WEB SCRAPING FOR VTU RESULTS

[1]Gauripriya Kakati, [2]Komal Revankar, [3]Pratidnya Kittur, [4]Asst. Prof. Shrikant Athanikar
[1,2,3]Computer Science and Engineering , S G Balekundri Institute of Technology
Belgaum, Karnataka, India
[1] gourirox@gmail.com, [2]komalrevankar4@gmail.com, [3]pratidnyakittur452@gmail.com
[4]athanikar.s@gmail.com

Abstract— Web scraping is essentially a mechanism that allows users to scrape and access data from websites and other internet sources. It combines software engineering technology and bespoke software programming to extract data or any other content from online sources, scrape a copy of the information, and preserve it in an external archive for review. Web scraping is often referred to as automated data collection, database exploration, database crawling, or content management mining. In this paper we discuss a Web Scraper Bot that scrapes the results from results.vtu.ac.in and automatically fills in the Captcha Code using Tesseract. The results will be parsed and stored in the form of Excel Files. This works for any VTU college.

Keywords—BeautifulSoup, ChromeDriver, OpenCV, Pandas, Pytesseract, Python, Selenium.

I. INTRODUCTION

Web scraping API is the framework that allows a college to expand its current internet-based infrastructure, as well as the collection of services designed to build new platforms, integrate designers or integrate collaborators. It allows delivering clean and organized data from current websites, such that different systems can access the data easily. Data disclosed over these APIs could be easily monitored, converted and managed. The underlying architecture lets developers implement improvements to the website without disrupting the structure of the abstraction by shifting them to modifications. Web Scraping Technology is the automated system used to necessarily make the automatic copying and pasting work. It often gathers vast quantities of data from websites such as directory pages, property investment pages, confidential pages and job sites and is stored on multiple servers. Web scraping could algorithmically optimize physical work by accessing each page, extracting information from websites, and html page decoding. There is also a range of Web Scraping Software on the marketplace which can enable to scrap information from every page you need.

A Web Scraper is used to scrape the results from results.vtu.ac.in and automatically fills in the Captcha Code using Tesseract. The results will be parsed and stored in the form of Excel Files. This works for any VTU college. Web scraping is a technique to fetch data from websites. Web Scraping is the automation of the data extraction process from websites. 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. PyTesseract or Python-tesseract is an Optical Character Recognition (OCR) tool for python. It will read and recognize the text in images, license plates, etc. Here, we will use the tesseract package to read the text from the given image.

II. EASE OF USE

Web scraping allows you to acquire non-tabular or poorly structured data from websites and convert it into a usable, structured format, such as a csv file or spreadsheet. Scraping is about more than just acquiring data: it can also help you archive data and track changes to data online.

Web Scraping is an automatic way to retrieve unstructured data from a website and store them in a structured format. For example, if you want to analyze what kind of face mask

can sell better in Singapore, you may want to scrape all the face mask information on an E-Commerce website like Lazada.

Web scraping just works like a bot person browsing different pages website and copy pastedown all the contents. When you run the code, it will send a request to the server and the data is contained in the response you get. What you then do is parse the response data and extract out the parts you want.

III. Literature Survey

Literature review is focused on the following works being done by an array scholar from the field of data security.

1. **Web Scraping of Social Networks:**

They 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.

2. **Detection of web scraping using machine learning:**

This technique is proposed in the paper where Web scraping solutions are aimed primarily at translating complex data obtained through networks into structured data that could be stored and examined in a central database. Web scraping solutions thus have a significant impact on the result of the cause.

3. **An Overview on Web Scraping Techniques and Tools:**

Web scraping is a quite important methodology used to produce structured data based on the unstructured data available on the internet. Scraping formed structured data, subsequently collected and evaluated in spreadsheets in central database.

IV. Problem Statement

The VTU website allows one student to view/obtain exam results at a time by entering the appropriate USN (University Location Number) in the user form. One of the traditional methods used by class teachers to collect data for 40 students is to print a result sheet for each student and then manually aggregate this data into an Excel spreadsheet. This gives rise to the consumption of more papers, large amount of time spent on data tabulation that is prone to errors during tabulation.

Solution:

A Web Scraper Bot

An intelligent way of handling things and is very much required in this space. An Algorithm that is the need of the day. People can expect a lot of accuracy in the results obtained. The work of gathering the required data will be completed within minutes/seconds.

Fig 1.1: Web Scraping Architecture

Web Scraping Bot (web harvesting or web data extraction) is a computer software technique to extract information from websites. Usually, such programming programs recreate human investigation of the World Wide Web by either executing low-level Hyper content Transfer Protocol (HTTP), or installing a completely fledged internet browser, like Internet Explorer or Mozilla Firefox. Web Scraping is firmly identified with web ordering, that lists data on the web utilizing about web crawler and is a widespread method received by most web indexes. Web Scraping is a technique to extract structured data from websites.

V. Methodology

The Python program contains a function, which takes a word as an argument. The URL of a certain website is used to navigate to the desired web page by using Selenium “webdriver” interface. We use the “headless” option in the webdriver to hide the browser automation process. The search bar of the web page serves as a navigation tool for changing the web page contents. It can be used by putting the HTML tag for the search bar in selenium code. We can send the desired search elements by using the “Keys” module in Selenium. The word taken as the argument is

then sent through Keys to the search bar. This word makes the website navigate to the desired web page. The current address of web page is taken as the source and all the required contents are then extracted using BeautifulSoup.

BeautifulSoup uses tags to identify the specific elements required for extraction. A Python dictionary element can be used to facilitate the proper storage and access of data elements. The dictionary element allows the data to be accessed in a structured manner using a key associated to the data element. During the execution of the Python program, an input is passed to the Scraper function that takes it as an argument and the data from the web page is extracted after function execution. We can now convert the extracted data into desired format using Pandas library. After the program execution, a data file is created in the previously determined format that can be accessed to view the stored information.

The world of Artificial Intelligence and machine learning has its common roots with data, which is primarily the most important entity on its own. Data has already impacted so many businesses worldwide and can never take a back seat when it comes to this technical world. To get access to data in its best form, web scraping was brought to use. Data provided on the internet is of so much use that the whole world is running after it. Web scraping was brought into practice long back and is still useful to date. This paper aims to make people aware of this technology to help them expand their knowledge. Tools and applications related to web scraping are also mentioned.

The World Wide Web is currently the largest data source in the history of mankind and consists of mostly unstructured data, which can be hard to collect. Extracting the World-Wide Webs unstructured data can be done with traditional copy-and-paste, as some websites provide protection against an automated machine accessing the website. However, this is a highly inefficient approach for larger projects. Sometimes websites or web services offer APIs to fetch or interact with the data. However, it is not uncommon that APIs are absent or that the available solutions do not cover the user needs. In essence, web scraping is used to fetch unstructured data from web pages and transform it to a structured presentation, or for storage in an external database. It is also considered an efficient technique for collecting big data, where gathering large amounts of data is important. Three different phases build up web scraping:

Fetching phase

First, in what is commonly called the fetching phase, the desired web site that contains the relevant data has to be accessed. This is done via the HTTP protocol; an Internet protocol used to send requests and receive responses from a web server. This is the same techniques used by web browsers to access web page content. Libraries such as curl 2 and wget 3 can be used in this phase by sending an HTTP GET request to the desired location (URL), getting the HTML document sent back in the response.

Extraction phase

Once the HTML document is fetched, the data of interest needs to be extracted. This phase is called the extraction phase, and the technologies used are regular expressions, HTML parsing libraries or XPath queries. XPath stands for XML Path Language and is used to find information in documents. This is considered the second phase.

Transformation phase

Now that only the data of interest is left it can be transformed into a structured version, either for storage or presentation. The process described above can be summarized in Figure 4.1

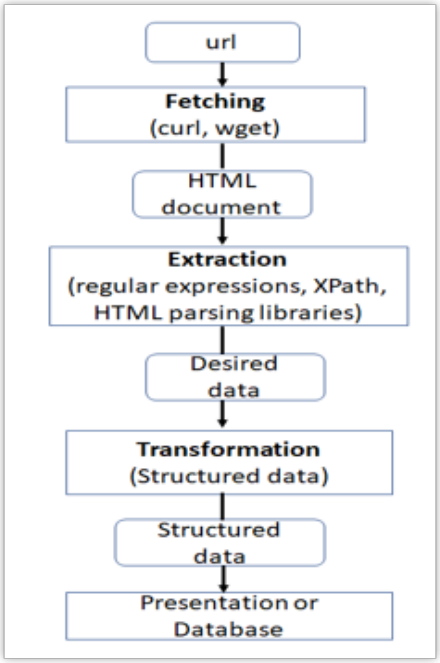

Fig 4.1: Web Scraping Process

VI. RESULT AND DISCUSSION

In web scraping, code reuse and maintenance are especially critical. Code reuse is important in web scrapping because when we develop program to mostly have some common feature that we used before. For example, for website data scraping, the program needs to access the website so if you built a good code to access a site using username and password, you can reuse it in a future project if you are going to run it in the same situation. Web scraping involves code maintenance because, no matter how reliably you construct your scraper, a significant change in the targeted site's design or behavior might make it extremely difficult for your program to find the appropriate data on the site. If you have a web scraping procedure that runs on a regular basis for any length of time, you'll need to be ready to make code changes to compensate for changes on the destination web site. Finally, Python is perfect for easy implementation. Web scraping is frequently required in order to go to more interesting aspects of a project. Simple web scraping in Python can be pulled up fast, especially if you have a large library of reusable code.

It was discovered that the web scraping has a wide variety of applications across the field of Business Intelligence. Using web scraping programs would involve a higher upfront cost, but would pay for themselves over and over again as a form of short-term loss, long term gain.

Although online scraping is an effective method for obtaining massive data sets, it has some ethical and legal consideration, scrapping may happen over copyrighted data and lead to copyright infringement furthermore aggressive spidering usually leads to large number of requests to sites that may cause loss of services which is naturally undesirable by site owners. The use of robots.txt allows some form of self-compliance in this method website owners place a text file in their root folder that lists all the files they forbid from being crawled. Compliance isn't necessarily enforced in these cases but ethical scrappers should abide by site owner wishes by abiding with robots.txt

The main difficulties that need to be examined before apply web scraping:

1. Are the data collected from human subjects? If yes, is data scraping ethical?
2. Is there an API available on the website?
3. Is web scraping permitted on the website?
4. Is the data presented in the form of an HTML table?
5. Is it simple to select CSS selectors using the Selector Gadget?
6. Is there any data that is not numerical? If yes, how manipulee is it?
7. Would the scraping procedure entail iteration across numerous pages? how much data do you intend to scrape if that is true? just a sample or the entire site?

VII. CONCLUSION

Extracting data through scraping technology is a new evolving activity in the technology harvesting arena. Though many colleges 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 colleges and companies will understand the value of scraping innovation and how it enables them remain ahead in the race dramatically.

In today's era, one can find the emergence of a new paradigm in every couple of years, so is the emergence of web scraping. This paradigm has its roots in the requirement of analysis of structured as well as unstructured data. There are various aspects related to web scraping. Some of them have been discussed in this paper. Initially we have discussed various applications of web scraping. This paper is majorly focused on tools of web scraping. The availability of these tools has made helped a lot of entrepreneurs to expand their business.

VIII. FUTURE SCOPE

A web scraper which is appropriately intended and executed, could assist analysts prevail over obstacle to data access, gather online information more resourcefully, and eventually respond investigation queries that cannot be answered by conventional means of assortment and examination."

One major issue which will need to be addressed very soon is the legal standing of web scraping. As with so many internet-based technologies, it's hard to define what is and isn't legal based on laws that predate the internet. As we move forward into the 21st

century, legal bodies around the world have continued to adapt themselves to the information age, and soon a consensus will have to be reached on whether any or all instances of web scraping violate the right to information privacy.

IX. REFERENCE

[1] Renita Crystal Pereira and Vanitha T, "Web Scraping of Social Networks," International Journal of Innovative Research in Computer and Communication Engineering, pp. 237-240, Vol. 3, 2015.

[2] Kaushal Parikh, Dilip Singh, Dinesh Yadav and Mansingh Rathod, "Detection of web scraping using machine learning," Open access international journal of Science and Engineering, pp.114-118, Vol. 3, 2018.

[3] Sameer Padghan, Satish Chigle and Rahul Handoo, "Web Scraping-Data Extraction Using Java Application and Visual Basics Macros," Journal of Advances and Scholarly Researches in Allied Education, pp. 691-695, Vol.15, 2018.

[4] Anand V. Saurkar, Kedar G. Pathare and Shweta A. Gode, "An Overview on Web Scraping Techniques and Tools," International Journal on Future Revolution in Computer Science & Communication Engineering, pp. 363 367, Vol. 4, 2018.

[5] Federico Polidoro, Riccardo Giannini, Rosanna Lo Conte, Stefano Mosca and Francesca Rossetti, "Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation," Statistical Journal of the IAOS, pp. 165-176, 2015.

[6] Jan Kinne and Janna Axenbeck, "Web Mining of Firm Websites: A Framework for Web Scraping and a Pilot Study for Germany," 2019.

[7] Ingolf Boettcher, "Automatic data collection on the Internet," pp. 1-9, 2015.

[8] Erin J. Farley and Lisa Pierotte, "An Emerging Data Collection Method for Criminal Justice Researchers," Justice Research and statistics association, pp. 1-9, 2017.

[9] Anjali Khute, Yash Roy, Yamita, Yashmeen Xalxo, "Dynamic Web Scraping Using Python"