[27] A. Khan, A. Laghari, A. Shaikh, S. Bourouis, A. Mamlouk, and H. Alshazly, (2021) *"Educational blockchain: A secure degree attestation and verification traceability architecture for higher education commission"* **Applied Sciences (Switzerland) *11***(22): DOI: 10.3390/app112210917.

[28] Y. Neil, (2016) *"Web Scraping the Easy Way"*:

[29] R. Egger, M. Kroner, and A. Stöckl. "Web scraping". In: *Applied Data Science in Tourism*. Springer, 2022, 67–82.

[30] V. Krotov and L. Silva. "Legality and ethics of web scraping". In: cited By 14. 2018.

[31] R. Sharma, (2020) *"DATA CRAPER"*:

[32] B. Sharma, A. Hashmi, C. Gupta, O. I. Khalaf, G. M. Abdulsahib, and M. M. Itani, (2022) *"Hybrid Sparrow Clustered (HSC) Algorithm for Top-N Recommendation System"* **Symmetry *14***(4): 793.

[33] Felix Speckmann, (2021) *"Web Scraping: A Useful Tool to Broaden and Extend Psychological Research"* **Zeitschrift für Psychologie *229***(4): 241–244. DOI: https://doi.org/10.1027/2151-2604/a000470.

[34] U. Janniekode, R. Somineni, O. Khalaf, M. Itani, J. Chinna Babu, and G. Abdulsahib, (2022) *"A Symmetric Novel 8T3R Non-Volatile SRAM Cell for Embedded Applications"* **Symmetry *14***(4): DOI: 10.3390/sym14040768.

[35] D. Goßen, I. H. Jonker, and I. E. Poll. *"Design and implementation of a stealthy OpenWPM web scraper"*. (phdthesis). Master's thesis, Radboud Universiteit Nijmegen, 2020.

[36] M. Edeh, O. Khalaf, C. Tavera, S. Tayeb, S. Ghouali, G. Abdulsahib, N. Richard-Nnabu, and A. Louni, (2022) *"A Classification Algorithm-Based Hybrid Diabetes Prediction Model"* **Frontiers in Public Health *10***: DOI: 10.3389/fpubh.2022.829519.

[37] Hemavathi, S. Akhila, Y. Alotaibi, O. Khalaf, and S. Alghamdi, (2022) *"Authentication and Resource Allocation Strategies during Handoff for 5G IoVs Using Deep Learning"* **Energies *15***(6): DOI: 10.3390/en15062006.

[38] X. Wang, J. Liu, X. Liu, Z. Liu, O. I. Khalaf, J. Ji, and Q. Ouyang, (2022) *"Ship feature recognition methods for deep learning in complex marine environments"* **Complex & Intelligent Systems**: 1–17.

[39] J. Jayapradha, M. Prakash, Y. Alotaibi, O. Khalaf, and S. Alghamdi, (2022) *"Heap Bucketization Anonymity - An Efficient Privacy-Preserving Data Publishing Model for Multiple Sensitive Attributes"* **IEEE Access *10***: 28773–28791. DOI: 10.1109/ACCESS.2022.3158312.

[40] C. Kavitha, V. Mani, S. Srividhya, O. Khalaf, and C. Tavera Romero, (2022) *"Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models"* **Frontiers in Public Health *10***: DOI: 10.3389/fpubh.2022.853294.

[41] T. Puri, M. Soni, G. Dhiman, O. Ibrahim Khalaf, M. alazzam, and I. Raza Khan, (2022) *"Detection of Emotion of Speech for RAVDESS Audio Using Hybrid Convolution Neural Network"* **Journal of Healthcare Engineering *2022***: DOI: 10.1155/2022/8472947.

[42] A. A. Khan, A. A. Laghari, and S. A. Awan, (2021) *"Machine learning in computer vision: A review"* **EAI Transactions on Scalable Information Systems**: e4.

[43] A. Khan, Z. Shaikh, L. Belinskaja, L. Baitenova, Y. Vlasova, Z. Gerzelieva, A. Laghari, A. Abro, and S. Barykin, (2022) *"A Blockchain and Metaheuristic-Enabled Distributed Architecture for Smart Agricultural Analysis and Ledger Preservation Solution: A Collaborative Approach"* **Applied Sciences (Switzerland) *12***(3): DOI: 10.3390/app12031487.

[44] A. Khan, Z. Shaikh, L. Baitenova, L. Mutaliyeva, N. Moiseev, A. Mikhaylov, A. Laghari, S. Idris, and H. Alshazly, (2021) *"QoS-ledger: Smart contracts and metaheuristic for secure quality-of-service and cost-efficient scheduling of medical-data processing"*

Electronics (Switzerland) ***10***(24): DOI: 10 . 3390 / electronics10243083.

[45] A. Khan, Z. Shaikh, A. Laghari, S. Bourouis, A. Wagan, and G. Ali, (2021) *"Blockchain-aware distributed dynamic monitoring: A smart contract for fog-based drone management in land surface changes"* **Atmosphere** ***12***(11): DOI: 10.3390/atmos12111525.

[46] A. Khan, A. Laghari, D.-S. Liu, A. Shaikh, D.-A. Ma, C.-Y. Wang, and A. Wagan, (2021) *"EPS-ledger: Blockchain hyperledger sawtooth-enabled distributed power systems chain of operation and control node privacy and security"* **Electronics (Switzerland)** ***10***(19): DOI: 10.3390/electronics10192395.