A Seminar Report On

WEB SCRAPING

Submitted in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING

from

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD

By

SRILEKHA BAYYAPU (21C11A05G0)

Under the guidance of

Mrs. M. Anusha M. Tech (Ph. D).
Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

ANURAG ENGINEERING COLLEGE
(An Autonomous Institution)
(Affiliated to JNTUH, Hyderabad& Approved by AICTE, New Delhi)
Ananthagiri(V&M), Kodad, Suryapet (Dt.),Telangana-508206.
2024 – 25

ANURAG ENGINEERING COLLEGE

(An Autonomous Institution)

(Affiliated to JNTUH, Hyderabad & Approved by AICTE, New Delhi)
Ananthagiri (V & M), Kodad, Suryapeta (Dt), Telangana-508206.

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

*This is to certify that the Seminar work entitled “**WEB SCRAPING**” is a Bonafide work done by “**SRILEKHA BAYYAPU (21C11A05G0)**” in the partial fulfillment for the award of Bachelor of Technology in Computer Science & Engineering from JNTUH, Hyderabad during the year**2024-25**.*

This work has not been submitted to any other university or institute or organization for the award of any degree or diploma.

Mrs. M. Anusha	Dr. Y.V.R. Naga Pawan
Supervisor	**H.O.D.**

ACKNOWLEDGEMENTS

This report will certainly not be completed without due acknowledgement paid to all those who helped me during this seminar.

I express sincere thanks to my supervisor **Mrs. M. Anusha,** for giving moral support, kind attention and valuable guidance throughout this seminar.

It is my privilege to thank **Dr.Y.V.R.Naga Pawan**, Head of the Department, for his encouragement during the progress of this seminar.

I am thankful to both Teaching and non-teaching staff members of **Department of Computer Science & Engineering** for their kind cooperation and all sorts of support in bringing out this seminar successfully.

I derive great pleasure in expressing my sincere gratitude to the principal **Dr.T. Suresh Kumar** for his timely suggestions, which helped me to complete this work successfully.

I am thankful to the **Management** of **ANURAG Engineering College** for providing required facilities during the seminar.

I would like to thank my parents and my friends for being supportive all the time, and I am very much obliged to them.

SRILEKHA BAYYAPU

(21C11A05G0)

CONTENTS

ABSTARCT

Web Scraping is the technique which allows user to fetch data from the World Wide Web. It gives a brief introduction to Web Scraping covering different technologies available and methods to prevent a website from getting scraped. Currently available software tools available for web scraping are also listed with a brief description. Web Scraping is explained with a practical example. Main objective of Web Scraping is to extract information from one or many websites and process it into simple structures such as spreadsheets, database or CSV file. However, in addition to be a very complicated task, Web Scraping is resource and time consuming, mainly when it is carried out manually.

List Of Figures

List Of Acronyms

API	Application Programming Interface
AWS	Amazon Web Service
DOM	Document Object Model
HTML	Hyper Text Markup Language
HTTP	Hyper Text Transfer Protocol
RAM	Random Access Memory
SEO	Search Engine Optimization
SQL	Structured Query Language
XML	Extensible Markup Language

CHAPTER 1
INTRODUCTION

1.1 what is web scraping

It is the automation of the data extraction process from websites. One way is to copy-paste the data, which is both tedious and time-consuming manually, So This event is done with the help of web scraping software known as web scrapers. They automatically load and extract data from the websites based on user requirements. These can be custom-built to work for one site or can be configured to work with any website. data is usually saved in a local file so that it can be manipulated and analyzed as needed. If you've ever copied and pasted content from a website into an Excel spreadsheet.

This is essentially what web scraping is, but on a very small scale. However, when people refer to 'web scrapers,' they're usually talking about software applications. Web scraping applications are programmed to visit websites, grab the relevant pages and extract useful information. By automating this process, these bots can extract huge amounts of data in a very short time. This has obvious benefits in the digital age.

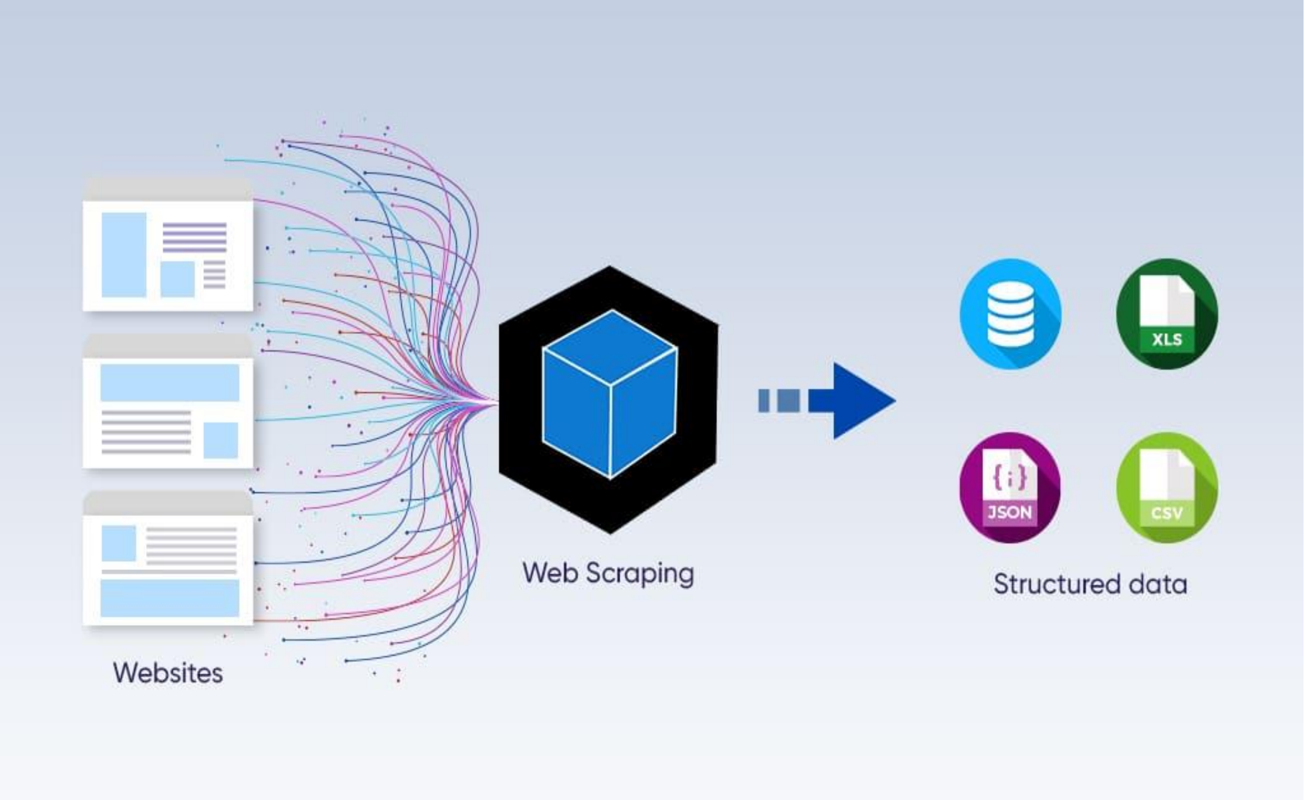


Fig 1.1: Web Scraping

1.2 Uses of Web Scraping

Web scraping finds many uses both at a professional and personal level. Having different needs at different levels, some popular uses of web scraping are:

- **Brand Monitoring and Competition Analysis:** Web Scraping is used to get customer feedback regarding a particular service or product to understand how a customer feels regarding that particular thing. It also extracts competitor data in a structural, usable format.
- **Machine Learning:** Machine Learning is a process of Artificial Intelligence in which the machine is allowed to learn and improve with its experience rather than being explicitly programmed. For that, a large amount of data is required from millions of sites which is extracted through web scraping software.
- **Financial Data Analysis:** Web Scraping is used to keep a record of the stock market in a usable format and hence employ the same for insights.
- **Social Media Analysis:** It is used to extract data from social media sites to gauge customer trends, and how they react to the campaign.
- **SEO monitoring:** Search Engine Optimization is the optimization of the visibility and ranking of a website among different search engines like Google, Yahoo, Bing, etc. Web scraping is used to understand how the ranking of the content over time.

1.3 Techniques of Web Scraping

There are two ways of extracting data from websites, the Manual extraction technique, and the automated extraction technique.

Manual Extraction Techniques: Manually copy-pasting the site content comes under this technique. Though tedious, time taking and repetitive it is an effective way to scrap data from the sites having good anti-scraping measures like bot detection.

Automated Extraction Techniques: Web scraping software is used to automatically extract data from sites based on user requirement.

HTML Parsing: Parsing means to make something understandable to be analyzing it part by part. To wit, it means to convert the information in one form to another form that is easy to that is easier to work on with. means taking in the code and extracting relevant information from it based on the user requirement. Mainly executed using JavaScript, the target as the name suggests are HTML pages.

DOM Parsing: The Document Object Model is the official recommendation of the World Wide Web Consortium. It defines an interface that enables a user to modify and update the style, structure, and content of the XML document.

Web Scraping Software: Nowadays, many web scraping tools are available or are custom build on users need to extract required desiring information from millions of websites.

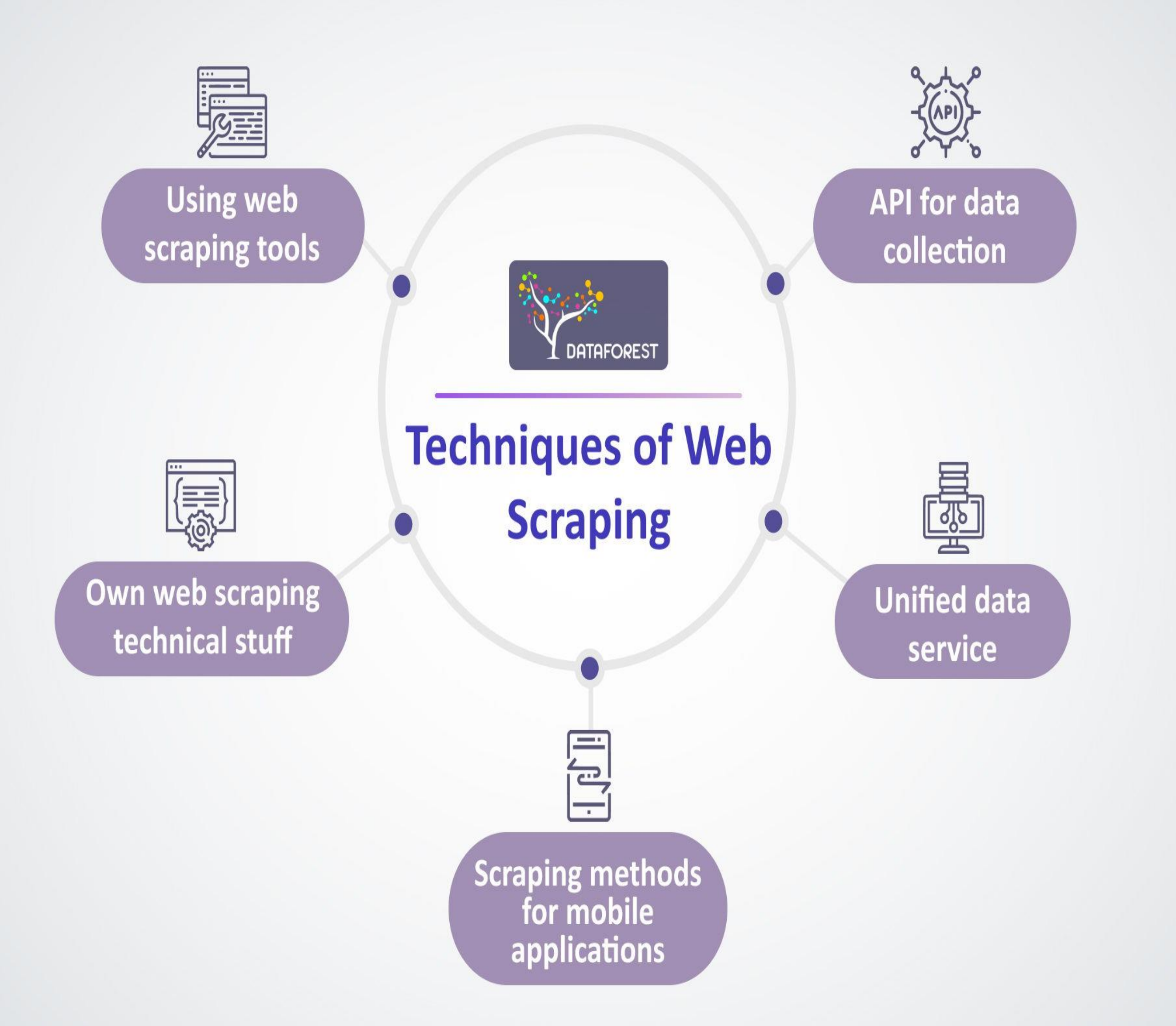

Using web scraping tools
API for data collection
DATAFOREST
Techniques of Web Scraping
Own web scraping technical stuff
Unified data service
Scraping methods for mobile applications

Fig 1.2 Techniques of web scraping

CHAPTER 2

LITERATURE SURVEY

1.Title: Web scraping with python

Author: Ryan Mitchell

Offered a description of online scraping approaches and tools, which confront a number of challenges since data extraction isn't straightforward. Because there is a great volume of data to manage and maintain, these tactics ensure that the data gathered is accurate, consistent, and has superior integrity. Although there are a few issues with functional approaches, such as the increased volume of web scraping, they may do serious damage to websites. The web scraper's measurement level will differ from the original source file's measurement units, making it impossible to comprehend the data.

2.Title: A Framework for Web Scraping

Author: Dimitrios Kouzis-Loukas

Web scraping has always been a tough assault to defend against. 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 relevance of machine learning emerges. Pattern detection is a skill that machine learning excels at. As a result, if we can teach the system to recognise an intruder's cadence, it will be able to prevent such dangers from happening. The primary goal of web scraping solutions is to convert complicated data collected over networks into structured data that can be saved and evaluated in a central database. As a consequence, web scraping technologies have a substantial influence on the cause's outcome.

3.Title: Web Scraping A beginner's Guide

Author: Diego copello

Proposed a method for extracting data from online pages in order 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. It will also assist the user in obtaining data from the site, saving it to their intent, and allowing them to utilise it as they desire. The scraped data may be utilised for database creation, research, and other similar operations. Scraping would become much more common, and it would often trespass on the structure in order to access the information. Scraping may be halted,however, by using effective and secure online scraping techniques. This approach should be seen as a gift that must be handled with caution in order to improve human races.

4.Title: Data Mining and Web Scraping

Author: Piatetsky-Shapiro

Web scraping is an essential approach for creating organised data from unstructured data accessible on the internet. Scraping created structured data, which was then gathered and analysed in a central database's spreadsheets. This study focuses on an overview of the web scraping data extraction process, numerous web scraping methodologies, and the majority of the most recent web scraping technologies.

This methodology's main goal has been to collect web-based data and incorporate it into a particular repository. In this paper, the writers covered the fundamentals of Web processing. They worked on scraping strategies for the web. The report concludes with a survey of the different technology options available in the industry for successful web scraping.

5.Title: Web Scraping with Beautiful soup and Python

Author: Sweigart

Focused on the results of online scraping assessment methodologies with a special focus on user electronics services and items. Despite the fact that the study was completed in a short focused on the results of online scraping assessment methodologies with a special focus on user electronics services and items. Despite the fact that the study was completed in a short period of time.

CHAPTER 3
ARCHITECTURE

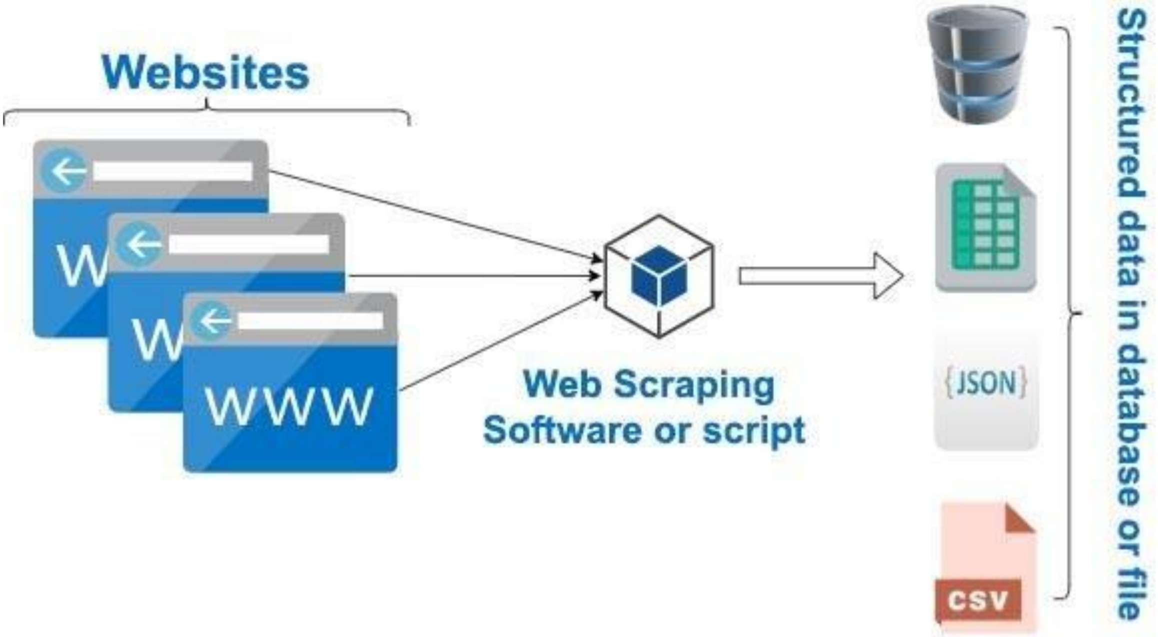


Fig 3.1 Web scraping Architecture

To build an entire web scraping software, a proper algorithm is not enough. We need to connect several services together that allows us to:

Communicate with the software some inputs (the websites we wish to scrape) and different parameters Save the scraped data on a storage unit Mange the EC2 instance, stopping it or activating it when we need . Also, because we need to store textual data, a tabular format like SQL is not the best choice. We need a data format that is flexible enough to have different fields, if needed, and can accept very long textual fields . There are many ways we can interact with an active Virtual Machine. Because the objective of the web scraper is quite simple, my objective is to only send new links to a table in DynamoDB that is scanned periodically by the web scraper. Every hour, the web scraper will read if there are new entries in that specific DynamoDB.

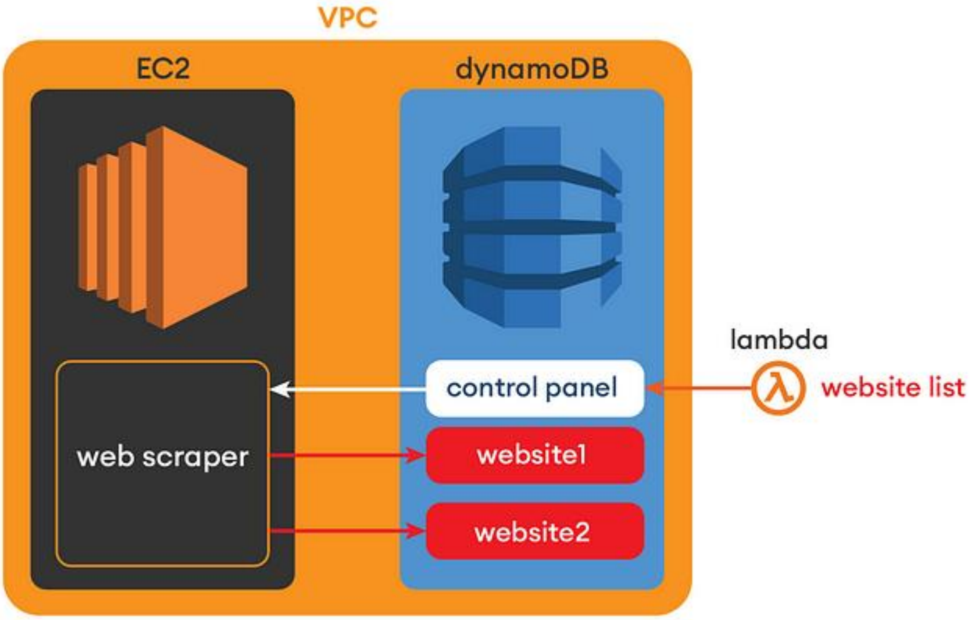


Fig 3.2 Architecture of a Cloud Web Scraper

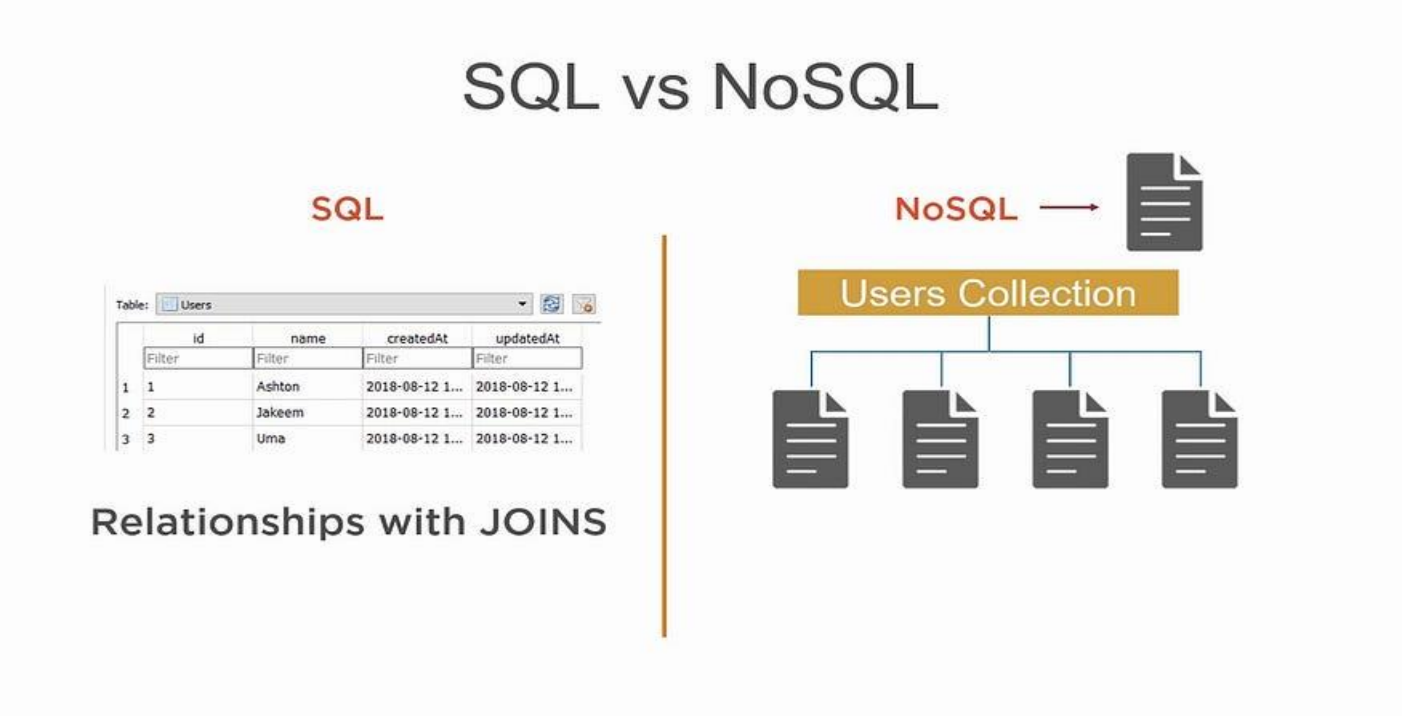
3.2.1 Managing the EC2 Instance

As mentioned, there are several EC2 instances we can choose from. While the cheapest option (about .14 USD per day) is the t2. micro, with 1 CPU and .5 GB of RAM may not be fit for web scraping. The next model is the t2. micro, which has 1 CPU and 1 GB of RAM, a bit more powerful and able to withstand more intensive power demands. Once activated, I have mounted my web scraping software on the virtual machine, so it will run 24/7 until the task is done. In case we encounter a bad link and the web scraper breaks, we can set up a task in the virtual machine that makes the software run again. The software will be unkillable.

3.2.2 Saving scraped files in storage

The web scraper will keep downloading files, but they cannot be stored in the EC2 memory. We could, but it would make much more sense, as we would need to connect to the EC2 each time to extract our data. The are several kinds of storage we can choose from, the most two popular formats in which we can store our data are the json format and SQL.The issue when using python is that is not a language that goes along very well with relational databases. Also, because we need to store textual data, a tabular format like SQL is not the best choice. We need a data format that is flexible enough to have different fields, if needed, and can accept very long textual fields. In our case, json is perfect. It is also less expensive, as using the service for NoSQL on AWS has no initial costs, unlike an SQL database.

The service that AWS uses for storing NoSQL data and allows us to store an unlimited amount of NoSQL tables, even offering a UI from which we can access it and easily visualize our data. For every web page, we are going to scrape, all the data will be sent and stored on our DynamoDB storage.



	id	name	createdAt	updatedAt
	Filter	Filter	Filter	Filter
1	1	Ashton	2018-08-12 1...	2018-08-12 1...
2	2	Jakeem	2018-08-12 1...	2018-08-12 1...
3	3	Uma	2018-08-12 1...	2018-08-12 1...

fig 3.3 sql vs Nosql

3.3.1 Communicating with the Web Scraper

Now that we have established where we will get the data from and where to store it, we need a way to communicate with the machine to give it proper directions. There are many ways we can interact with an active Virtual Machine. Because the objective of the web scraper is quite simple, my objective is to only send new links to a table in DynamoDB that is scanned periodically by the web scraper. Every hour, the web scraper will read if there are new entries in that specific DynamoDB table, and if there are it will start scraping all the links of that web page.To update the dataset and send small information to the DynamoDB service. There are two main advantages of using lambda

CHAPTER 4

FUNCTIONING

How does a web scraper function?

web scraping is, and why different organizations use it. While the exact method differs depending on the software or tools you're using, all web scraping bots follow three basic principles:

- Step 1: Making an HTTP request to a server
- Step 2: Extracting and parsing (or breaking down) the website's code
- Step 3: Saving the relevant data locally

Now let's take a look at each of these in a little more detail.

Step 1: Making an HTTP request to a server

As an individual, when you visit a website via your browser, you send what's called an HTTP request. This is basically the digital equivalent of knocking on the door, asking to come in. Once your request is approved, you can then access that site and all the information on it. Just like a person, a web scraper needs permission to access a site. Therefore, the first thing a web scraper.

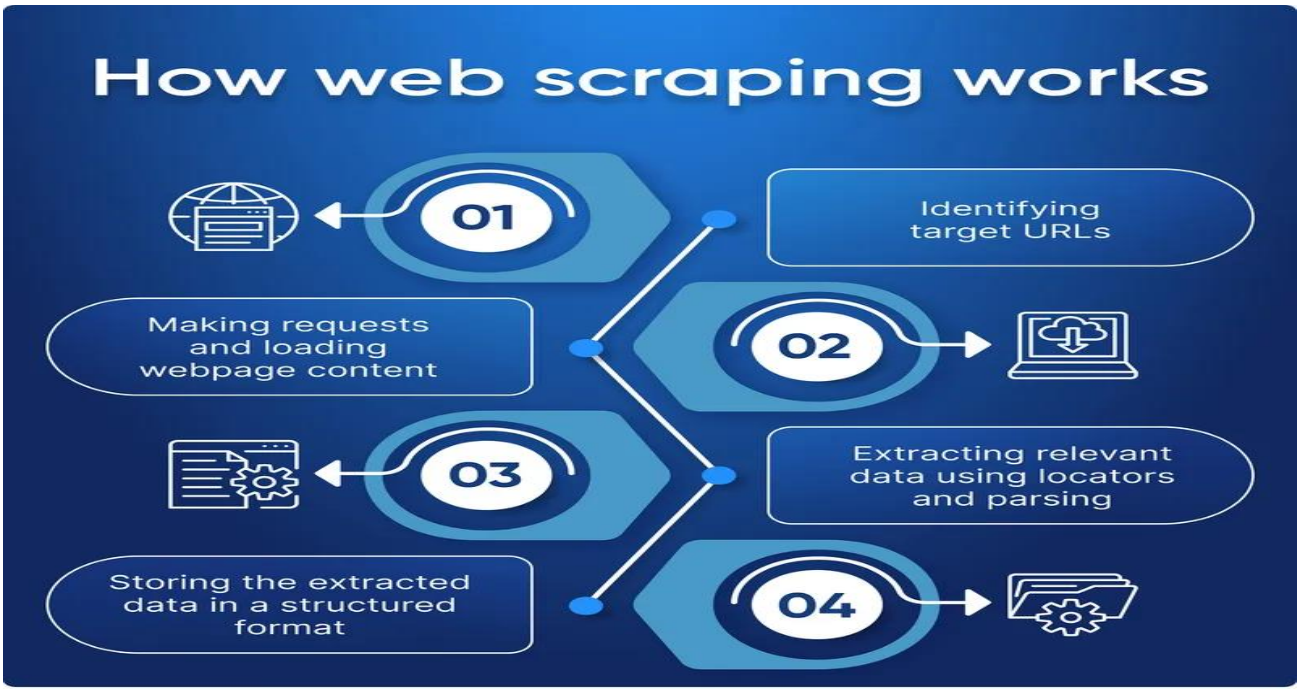

How web scraping works
01
Identifying target URLs
Making requests and loading webpage content
02
03
Extracting relevant data using locators and parsing
Storing the extracted data in a structured format
04

Fig 4.1 working of web scraping

Step 2: Extracting and parsing website's code

Once a website gives a scraper access, the bot can read and extract the site's HTML or XML code. This code determines the website's content structure. The scraper will then parse the code (which basically means breaking it down into its constituent parts) so that it can identify and extract elements or objects that have been predefined by whoever set the bot loose! These might include specific text, ratings, classes, tags, IDs, or other information.

Step 3: Saving the relevant data locally

Once the HTML or XML has been accessed, scraped, and parsed, the web scraper will then store the relevant data locally. As mentioned, the data extracted is predefined by you (having told the bot what you want it to collect). Data is usually stored as structured data, often in an Excel file, such as a .csv or .xls format.With these steps complete, you're ready to start using the data for your intended purposes.

How to scrape the web (step-by-step):

The exact method for carrying out these steps depends on the tools you're using, so we'll focus on the (non-technical) basics.

Step 1: Find the URLs you want to scrape

It might sound obvious, but the first thing you need to do is to figure out which website(s) you want to scrape. If you're investigating customer book reviews, for instance, you might want to scrape relevant data from sites like Amazon, Goodreads, or Library Thing.

Step 2: Inspect the page

Before coding your web scraper, you need to identify what it has to scrape. Right-clicking anywhere on the frontend of a website gives you the option to 'inspect element' or 'view page source.' This reveals the site's backend code, which is what the scraper will read.

Step 3: Identify the data you want to extract

If you're looking at book reviews on Amazon, you'll need to identify where these are located in the backend code. Most browsers automatically highlight selected frontend content with its corresponding code on the backend. Your aim is to identify the unique tags that enclose (or 'nest') the relevant content (e.g. <div> tags).

Step 4: Write the necessary code

Once you've found the appropriate nest tags, you'll need to incorporate these into your preferred scraping software. This basically tells the bot where to look and what to extract. It's commonly done using Python libraries, which do much of the heavy lifting. You need to specify exactly what data types you want the scraper to parse and store. For instance, if you're looking for book reviews, you'll want information such as the book title, author name, and rating.

Step 5: Execute the code

Once you’ve written the code, the next step is to execute it. Now to play the waiting game! This is where the scraper requests site access, extracts the data, and parses it (as per the steps outlined in the previous section).

Step 6: Storing the data

After extracting, parsing, and collecting the relevant data, you’ll need to store it. You can instruct your algorithm to do this by adding extra lines to your code. Which format you choose is up to you, but as mentioned, Excel formats are the most common. You can also run your code through a Python (short for ‘regular expressions’) to extract a cleaner set of data that’s easier to read.Now you’ve got the data you need, you’re free to play around with it.Of course, as we often learn in our explorations, web scraping isn’t always as straightforward as it at first seems. It’s common to make mistakes and you may need to repeat some steps. But don’t worry, this is normal, and practice makes perfect!

web scraping requires some knowledge of programming languages, the most popular for the task being python. Luckily, Python comes with a huge number of that make web scraping much easier.

CHAPTER 5
APPLICATIONS

Web Scraping Web scraping is widely utilized for a variety of purposes, including comparing prices online, observing changes in weather data, website change detection, research, integrating data from multiple sources, extracting offers and discounts, scraping job postings information from job portals, brand monitoring, and market analysis. It is also used as a means of data collection quickly and efficiently. Web scraping has myriad applications in various domains. It acts as a prerequisite to big data analytics. Discussed below are a few of the several domains where web scraping is used.

5.1 Healthcare:

It is no longer a domain that relies wholly on physical contact. Instead, in its unique manner, it has gone digital. In this data-driven environment, web scraping in healthcare can save many lives by allowing sensible decisions to be made. Healthcare workers typically regard data collecting engaging many patients as a tedious and arduous process. Even while clinical data is needed more than ever, the current patient load makes gathering it nearly impossible. To that end, the author proposes implementing a system that collects clinical data from SARS-CoV2 patients who visit the hospital automatically and autonomously for future research . Another application of web data extraction techniques in the healthcare domain is research conducted by Dascalu et al., where crawlers extract drug leaflets.

5.2 Social media:

Extracting data from social media proves to be a great help in improving the marketing campaigns for companies. In this fast-paced world, companies can quickly analyze the customers' sentiment towards their products, improve public relations and audience engagement. For this purpose, they created a web-based Instagram account data download application that may be utilized by numerous parties for this purpose, using a web scraping technology. The web scraping method was chosen by the researchers eliminating the need to use Instagram's Application Programming Interface (API), which has several restrictions for accessing and retrieving data on the platform. The web scraping method successfully created an Instagram account data grabber application. Application testing was carried out on 15 accounts with a total number of publications ranging from 100 to 11000 in this study. The web scraping solution was able to successfully capture Instagram account data for 2412 accounts, based on the results of the analysis. This application can help users save Instagram account data to a database manager and export data to several formats, namely Excel, JSON, or CSV .

5.3 Finance:

The author proposed a first approach to develop web-based innovation indicators that could address some of the drawbacks of existing indicators. In particular, they created a strategy for identifying product innovator enterprises on a wide scale at a minimal cost. Then, utilizing traditional company-level data from a questionnaire-based innovation survey, trained an ANN classification model using labeled (product innovator/no product innovator) online texts of surveyed enterprises (German Community Innovation Survey). They then used their categorization model to forecast whether or not hundreds of thousands of German companies are product innovators by analyzing their online texts. Next, they compared their predictions against patent statistics at the firm level, benchmark data derived from survey analysis, and regional innovation indicators. Given its breadth and geographic granularity, the findings show that this method yields solid projections and has the potential to be a valuable and cost-effective addition to the existing set of innovation indicators . The research conducted by Tharanya et al.

5.4 Marketing:

The vast amount of customer data in the form of a digital footprint available to analyze customer behavior and to answer customer research questions. In their paper, Saranya et al.propose to predict customer purchase intention during online purchases using machine learning models. The data is collected using web scraping since the information on the Web is in an unstructured format. The data is further analyzed to predict the purchase intent. Nguyen et al. analyze social media engagement of Australian SMEs using web scraping. They collect the data from Instagram using Instagram API and use the data to further find that tagging instead of hashtags garner more engagement as it is more trustworthy

CHAPTER 6
ADVANTAGES

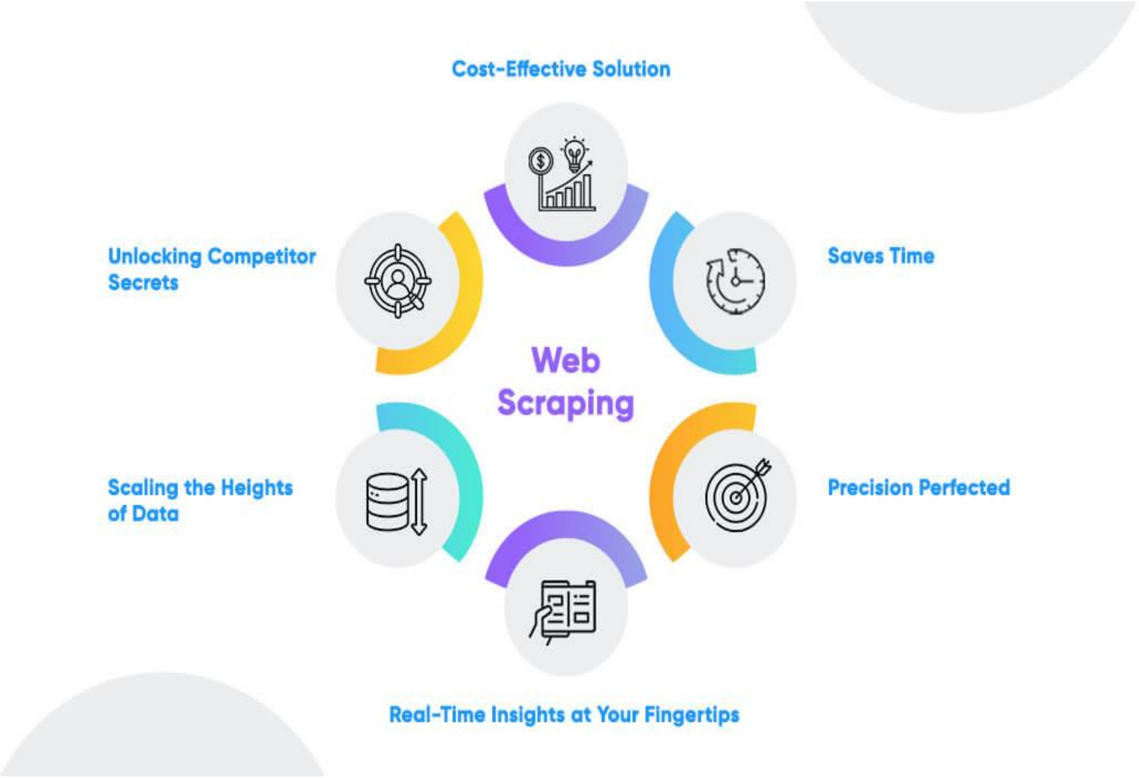

Cost-Effective Solution
Unlocking Competitor Secrets
Saves Time
Web Scraping
Scaling the Heights of Data
Precision Perfected
Real-Time Insights at Your Fingertips

Fig:6.1 Advantages of web scraping

6.1.1 Cost-Effective

Web scraping services provide an essential service at a competitive cost. The data will have to be collected back from websites and analyzed so that the internet functions regularly. Web scraping services manage to do this in a cost-effective and budget-friendly manner.

6.1.2 Low Maintenance and Speed

Web Scraping does have a very low maintenance cost associated with it over a while. In that way, it helps plan the budget accurately. Also, scraping web saves a lot of time, as it can do a day's manual work in a few hours.

6.1.3 Data Accuracy

Simple errors in data extraction can lead to major issues. Hence it is needed to ensure that the data is correct. Data scraping is not only a fast process, but it's accurate too. This reputation helps while collecting important data such as sales price, financial data to name a few. storing data with automated software and programs, your company or employees will be able to spend no time copying and pasting data

6.1.4 Easy to Implement

Once a website scraping service starts collecting data, you can rest assured that you are getting data from not just a single page but from the whole domain. With a one time investment ,it can have a high volume of data.

6.1.5 Effective Management of Data

By storing data with automated software and programs, your company or employees will be able to spend no time copying and pasting data. So they can focus more time on creative work, for example. Instead of this tedious work, web scraping allows you to pick and choose which data you want to collect from various websites and then use the right tools to collect it properly. Moreover, using automated software and programs to store data ensures that your information is secure.

CHAPTER 7
DISADVANTAGES

7.1 Data Analysis of Data Retrieved

To analyze the retrieved data, it needs to be treated first. This often becomes a time-consuming work.

7.2 Difficult to Analyze

For those who are not much tech-savvy and aren’t an expert, web scrapers can be confusing. Even though it’s not a major issue.

7.3 Speed and Protection Policies

Most of the web scraping are slower than API calls. Many websites don’t allow screen scraping. It is a huge challenge to Web Scraping. Also, if any code of the target website gets changed, web scrapers stops capture the data.

7.4 Data Analysis

Processing the extracted data through web scraping can be a time-consuming and energy-intensive process. This is because the information comes as HTML code and that can be difficult for some to read. Don’t worry, though, there is software that can take care of that too!.

7.5 Website Changes and Protection Policies

Because websites’ HTML structures change regularly, your crawlers will sometimes break. Whether you use web scraping software or write your own web scraping code, you’ll need to perform some maintenance periodically to ensure your data collection pipelines are clean and operational.

Moreover, it’s a good idea to invest in proxies if you want to do data scraping or crawling on multiple pages on the same website. Sending plenty of HTTP requests from the same IP in just a few moments looks suspicious and it could get the IP banned. If you have a proxy pool, though, each request can come from a different IP.

7.6 Learning Curve

Web scraping is not just about one way of extracting data. And here, I mean only one tool or the most appropriate method. Whether you use a visual web scraping tool, an API, or a framework, you’ll still have to learn the ropes. This can sometimes be difficult, depending on the knowledge level of each user.

CHAPTER 8
CONCLUSION

The user to log in and do new searches in addition to viewing previously conducted searches and results. When using normal methods through browsers or article sites, literature search can take weeks to complete. The developed program makes it possible to complete this task considerably more quickly. Furthermore, when conventional procedures are followed, the application built gives the user access to content that they might otherwise overlook but find useful. This increases the literature review's effectiveness and scope. Academicians and graduate students in our department have tested and begun using our program, which can be searched using a search engine that may be customized by the user as needed. The established system makes it possible to access the vast majority of previously published studies in the topic under investigation, and it streamlines and regularizes the process of conducting a literature review. Web scraping is a very popular technology to get some information from the web. In the literature there are few examples of academic research being scraped. They did this kind of study for analysis on the sites. There is no example for dynamic scraping of academic research sites by user keywords. But we give a solution, an application to scrape academic research by user's own. The user can scrape academic sites at any time with his own keywords by using our "Smart Literature Search App". The developed application allows users to conduct literature searches quickly and easily. While our work provides valuable insights into the development and implementation of a web scraping.

CHAPTER 9
FUTURE ENHANCEMENT

The field of web scraping for literature research is continuously evolving, driven by advancements in technology and the increasing demand for efficient data collection methods. This section discusses potential future trends and technological advances that could further enhance the effectiveness and efficiency of web scraping applications. Future web scraping applications are expected to leverage more advanced AI techniques, such as machine learning and natural language processing (NLP), to automate and improve the accuracy of literature searches. AI-powered web scrapers can better interpret complex web pages and extract relevant data more accurately. Additionally, NLP can provide a deeper contextual understanding of search queries, resulting in more relevant and precise search results. Advancements in UI and UX design will make web scraping tools more accessible and user-friendly for researchers with varying levels of technical expertise. Interactive dashboards can provide intuitive data visualization, and personalized search features can allow users to customize and save their search preferences, improving the overall user experience.

As web scraping technologies advance, it is crucial to adhere to legal and ethical guidelines. Future improvements may include compliance automation tools that ensure web scraping activities comply with legal requirements and website terms of service. Additionally, implementing ethical scraping practices will help protect privacy and sensitive information.The integration of advanced technologies and adherence to ethical practices will play a significant role in the future of web scraping for literature research. These advancements have the potential to enhance the efficiency, accuracy, and accessibility of literature reviews, benefiting researchers across diverse academic disciplines.

REFERENCES

[1] Gunawan, R. et al. (2019). Comparison of web scraping techniques: regular expression, HTML DOM and Xpath. In International Conference on Industrial Enterprise and System Engineering.

[2] Sirisuriya, D. S. (2015). A comparative study on web scraping. In the Proc. 8th Int. Res. Conf. KDU, 135– 140.

[3] Spangher, A. and May, J. (2021). A Web Application for Consuming and Annotating Legal Discourse Learning. arXiv preprint arXiv.

[4] Phan, H. (2019). Building Application Powered by Web Scraping. Doctoral Thesis.

[5] Saleh, A. I. et al. (2017). A web page distillation strategy for efficient focused crawling based on optimized Naïve bayes (ONB) classifier. Applied Soft Computing.

[6] Tharaniya, B. et al. (2018). Extracting Unstructured Data and Analysis and Prediction of Financial Event Modeling. In Conference proceedings of the Annual Conference IET

[7] Boegershausen, J. et al. (2021). Fields of Gold: Web Scraping for Consumer Research. Marketing Science Institute Working Paper Series.

[8] Saranya, G. et al. (2020). Prediction of Customer Purchase Intention Using Linear Support Vector Machine in Digital Marketing. In Journal of Physics: Conference Series, IOP Publishing, 1712(1):012024.

[9] Nguyen, V. H., Sinnappan, S. and Huynh, M. (2021). Analyzing Australian SME Instagram Engagement via Web Scraping. Pacific Asia Journal of the Association for Information Systems, 13(2):11-43.

[10] Deng, S. (2020). Research on the Focused Crawler of Mineral Intelligence Service Based on Semantic Similarity. In Journal of Physics: Conference Series, IOP Publishing