



DATA ANALYSIS BY WEB SCRAPING

S.Soumya¹, D.Shravani¹, V.Shivani¹, L.Swathi¹, G.Ranjith kumar², Dr.R.Jegadeesan³

¹Students of IV B.Tech, Department of CSE, Jyothishmathi Institute of Technology & Science, karimnagr(TS)

²Assistant Professor, Department of CSE, Jyothishmathi Institute of Technology & Science, karimnagr(TS)

³Associate Professor, Head of CSE department, Jyothishmathi Institute of Technology & Science, karimnagr(TS)

ABSTRACT:

The standard information investigation is built on the root and impact relationship, shaped an example minuscule examination, subjective and quantitative examination, the rationality approach of creating extrapolation examination. The Web Scraper's conniving ethics and procedures are juxtaposed, it explains about the working of how the scraper is premeditated. The technique of it is allocated into three fragments: the web scraper draws the desired links from web, and then the data is extracted to get the data from the source links and finally stowing that data into a csv file. The Python language is implemented for the carrying out. By doing so, linking all these with the moral knowledge of libraries and working know-how, we can have an adequate Scraper in our hand to produce the desired result. Due to an enormous community and library resources for Python and the exquisiteness of coding chic of python language, it is most appropriate one for Scraping desired data from the desired website.

Keywords: Data Analysis, Web Scrapping, data scrape.

[1] INTRODUCTION

Data analysis is the method of extracting solutions to the problems via interrogation and interpretation of data. The analysis process consists of discovering problems, resolving the accessibility of suitable data, determining which method can help in finding the solution to the interesting problem and convey the result. For the purpose of analysis, the data has to segregate into various steps further on such as starting with its specification assembling, organizing, cleaning, re-analyzing, applying models and algorithms and the final result. Web information scraping [1] and publicly supporting are outstanding strategies for naturally creating substance

on the web. A considerable number of individuals utilized these strategies in research and business for creating substance or offering criticisms to expand the exactness of business advertising that enables individuals to deliver resources in advancing and developing the business [2] by and large, web scraping is notable for a "Screen Scraping", "Web Data Extraction". The web scrubber programming is planned to be exhaustive for all noteworthy data from different online stores and mining, and collecting it into the new website. The scraper tool for the web is utilized for derived information from the web host, and as a portion of uses used for web orders, web mining and data mining, online esteem change observing and value correlation, element survey scratching (to watch the challenge), gathering land postings, atmosphere data checking, webpage change area, inspect, following on the web closeness and reputation, web mash up and, web data joining. [3] Pages are manufactured utilizing content-based increase dialects (HTML and XHTML), and much of the time contain a profusion of cooperative info in the content structure. Be that it may be as most website pages are anticipated for human end users and not for minimalism of robotized use. Thus, the toolbox that scrapes web info was made.

[2] LITERATURE SURVEY

Renita Crystal Pereira et. al.,[1] provided web scraping summary and techniques and tools that face several complexities as data extraction isn't that simple. These strategies guarantee that the data collected is correct, consistent and has better integrity, because there is a large amount of data present which is hard to handle and retain. Although there are a few problems faced by functional techniques that can be such as the elevated amount of web scraping be able to cause rigid harm to the websites. The measurement level of the web scraper will vary with the measurement units of the original source file, making it very difficult to interpret the data.

Eloisa Vargiu, Mirko Urru [4] discussed misuse web scraping in a synergetic filtering-based accession to web broadcasting. Usually, we accept various web Promotion fields that affects the Web Page The proposed system, depends on the synergetic filtering by exploiting peer pages and, subsequently, it resorts to Web scraping to perform the page content Analysis. Even we have showed the Case Diagram for the knowing the process. i.e., referring the Backgrounds of the Web Page of a German Portal. Till now it is the most recent technique used in web scraping for Web Broadcasting. As for the future work, we are setting up experiments aimed at calculating the performances of the proposed system in term of precision at k, i.e., the ability of the system in suggesting k relevant ads.

Yun Fei Xu, Jingnong Weng, Ananta Raj Sharma, Dilshod Yussupov[5] suggested Web Content Aggregation Service on the basis Of Geospatial Web Content Aggregation service and also the Area and Level of Data Extraction. It has all its emphasis on Data Aggregation. In this there are thousands of Users on the Website everyday which is very difficult to know the Actual User and fake users. The given result is based on the Digital Earth. Content Acquisition is done on the Precise Data on the Website, as Internet is very Common Medium for collecting any Sought of Information so there are Chances of Exploitation of Data.

Debahuti Mishra and Niharika Pujari [6] proposed data drives present's businesses and the internet is a Powerful origin of information. Data combiner gives the user with a complete view of all mixed data sources. The foundation service given by data integration is query processing. But if we are considering a query that include multiple domains, then we find that generic purpose search engines failure to provide solution of such query. Such queries and domain specific search services cover complete only one domain. Hence presently the only answer to this challenge is to pose the query distinctly to devoted services and feed the result of single as input to extra. Our thought can be tense from the task in data integration, wherever two foundation methods has been scheduled to involve the mapping between global ontology and a regular of services, that are GAV and LAV.

Robert Baumgartner, Sergio Flesca, and Georg Gottlo, [7] presented Extracting useful information from the web is the most significant issue of concern for the realization of semantic web. This may be achieved by several ways among which Web Usage Mining, Web Scrapping and Semantic Annotation plays an important role. Web mining enables to find out the relevant results from the web and is used to extract meaningful information from the discovery patterns kept back in the servers. Web usage mining is a type of web mining which mines the information of access routes/manners of users visiting the web sites. Web scraping, another technique, is a process of extracting useful information from HTML pages which may be implemented using a scripting language known as Prolog Server Pages (PSP) based on Prolog.

[3] PROBLEM STATEMENT

The world of retail is changing rapidly. Many brick and mortar locations are closing and being replaced by online stores, direct to consumer brands, and subscription services. However, while the breadth of assortment is something that drives customers to website, a lot of E-Commerce platforms fail to sell through a high percentage of merchandise. In this paper we propose a reliable, convenient and accurate detection system. Our study has the following specific objective:

The Objective is to get the information from different sources with the assistance of programming known as web crawler Scrapy. The software is used to extract data using an application programming interface or as a general-purpose web crawler required by the desired customer to analyze the variation, comments, ratings or anything else with innumerable options.

[4] EXISTING SYSTEM

The Existing system is the manual web data extraction process has two major problems. Firstly, it can't measure costs efficiently and can escalate it very quickly. The data collection costs increase as more data is collected from each website. In order to conduct a manual extraction, businesses need to hire large number of staffs; this increases the cost of labor significantly. Secondly, each manual extraction is known to be error prone. Further, if any business process is very complex then cleaning up the data can get expensive and time consuming. The below figure explains the errors and data cleanup processes problems with the Manual method.

[5] PROPOSED SYSTEM

The proposed system is a **web scraper** that is able to access and extract data from websites using a web application as an interface for user interaction. The extracted data is then stored in a database, as the web application allows the user to search through and query the saved findings.

[6] ARCHITECTURE

Web Scraping is a strategy to separate organized information from sites. WSAPI is the stage that empowers an association to expand their current electronic framework, too planned arrangement of administrations for making new channels, designer mix or accomplice joining. It assists with offering spotless and organized information from existing sites, so the information can be easily devoured by unique frameworks. The innate

plan assists engineers with fusing site changes without influencing the extraction rationale by moving them to designs. There are numerous particular reasons why organizations might need to scratch their site; one of the essential explanation being the inaccessibility of APIs.

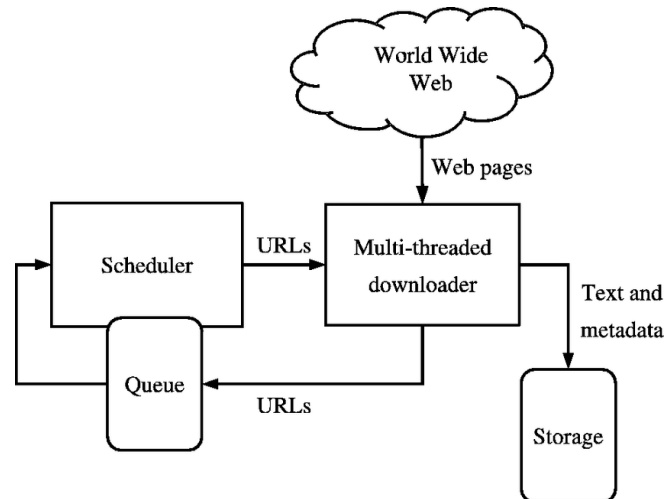


Figure 1 : Architecture of web scrapping

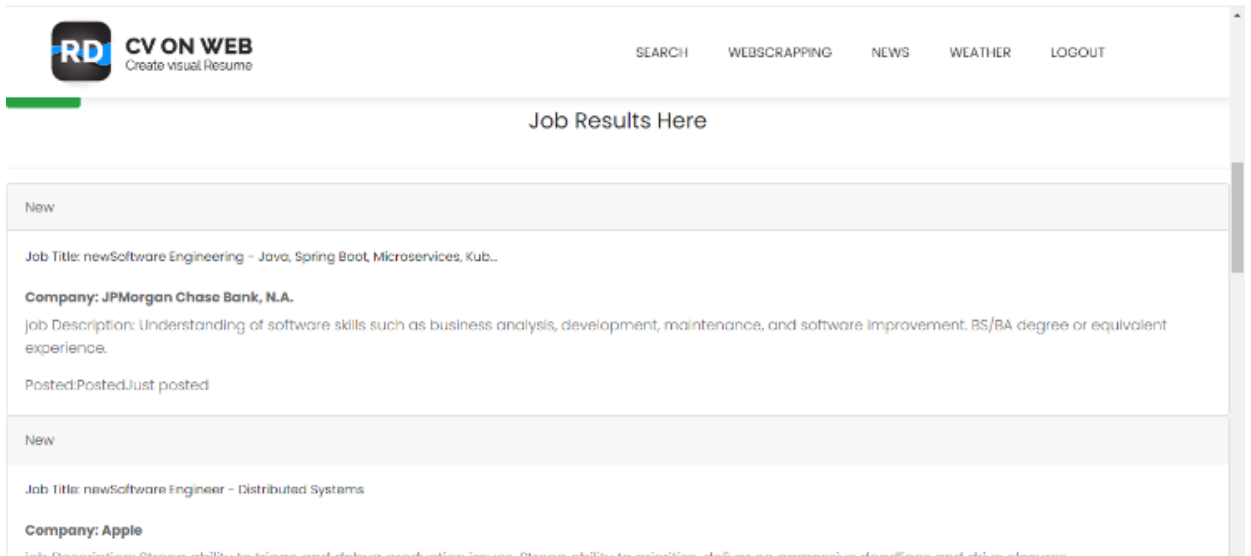
[7] IMPLEMENTATION

1. Identify the target website.
2. Collect URLs of the pages where you want to extract data from.
3. Send an HTTP request to the URL of the webpage you want to access. The server responds to the request by returning the HTML content of the webpage. For this task, we will use a third-party HTTP library for python-requests.
4. Once we have accessed the HTML content, we are left with the task of parsing the data. Since most of the HTML data is nested, we cannot extract data simply through string processing. One needs a parser which can create a nested/tree structure of the HTML data. There are many HTML parser libraries available but the most advanced one is html5lib.
5. Now, all we need to do is navigating and searching the parse tree that we created, i.e. tree traversal. For this task, we will be using another third-party python library, BeautifulSoup. It is a Python library for pulling data out of HTML and XML files.

The result will be having following modules.

User:

The User can register the first. While registering he required a valid user email and password for further communications. Once the user registers, then admin can activate the customer. Once the admin activates the customer then the customer can login into our system. After login he can search all the company's details. For searching the company details we will get company rating and reviews and total number of employees based on our dataset. After that login if we click on web scrapping, we can find the job portal based on our title and job location.in the job portal completely it provides job description and requirements of the particular company.



The screenshot shows the CV ON WEB website interface. At the top left is the logo 'RD CV ON WEB' with the tagline 'Create visual Resume'. To the right are navigation links: SEARCH, WEBCRAPPING, NEWS, WEATHER, and LOGOUT. Below the navigation is a green bar with the text 'Job Results Here'. The main content area displays two job listings. The first listing is for 'newSoftware Engineering - Java, Spring Boot, Microservices, Kub...' at 'JPMorgan Chase Bank, N.A.'. The second listing is for 'newSoftware Engineer - Distributed Systems' at 'Apple'.

Result 1 : Job Portal



The screenshot shows the MIRACLE NEWS website. The header features the text 'MIRACLE NEWS' in green and a 'back' link. Below the header is a large banner with the text 'Latest News using webscraping' in white and blue. In the center of the banner is a 'BREAKING NEWS!' logo. Below the banner is a photograph of a grocery store aisle with price tags for '\$369', '\$1.2', and '\$4.09' visible. A sign for 'PLUMS' is also visible in the background.

Result 2:News

Find the Weather

Karimnagar, Telangana
Monday 12:58 pm
32°C
Cloudy

[back to home](#)

Result 3 : Weather

Admin:

Admin can login with his credentials. Once he login he can activate the users. The activated user only login in our applications. The admin can set the data set by the company details. In this report the data is considered as company reviews and company rating and headquarters and total number of employees. The admin can add new data to the dataset. So this data user can perform the testing process.

Rank	Company Name	Rating	Reviews	Type	Location	Employees
1	TCS	3.9	18.1k	Public	Mumbai	10000+ employees
2	Accenture	4	14.1k	Private	Dublin	10000+ employees
3	ICICI Bank	4.1	12.7k	Public	Mumbai	10000+ employees
4	Cognizant	3.9	12.1k	Private	Tecneck	10000+ employees
5	HDFC Bank	4	10.8k	Public	Mumbai	10000+ employees
6	Infosys	3.9	10.8k	Public	Bangalore	10000+ employees
7	L&T	4.1	10.2k	Public	Mumbai	10000+ employees
8	Capgemini	3.3	9.5k	Private	Paris	10000+ employees
9	Tech Mahindra	3.5	9.1k	Public	Pune	10000+ employees
10	HCL Technologies	3.7	8.7k	Public	Noida	10000+ employees
11	Tata Motors	4.1	8.3k	Public	Pune	10000+ employees
12	IBM	4	8.2k	Private	New York	10000+ employees
13	Reliance jio	4	7.9k	Public	Navi Mumbai	10000+ employees
14	Genpact	4	7.6k	Private	New York	10000+ employees
15	Axis Bank	4	7.6k	Public	Mumbai	10000+ employees
16	Wipro	3.8	6.5k	Public	Bangalore	10000+ employees
17	Praxaris Parikh Consulting	4	6.5k	Public	New Delhi	10000+ employees

Result 4 : Web Scrapping Details

[8] CONCLUSION

Extracting data through scraping technology is a new evolving activity in the technology harvesting arena. Though many companies are still using manual process of extracting data but Web Scraping solutions will transform the traditional method of extracting data. The day is not that far with exponential growth throughout this field when it can become a phenomenon and most companies will understand the value of scraping innovation and how it enables them remain ahead in the race dramatically. This paper presents the survey of Web scraping technology incorporating what it is, how it works, the popular tools and technologies of web scraping, the websites used for this technology and the top most fields which are making use of this technology.

REFERENCES

- [1] Renita Crystal Pereira, Vanitha T. “WebScraping of Social Networks.” International Journal of Innovative Research in Computer and Communication Engineering, vol. 3, pp.237-239, Oct. 7, 2018”
- [2] Bellarosey. “Crowdsourcing-Definition.” Internet: http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html, Jun. 02, 2006”
- [3] Ghazvinian, Holbert, Viswanathan. “Simple WebScraping.” Internet: <https://seanolbert.wordpress.com/2011/07/15/scrappy-simple-webscraping/>, Jun. 2015”
- [4] Eloisa Vargiu , Mirko Urru “Exploiting web scraping in a collaborative filtering-based approach to web advertising” December 5, 2012
- [5] Yunfei Xu, Jingnong Weng, Ananta Raj Sharma, Dilshod Yussupov” Web Content Acquisition in Web Content Aggregation Service Based on Digital Earth Geospatial Framework” Beihang University Beijing 100191, China P.R
- [6] Debahuti Mishra and Niharika Pujari “Cross-Domain Query Answering: Using Web Scraper and Data Integration.
- [7] Robert Baumgartner, Sergio Flesca, and Georg Gottlob, 'Visual web information extraction with(lixtol)', In VLDB Journal, Proceedings of 27th International Conference on Very Large Data Bases, September 11-14, 2001, pp. 119-128.
- [8] Naveen Ashish and Craig Knoblock. Wrapper Generation for semi-structured Internet Sources. In Proc” ACM SIGMOD Workshop on Management of Semi Structured Data, Tucson, Arizona, May 1997.”
- [9] Datahen.”3 Advantages of web scraping for your enterprise” Internet: <https://www.datahen.com/3-advantages-web-scraping-enterprise/>, May. 17, 2017”
- [10] https://en.wikipedia.org/wiki/Web_scraping”
- [11] <https://www.quora.com/What-is-the-legality-of-web-scraping>”
- [12] https://en.wikipedia.org/wiki/Web_crawler
- [13] Kolari, Pand Joshi A. , “Web mining :research and practice , Computing in Science & Engineering”, IEEE Transactions on Knowledge and Data Engineering, vol. 6, no. 2, Vol. 6 , No. 4, 2004”
- [14] Bright Planet.com Deep web White Paper. <http://www.completeplanet.com/Tutorials/DeepWeb/index.asp>.”