Web-Scraping

Vidit Agrawal[1], Swapnil Shinde[2], Prajwal Thokal[3], Mohammadsufiyan Shah[4]
[1, 2, 3, 4] Dept Of Electronics And Telecommunication Engineering
[1, 2, 3, 4] Pimpri Chinchwad College Of Engineering

Abstract- *The main objective of web scraping is to take out information from one or many website and helps to store into way simple form such as spreadsheet, database or CSV file. When it comes to a very complicated task, web scraping can be used as the time consuming resource and mainly when it is carried out manually. Since previous studies has generated some of the automated solution.*

I. INTRODUCTION

Web scraping can be also known as web extraction which is used to extract data from the one of the website that you want and can be save in the form of spread sheet or database. Generally web data is scraped from the web browser or from Hypertext Transfer Protocol (HTTP). This is executed manually by a user or automatically by a bot. As we know, there is enormous amount of data available on the WWW and here web scraping comes as the main factor to collect the large data from the website. As the usage of web a new marketing and sales channel the quantity if content by the user. Online merchants offers large number of data to describe their products. Knowledge base providers also grant access to their database for learning.

II. IDENTIFY, RESEARCH AND COLLECTIDEA

There are many different required functions, modules, software and platform for the web scraping , some of them are

- Beautiful soup: Beautiful soup is a python library. It get from xml, HTML and markup language. It helps to get a particular content from web pages and remove the HTML markup language .It used as tool to clean up and paste document. It helps in isolating titles and links, to get all the text from HTML tags, to alter the HTML with the document.
- Install: To install beautiful soup you need to install pip or python compiler. First open command prompt to install. Write 'pip install beautiful soup. Beautiful soup create hierarchical and readable manner of parse tree ehich extract data from code beautiful soup is very fast, prettify the source code, extremely lenient. "prettify()" give the visual representation of parse url.
- CSV FILE:CSV file is also known as common separated values. It uses OD contexts for stimulating the creation of open government world-wide. The OGP launched in2011. CSV is not used in INDONESHIA because there are only PDF file. CSV is machine-readable format, basic non-proprietary format, simple tabular format.CSV is just like a text file, in a human format which is extensively used to store tabular data, in a spared-sheet or database. The separator character of CSV files called a delimiter. Default delimiter is comma (,).Other delimiters are tab(t) ,colon(:) character. Each record consist of field separated by commas.
- Advantages of CSV file: Easier to create, preferred import and export format for database and spreadsheet capable of storing large amount of data.
- Python CSV module: CSV module provide too types of objects: reader-to read from the CSV files writer-to write in to the CSV file to import CSV module in our programme ,write the following statement :import CSV Opening\Closing CSV file Open a CSV file: f=open("stu,CSV","W") or f=open("stu.csv","r") close a csv file: f.close() Role of argument newline in opening of SCV file : Newline argument specifies how would python handle new line character while working with CSV file, on different operating system. Different operating system store EOL character differently. Writing in CSV files: csv.Writer()=returns a writer object which writes data into csv files.write row()=write one row of data on the writer object. Role of writer object: THE csv.writer() function returns a writer object that converts the users data into a delimited string this string can later be used.
- Pandas: Pandas is the name derive from Panel data. Pandas is a free software under three clauses BSD. Python library pandas is used for data analysis. Pandas is built by matplolib and numpy. mataplotib is used for data visualization. numpy is used for mathematical operation use of this is we can write brief code into short pandas have two new type of object for storing data that makes easier series and data forms pandas is the data sheet data forms store data in the table format like rows and columns like a spreadsheet data frames is used find average per column combine form a table.by using group by() function you can apply array method to sub groups for example split- apply-combine. Pandas is used to reshaping and pivoting datasheet, subsisting, label-based slicing. pandas having some factors like drat set joining, data filtration, data lagging, moving windows, linear regeneration numpy, scipy have maturity and stability of fundamental numerical libraries. pandas give

scientific python to be more attractive and practical computing environment numpy and array strutted type's can be used to hold this data collection of independent columns data. frames class which store mixed data type. Pandas have data frames which implement functions of r counterpart. data structure of panda have a index object which store labeling information about tick.

III. WRITE DOWN YOUR STUDIES ANDFINDINGS

- **The URL generator block:**
  Explanation: The block contains the function customURL() definition. In this block we take two inputs from user which are job and location respectively. The input is immediately stripped before storing into a local variable in order to get data without any spaces in front or back. This is simply done to generate proper URL which will be then used to fetch the required HTML document. In the URL, each keyword is separated with "+". Hence, if job profile consists of more than one word, replace function in python will replace blank space with "+" operator. Example : Job = "Data scientist" the resultant job variable will be "Data+scientist". Finally the processed local variables job and FinalURL are appended to home page link of indeed.com to create custom job Search URL.

- **URL call and custom functions:**
  Explanation: The get() function from requests library will use the custom URL to make a call to indeed.com server. The HTTP request method used is "GET". The BeautifulSoap() method call will use Html parser to parse the fetched document and will allow python To read the various components of page. The function cleanhtml() is written to remove html elements inside the text. For this the re(regular expression) library is used. It will clean the data to represent it properly on html document. The function input_not_type_hidden() takes tag as input and returns all elements inside it except for certain type of attribute

  Explanation: The function extract_job_title_from_result() takes in input argument "soup" object generated by beautifulsoap with use of html parser. The function will target specific html elements of the fetched html document to scrape the required data. In our case, the document contains 15 entries of job profiles in single html document so to scrape data through every job section, we iterate same logic used for scraping one job profile section multiple times using for loop. The data from every html element is accessed using "text" attribute. The data is then appended to list in sequential order as a row. This row will have all details related to single job profile. Hence to scrape all job profiles and store, we will be using nested list structure where the parent list "job" will have list of "rows" inside it.

  Explanation: The library "io" which is used for file operations takes in name of file and action to open the file. The "w" shows that file is opened for writing. Now with help of csv library we will create a writeheader Which will be used to write the data to csv file from the nested list "FinalData". The panda library's function read_csv() is used to read csv file generated during last operation. The csv file is then generated to html file with the help of to_html() method which takes in input argument as name of the output file. Hence the output file "indexindeed.html" is generated which can be viewed in any browser.

- **You can access the entire code by clicking the link down below**

{https://github.com/SwapnilShinde47/Web-scraping-of-job-portal}

IV. LITERATUREREVIEW

Web Scraping Web Scraping is a very beneficial or useful technology in the field of scrap the data from the different websites or

WWW (World Wide Web).The great or notable features of web scraping are that , it can scrap that data or content which we required. Much more companies also used the web scraping for their business. With the help of Web Scraping All the machine learning industries scrap data or content.

The Website creators also get help of the Web Scraping to collect or scrap data from the different social media websites, and they can also see that what is trending and what is going on. Web Scraping is used in the one project in which it used to get the data from a particular category of shoes in amazon store. In another different project of web scraping is also used to get the data from the Twitter on the of hash tag or by searching the keywords on the twitter.

With the help of web Scraping in the technologies as market research using web data in any of the company or industries. Web Scraping also help in price comparing from different websites. In Advance these type of applications use the web scraping to get and scrap the data from the dependant websites.

In other way in government and private watchdog also take help the application of Web Scraping to monitor the malicious and unexpected activities going on the WWW (World Wide Web) or internet.

V. CONCLUSION

Since Web scraping can be used for the legit purpose and no intent to bring website down but if not careful while doing web scraping can lead to banned for using that website for the particular user.

REFERENCES

[1] https://towardsdatascience.com/python-and-worldliterature-elementary-web-scraping-with-beautifulsoupa43daaf46ab
[2] http://wthtjsjs.cn/gallery/1-whjj-june-5412.pdf
[3] https://www.ijert.org/call-for-papers
[4] https://www.youtube.com/watch?v=XQgXKtPSzUI&list=PLsZJ kZL-rNTsUGo6d2sd_8NxW_iOOcjuw&index=11&t=13s