

Smart Job Search Engine Using Web Scrapping

Saumya Sinha, Rashee Saxena, Hitesh Kr Garg

*Department of Electronics and Communication Galgotias college of engineering and technology
Gr. Noida, India*

*Department of Electronics and Communication Galgotias college of engineering and technology
Gr. Noida, India*

*Department of Electronics and Communication Galgotias college of engineering and technology
Gr. Noida, India*

Submitted: 25-05-2021

Revised: 01-06-2021

Accepted: 05-06-2021

ABSTRACT—With the progression in innovation and prominence of web based business, the quantity of job searching sites have been expanded quickly in the digital world. This made individuals' life simple in this present COVID scenario where people are losing jobs daily and its very difficult to find jobs with preferred field and location. Be that as it may, this too bring exertion for individuals as they invest a great deal of energy and endeavors to search best jobs for them in different sites. They need to channel and look at information without help from anyone else. It takes a ton of time and still there are odds of questionable outcomes. This paper depends on web automation and web scrapping applied for distinguishing best arrangements from the site indeed.com. The scrapping scripts are composed of python libraries and web crawling that works on HTML labels. The difference in this work is that we are not keeping scrapped data on local database. Rather the outcomes are fetched and showed each time the client input the data. Besides, the information recovery measure precision is 95% with least calculation and less time .

Keywords— Web Scrapping, Web Automation, Web drivers, DOM.

I. INTRODUCTION

The web world is incredibly wealthy as far as useful information furthermore, substance open in different organizations like numbers, text, pictures, sound and video and so on website pages which prompts abnormality in recovery of data on account of its unimportance for which the client looking for . The enormous measure of information can be favouring and hostile too. The developing interest of information to ask people to become new strategies and innovations with the objective that the entrance of information can be rapid and straightforward. The cycle of data recovery depends on putting away information on information base, returning data as indicated by user field and location. By and by,

53.7% of Internet utilized by individuals searching for information about product or organizations, 47.4% information is looked for instructive reason, 39% substance are looked for wellbeing and clinical information , 27.9% for job purpose. Through this users easily find their suitable jobs without visiting different sites. With the acclaim of Internet, the quantity of job searching sites has immediately developed , and this engages people to search adequately through the Internet. Users put huge loads of energy in looking doable jobs. Due to difficulty in searching from different sites, user need to have a single site where he can easily get all the required data. Web scrapping is a process to extract content and data from a website. It extracts underlying HTML code and with it data stored in a database. It is a effective way of getting information for working with massive data on web on the web. Web Scrapping is a strategy utilized to gather a ton of data from sites and sent to your local framework. The technique of web scrapping can be divided into two phases progressively:

- 1) Getting the web sources
- 2) Fetching of important information from acquired sources.

Customarily, this sort of projects mimicked user investigation of web by executing hypertext transfer protocol (HHTP), or introducing a totally fledged web program, for instance, Google chrome. Web harvesting is firmly identified with web indexing, which records data on the web using a bot or web crawler and is an inclusive method implemented in most web search tools. Additionally it stimulates user browsing using application software. Web Scrapping isn't new to us, for what it's worth getting more famous these days due to the new on the web business and Start ups, as they don't have to do a lot of steady work to get the data. Ideally, they utilized scratched information from other comparable sources and the change it as indicated by their need. The utilizations of web scratching are by and large noticing climate information, research information

gathering, looking for wellbeing data and finding intriguing examples for business or web information incorporation..

II. LITERATURE REVIEW

Information extraction using usage mining, web scraping and annotation[1]. In this paper, we return to, investigate and examine some data extraction techniques on web like web utilization mining, web rejecting and semantic explanation for a superior or productive data extraction on the web delineated with examples. Web scratching, another technique, is a cycle of removing helpful data from HTML pages which might be executed utilizing a scripting language known as Prolog Server Pages(PSP) in light of Prolog.

Exploiting filtering approach with web scrapping for smart online shopping[2]. This paper depends on web crawling and scratching techniques applied for distinguishing best arrangements from five online business sites. The structure is planned utilizing HTML (Hypertext markup language) and CSS (Cascading template) as front-end and PHP: Hypertext preprocessor language as back-end support. the quantity of web based shopping sites have been expanded quickly in the digital world. This made individuals' life simple since it is anything but difficult to shop through internet .But they need to channel and look at information without anyone else. It requires some investment and still there are odds of uncertain outcomes.

An Overview On Web Scraping Techniques And Tools [3]. Because of trade, offer and store information on web, another issue is emerge that how to deal with such information over-burden and how the client will get or get to the best data in least endeavors. To tackle this issues, scientist spotout new procedure called Web Scraping. Web scratching is basic strategy which is utilized to create organized information based on accessible unstructured information on the web.

A review on web scraping and its applications [4]. This paper will focus in on different parts of web scratching, starting with the essential presentation and a concise conversation on different programming's and apparatuses for web scrapping. We had likewise clarified the cycle of web scratching with an elaboration on the different kinds of web scratching procedures lastly finished up with the advantages and disadvantages of web scratching and an in detail depiction on the different fields where it very well may be applied.

Web crawling based search engine using python [5]. The purpose of the paper is to eliminate the data from various sources with the help of programming known as the web crawler Scrapy using the programming language Python variation 3.6. The Database is made which gathers all the unstructured information from different sources and afterward dissects them by the logical cycle of its particulars, collecting, coordinating, cleaning, re-examining, applying models and calculations lastly giving the ideal outcomes.

Data analysis by web scraping using python[6]. The purpose of the paper is to eliminate the data from various sources with the help of programming known as the web crawler Scrapy using the programming language Python variation 3.6. The Database is made which gathers all the unstructured information from different sources and afterward investigates them by the insightful cycle of its determinations, amassing, arranging, cleaning, re-examining, applying models and calculations lastly giving the ideal results.

Pro Circle: A promotion platform using crawling and web data scraping technique[7]. there is a ton of advertising advancements distributed on different sites every day. Individuals frequently need to look through a few sites to discover the advancements that they want. Our versatile application, 'Pro Circle', was created as a stage to gather advancement news into one spot. We utilized web information scratching strategy for scratching advancement news from solid sites. Likewise, we additionally applied publicly supporting strategy to get supporting information of advancement news which can be quickly evolved from publicly supporting will permit the aggregate hunt of advancement data in our Pro Circle platform..

Cross-Domain query answering :Using web scraping and data Integration [8]. The basic service provided by data integration is query processing. But if we are considering a query that involves multiple domains, then we find that general purpose search engines fail to answer such queries and domain specific search services cover only one domain.the only solution to this problem is to pose the query separately to dedicated services and feed the result of one as input to another.

Increased information retrieval capabilities on e-commerce websites using scraping techniques [9]. It takes a procedure that can accumulated data from numerous sources into a single element to encourage the cycle of data retrieval. This

examination utilize 3 internet business site as a wellspring of information, by using slithering technique will create new factor they can store information from source data, at that point these information will put away in a database.

Board forum crawling: A web crawling method for web forum[10]. The strategy begins creeping from the landing page, and afterward enters each leading body of the site, and afterward slithers all the posts of the site straightforwardly. Board Forum Crawling can slither most significant data of a Web gathering website productively and simply They tentatively assessed the adequacy of the strategy on genuine Web gathering destinations by contrasting and the conventional broadness first slithering.

Criminals And Missing Children Identification Using Face Recognition And Web Scrapping[11]. This undertaking proposes to utilize this innovation for distinguishing crooks who are on the run from their past records. These hoodlums can be distinguished by the face acknowledgment from a picture or video outline which is caught by the cameras which are introduced in different areas and it can likewise be utilized for recognizing missing youngsters.

Web Scraping And Data Acquisition Using Google Scholar[12]. This paper attempts to set up an interface that would use web scraping techniques and Python modules to link a researcher's list of publications present on Google Scholar to a MySQL database and Excel application, allowing them to access and manipulate their works in minimal steps.

Web Information Retrieval Using Python and Beautiful Soup[13]. In this paper, we have created technique for recovering web data utilizing Beautiful Soup and python content. Wonderful Soup is apparatus for web data recovery. The majority of the web data presents in unstructured configuration. The proposed framework recovers the unstructured information in client's example and makes it helpful.

Advanced deep web crawler based on Dom[14]. Because of the way that enormous measure of information today must be put away in profound web. Considering the work done by others on profound web crawlers, it is wiped out that no ideal, or even total crawlers for profound web information has been made. To address the issues of profound web search, we have worked out another structure of crawler, right now concerned most on removing

information from structures - the most well-known kind of profound web interface.

Testing using Selenium Web Driver [15]. This paper centers around the utilization of Selenium Web driver to test web application and to exhibit the utilization of hardware in mix with different apparatuses like the Maven, TestNG, and so on, for more simpler way to deal with testing and to improve the nature of testing process.

III. METHOD

The framework is executed to assemble a site that look for skill based preferred jobs through HTML DOM-based design by utilizing web scratching and web crawling methodologies. We utilized two internet jobs site i.e Naukri.com and indeed.com . Both these sites are analyzed on the basis of your skills. The web content will be scrapped whenever user input the asked data.

A. Working Principle

The web application is actualized by following advances given underneath:

1. Import the Python libraries
2. Bringing the URL using request and selenium libraries and save it into temporary variable
3. Parse the HTML in temp variable and convert it into BeautifulSoup4 design
4. Scrap job post and location
5. Look at the job preference
6. View scratched information according to input field and location

The first two stages go under the umbrella of web crawling that is finished by utilizing python libraries and third and fourth steps are known as web scratching

B. Web Scrapping implimentation using python

In this usage we utilized python as coding language since python gives quick and amazing libraries and it offers local area uphold for web crawling and scrapping. We used request, Beautiful Soup 4 and Web Selenium Driver python libraries for scratching for various stages.

1. Python Requests

For opening and passing HTTP URLs python, demand library is imported. Requests Library is a direct and straightforward to use HHTP library written in python. It makes correspondence with web administrations reliable and helps in association pooling. This licenses to proceed with boundaries and treats over all demands that created utilizing the meeting event. Python requests encode the boundary naturally and interprets the reaction in Unicode. On the other hand, various document sharing could be dealt with.

2. Beautifulsoup4

It is a library that brings information of the page either in HTML or XML design. In mix with parser it creates a parse tree dependent on DOM (Document Object Model) approach and afterward unique channel capacities can be applied to discover specific tag, string, characteristics or blend of all. Let talk about the cycle initially the archive to be parsed is given as contention to lovely soup work at that point the particular report is changed over to UNICODE and components in it become UNICODE characters. These characters are given as contribution to parser naturally HTML parser runs anyway you can likewise determine the parser you need to utilize. At the point when we talk about this library, it considers archive and its components as items for example labels, safe String, excellent soup and remark.

3. Selenium

Selenium is best of its sort system for testing, it gives full help to numerous programs like Google, Firefox and so on It gives many testing activities to testing web applications. Selenium is a web driver that supports web pages that are dynamic in nature for example they have capacity to uphold a page whose components may change without reloading of page itself . Despite the fact that it has been utilized to make web tests for online application however it very well may be utilized for pages that have JavaScript on them. It is accessible in Python Web driver Selenium bundle.

C. Website module

I. User Interface

We planned a straightforward and easy to use interface. By utilizing this interface client can get the preferred job by adding the job preference and location. The result shows images, skills, and preferred jobs according to information source.



Fig.1.User Interface

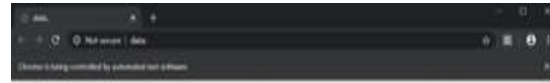
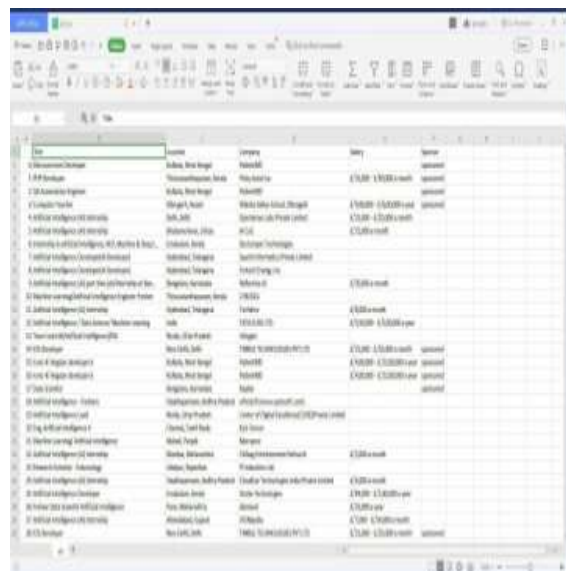


Fig.2.Automated web browser



Job ID	Job Title	Location	Salary	Status
1	Software Engineer	Atlanta, GA	\$75,000 - \$100,000/year	open
2	Data Analyst	Atlanta, GA	\$50,000 - \$70,000/year	open
3	Product Manager	Atlanta, GA	\$80,000 - \$100,000/year	open
4	Marketing Specialist	Atlanta, GA	\$40,000 - \$55,000/year	open
5	Business Development	Atlanta, GA	\$60,000 - \$80,000/year	open
6	Operations Manager	Atlanta, GA	\$70,000 - \$90,000/year	open
7	Systems Administrator	Atlanta, GA	\$55,000 - \$75,000/year	open
8	Quality Assurance	Atlanta, GA	\$45,000 - \$60,000/year	open
9	Customer Support	Atlanta, GA	\$30,000 - \$40,000/year	open
10	Human Resources	Atlanta, GA	\$40,000 - \$55,000/year	open
11	Finance Analyst	Atlanta, GA	\$50,000 - \$70,000/year	open
12	Legal Counsel	Atlanta, GA	\$100,000 - \$150,000/year	open
13	Public Relations	Atlanta, GA	\$45,000 - \$65,000/year	open
14	Sales Representative	Atlanta, GA	\$35,000 - \$50,000/year	open
15	Project Manager	Atlanta, GA	\$70,000 - \$95,000/year	open
16	UX Designer	Atlanta, GA	\$65,000 - \$90,000/year	open
17	Business Development	Atlanta, GA	\$60,000 - \$80,000/year	open
18	Marketing Specialist	Atlanta, GA	\$40,000 - \$55,000/year	open
19	Operations Manager	Atlanta, GA	\$70,000 - \$90,000/year	open
20	Systems Administrator	Atlanta, GA	\$55,000 - \$75,000/year	open

Fig.3.Desired job results

II. Business logic

The application layer is liable for recovering chosen ascribes from site. It depends on exceptional contents that have been written in python and utilizations BeautifulSoup library to parse information. The web application is liable for connecting with client and introducing them required outcomes. The working progression of web application and information scratching is appeared in fig.1 and fig.2 and 3. The cycle begins by client produced inquiry this question is inserted with URL to discover a least 6 most minimal value occurrences of the specific item from each web space and got the HTML marks. After discovering examples there information and metadata including cost, name, picture and item particular are gotten for cost correlation by utilizing web scratching. The HTML names that have been gotten from brought page will be thus parsed utilizing HTML DOM (Document Object Model). DOM parsing used by

By incorporating web program with the objective that the program can take dynamic substance created from web content on the client side. Information that has been successfully parsed will be select with the marking in each target site and looked at set variable as needs be case taken from a web space is contrasted and each other example of different spaces lastly 5 cases with most minimal cost of all are shown as consequence of looked through item on our web application.

Increased information retrieval capabilities on e-commerce websites using scraping techniques.

IV. RESULT AND DISCUSSION

This research actualizes a web application that gives facility of getting to and choosing the best job for you. This application is need of the time as in this covid situation various of us are loosing their job. In view of remaining burden and human race users need more an ideal opportunity to get to various destinations to get the ideal job in minimum time. Prior to moving towards results, legitimacy issues should be taken in thought while actualizing such kind of web application. This webpage is getting to information of various web areas to think about the best job they are offering to users. The inquiry that is it lawful to get to their information or not emerges yet as far legitimacy issue is thought of, this application is getting to just information that is on the web interface of internet business destinations and as indicated by reality that it is free for any individual who is getting to this site it is clear that no infringement is finished. Other significant issue to be considered is refreshing of web spaces by their executives.

The web spaces that have been gotten to are refreshed time by time by their managers interim, to get to these refreshed changes, the technique that has been utilized is getting to page each time the client inquiry for a specific job. With the goal that each time refreshed page is utilized for additional cycle. This system improves our application by diminishing memory necessity, as nearby information base isn't needed to save information intermittently. Alongside this mistake in feeling of item accessibility is likewise diminished i.e., on the off chance that we have utilized information base and a individual pursuets a job which is unavailable in web area in any case, present in our information base shows wrong outcomes and produce equivocallness. At the point when site page from each web space for an exchange started by client is scratching.

V. CONCLUSION

The web is wealthy in term of huge information. With the entry of time, the information

burial chamber is expanding. The need emerges to get data utilizing web index which prompts more bother in precise discoveries from the unique sources over web and it is realized that the substance accessible on pages talks a great deal on massive topics. Web scrapping methods could assist with diminishing this issue Here, a framework is created to get your dream job. Be that as it may, People who are doing Scratching should consider that they are not breaking any sort of law which could make them subject for any offense. In our situation, we approach just information that is available by all watchers so it makes execution of this application legitimately substantial. In future, we can develop mobile application to facilitate mobile users.

REFERENCES

- [1]. P. Lambrix, "towards a semantic web for bioinformatics using ontology-based annotation", in: proceedings of the 14th IEEE international workshops on enabling technologies: infrastructures for collaborative enterprises, 2005, pp. 3-7.
- [2]. V. Bhagwan and T. Grandison, "injection," in 2009 IEEE international conference on web services deactivation, 2009, pp. 2-3
- [3]. Deepak Kumar Mahto, Lisha Singh, a dive into web scraper world, 2016 international conference on computing for sustainable global development (indiacom), 2016 IEEE.
- [4]. Osmarcastrillo-Fernández, "web scraping: applications and tools" European public sector information platform topic report no. 2015 / 10, December 2015.
- [5]. S. Amudha, "Web Crawler for mining web data" in International research journal of engineering and technology volume: 04 issues: 02, Feb. 2017.
- [6]. "Renita Crystal Pereira, Vanitha T. "web scraping of social networks." International Journal of innovative research in computer and communication engineering, vol. 3, pp.237-239, oct. 7, 2018".
- [7]. "Rohita Chopra Prem, Vinit D. "web scraping of social networks." International Journal of innovative research in computer, vol. 3, pp.237-239, oct. 7, 2018".
- [8]. Robert Baumgartner, Sergio Flesca, And Georg Gottlob, 'visual web information extraction with (lixto)', in vldb journal, proceedings of 27th international conference on very large data bases, september 11-14, 2001, pp. 119-128.
- [9]. .M.Mangala, M.B. Chandak, C. Nekita, "information retrieval system and machine

- translation : a review", proceeding computer science 78(2016) 845-850.
- [10]. J R. Baeza-Yates And C. Castillo. "crawling the infinite web: five level are enough". proceedings of the third workshop on web, 2004.
- [11]. Apoorva.P, Ramesh.B And Varshitha.M.R "automated criminal identification by face recognition using open computer vision classifiers" third international conference on computing methodologies and communication(ICC WC 2019).
- [12]. C. Yang, Y. Yan And Q. Zhu, "the face database development of science and technology expects based on web mining," 2012 fourth international conference on multimedia information networking and security, Nanjing, 2012, pp. 388-391.
- [13]. Pang, B., & Lee, L. (2004, july). a sentimental education: sentiment using subjectivity summarization based on minimum cuts. in proceedings of the 42nd annual meeting on association for computational linguistics (p. 271).association for computational linguistics.
- [14]. A.Ntoulas, P. Zerkos And J. Cho, "downloading textual hidden web content through keyword queries," proceedings of the 5th ACM/IEEE-CS jointconference, pp. 100-109, 2005.
- [15]. A. Holmes ; Digital Focus, Herndon, Va ; M. Kellogg "automating functional tests using selenium" agile conference, 2006.