IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Voice Assistant Based Corona Virus Web Scraper

[1]Indranil Amar Sawant, [2]Rohan Vikas Mohite, [3]Shubham Kishor Pawar

[1,2,3] U. G. Student, Department of Information Technology, Finolex Academy of Management and Technology, Ratnagiri, Maharashtra, India

Abstract: Corona virus pandemic has been recognized as a global threat across the world and many methods were adopted for the prevention of this disease. This pandemic has caused global and economic disruption which resulted in numerous Covid-19 cases across the world. To know the number of cases and keep a track of this pandemic situation we need to collect the live data sets from the worldwide corona virus records. This can be achieved by the technique of Web Scraping which enables the extraction of live data sets from a specific platform. It facilitates the user to access the World Wide Web wherein specific data is gathered, copied from the web and then it is stored in a central local database then provides ways to retrieve and analyse the data. This research focuses on the implementation of Voice Assistant based corona web scraper and explains how a single system is easy to use by all users. This system is useful for get the information regarding corona virus updates. The user can read and listen covid updates through voice assistant.

Index Terms **– Web scraping, Covid-19, Voice assistant**

I. INTRODUCTION

With massive outbreaks and news spreading faster than the Coronavirus (COVID-19) itself, we have to be aware of the actual stats that are affecting the areas that we live in Our project will help to get Covid-19 updates that happens in our areas as well as all over the world. The proposed system gives the updates regarding covid 19 virus. The user can read the covid updates, listen them through voice assistant, and able to search the covid updates whatever region they want. This system is to design a platform where you can obtain the live data sets and have a compact knowledge about the present scenario. This is an elementary approach to scrap the live data sets through a user interface from the "Worldometer" Covid-19 data set with the aid of a voice assistant. To implement this scheme, we use the Python programming language. To effectuate this task, we acquire the process of making API calls to the "Worldometer" Covid-19 website and simultaneously we will make use of the regular expressions to extract the data from the web page. However, this action includes a series of tactics that has to be recognized analysed and accomplished sequentially. Initially the input is given by the user. Then the required contents are searched and matched with the user's input. If the contents match then with the help of a voice assistant result is obtained which is the output in turn.

II. LITERATURE REVIEW

Python has rich set of libraries that are available to extract digital contents from the internet. Amongst the libraries available, the following three are popularly known: BeautifulSoup, LXml and RegEx. A statistical study was carried out on the datasets that was available, it shows that RegEx was capable to deliver the requested data on an average of 153.6 ms. But RegEx has that drawbacks of limited data extraction for web pages with more of HTML inner tags. Due to this demerit RegEx is used to perform only moderate complex data extraction. Some of the other libraries like BeautifulSoup and LXml are able to extract content of the web pages under complex environment which yielded a response rate of 457.66 ms and 203 ms respectively. These libraries are based on the Document Object Model (DOM) proved to be accessible libraries. Web scrappers are the one by which regional languages available in social media are influenced and hence modern content grading system are developed. The survey conducted in this paper proves the overwhelming performance of RegEx under varying situations [1].

The main goal of data analysis is to obtain the information which is useful from data and take decisions on the basis of data analysis. Web scraping refers to collecting data from the web. Web scraping is also popularly known as data extraction. For the purpose of analysis the data can be divided into several steps such as cleaning, organizing etc. Scrapy is most popularly used open source for collecting the data that is needed by the user. The main purpose of using scrapy is to scrape the data from its sources. Scrapy, which is a web crawler and based on a python programming language, is very helpful to get the data which we require by making use of URLs which are necessary for scraping data from its sources. Web scraper is an API which is useful to get data from a website. Scrapy provides all the tools which are necessary for extracting the data from a website and then process the data as per the needs of the user and store the data in specific format as specified by the users [2].

Regular expressions can be used in extraction, classification and validation of data. In order to understand the concept of regular expressions one must have good knowledge on it. Our main aim is to learn about how the regular expression is changing with time. If the scope of regular expressions increases as time goes on then we need to match more strings outside the regular expressions language. Otherwise if its importance does not increase as passes then we need to focus on matching strings inside language of regular expressions. It is a language which can be used in describing a set of strings which can be matched by it, and

generally there can be numerous ways for expressing this. The regular expressions can be generated from huge number of strings which are labelled or it can be generated by using the previously existing ones. Usually the editing in regular expressions involves changing various features, and usually the adding of new features is more when compared to deletion of the existing features. Based on the range of the features which are added newly and the features which are removed from the existing ones we can construct mutation operators in new regular expression which will be generated and it also provides guidelines for the process of mutation generations. The editing in regular expressions however do not affect escape characters such as backslash, dot, question mark etc. In general, the process of defining operators in mutation depends mainly on the changes which are likely to be present among features of regular expressions and also depend on changes which are most likely to be present inside regular expressions features. Mutation testing: The ability of the generic set of the strings in exposing all the faults which are possible to be present in the regular expression which is user tested [3].

1. Web Scraping

Web Scraping is a process of fetching and mining for the essential data by scrawling through a web page. Web scrapers function in an ideal way wherein the content of the page maybe parsed, searched or reformatted. Then the data which is collected is copied to a spreadsheet or it can also be stored in a database for further analysis.

With the end goal of analysis, the data needs to be segregated into different advances further on, for example, beginning with its specification collection, organizing process, cleaning process, redissecting, applying different models and various algorithms and the eventual outcome. There are two ways of extracting data from websites, first one is manual extraction technique and the second one is automated extraction technique. Web scrapers assemble site information similarly to how a human would do that is the scraper goes onto a webpage of the site, gets the pertinent data, and push ahead to the following website page.

Each website has an alternate structure that is the reason web scrapers are generally built to search through a website. Web scraping can assist in obtaining any sort of information that is intended. We would then have the option to retrieve, analyse and utilize the information in the manner we need. So web scraping streamlines the way towards deriving information, speeds it up via automation and makes simple to access the extracted data by offering it in a CSV pattern.

Web scraping commonly extracts a lot of data from websites for example, monitoring interests of consumers, monitoring price i.e. value observing, advancing AI models, monetary information accumulation, tracking news, and so forth. Hence no doubt that web scraping is a programmed technique to acquire a lot of data from websites. Web pages contain information in a unstructured format in a HTML design which is then changed over into organized information in a spread sheet or a data set so it tends to be utilized for different applications.

Web scraping requires two sections in particular the crawler and a scraper. The crawler is a man-made AI algorithm that parses the web to look through the specific information needed by following links over the internet. The scraper, is a particular tool made to extract data from the sites.

2. Data analysis

Data analysis is a method to extract solutions to the problems by interrogation and interpretation of data. The web scraper program is planned to be thorough for all significant information from various online stores and mining, and gathering it into the new site. A web scraper works as an API to extract data from a website here it is Worldometer in which data is accessible free of cost to end users.

Web data scraping techniques are used by extensive number of people used in exploration and business for making content or offering reactions to grow the precision of business for promoting that empowers people to convey assets in progressing and building up the business.

3. Technologies

The project mainly covers the fundamentals of web scraping, voice assistance using python. It makes use of a tool called Parse Hub.

3.1 API source

We will be scraping from the very famous website for statistics which is Worldometer that is regularly updated with the coronavirus information.

3.2 Web Scraper

It makes use of a tool called Parse Hub. Parse Hub is a free web scraper that is very powerful and is easy to use. This tool allows you to scrap the web merely just by clicking on the elements you would like to get out from that website. The Parse Hub is clearly very efficient and uses an Artificial Intelligence technology to understand which elements you would want.

3.3 Python Libraries

The implementation is done by the programming language Python. It has great tools for extraction of data. It requests library for retrieving content from a web page, and bs4 (BeautifulSoup) for extracting the relevant information. These 2 libraries are often used wherein first a GET request is made to a website. Then a BeautifulSoup object is created from the content that is returned and parses it using several methods.

3.4 Regular Expression

Regular expressions are used for matching patterns. The basic idea for applying regular expressions is to define a pattern that is to be matched in a string and then search in that string to return the pattern matched. Regular expressions come up while parsing string information.

Regular Expression is a unique succession of characters that causes match or search different strings or strings set, utilizing a specific syntax to match a pattern. Pattern matching is search for substring in a given string. Regular expressions are used in validation, extracting and classification. Writing and understanding regular expressions requires to have knowledge and experience. If a single character written wrong would cause different pattern to be matched. Regular Expressions are basically a highly specific programming language that is embedded in Python and made accessible through the re module. Utilizing this little expressions, we can indicate the principles for the arrangement of strings that needs to be matched; this set may contain English sentences, or email patterns, TeX commands, number matching and many more.

Python offers two diverse operations dependent on these expressions: Match that checks for a matching pattern just towards the beginning of a string and Search that checks for a matching pattern that can exist anywhere in the string. The Python module re offers full help for Perl-like expressions for pattern matching in Python. The re module raises the special case re-error if a mistake happens while accumulating or utilizing regular expressions. Pattern matching is checking whether a particular grouping of characters, sentences or tokens which exists among the given data. For this situation regular expressions permits us to locate the particular patterns of words and obtain the data we need more effectively than physically looking for explicit characters in the website. Pattern matching is used for checking a given group of tokens for the presence of certain pattern. A traditional approach to perform pattern matching is to look for a sub string in a given string.

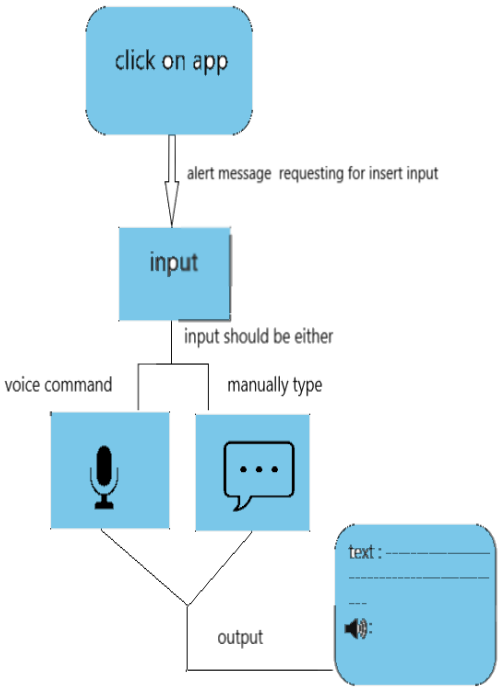
It can be got via looking a "sub string or pattern" through the given string. Contrasting and the other algorithms used for pattern matching RE can uphold in a more extensive way to describe patterns. A Regular Expression is called as regex or regexp, which is string and its layout depicts a bunch of strings. With these strategies the execution of this project is worked out. The web scraper program is anticipated for comprehensive for all critical information from various online sites, collecting and mining. The tool used for scraping is programmed to derive a lot of sensible data from the web.

III. Proposed Work

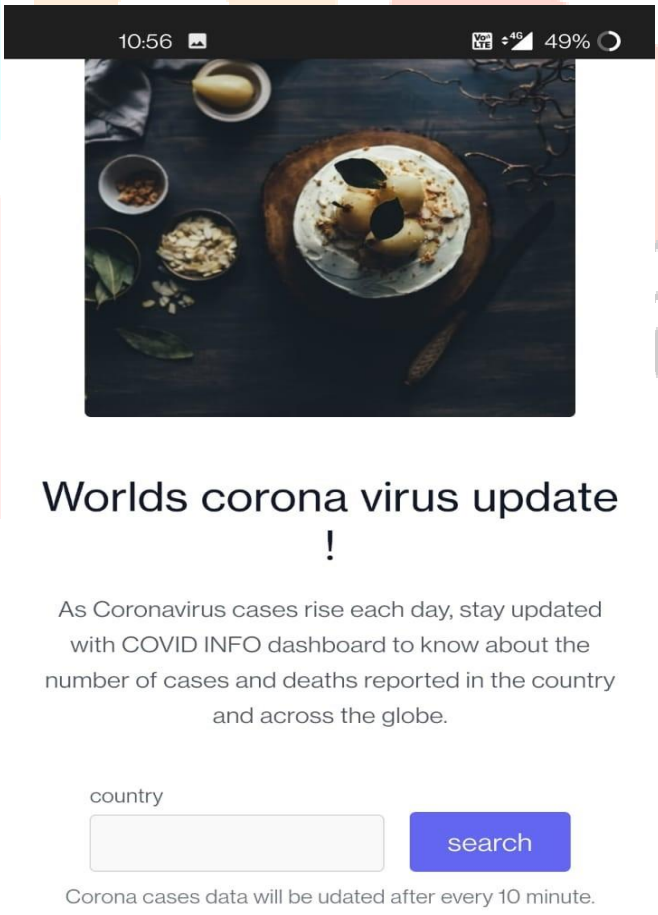
In this approach, by utilizing the Worldometer COVID19 dataset to fetch the data on Covid-19 updates and also for optimizing the search results. The use ParseHub, a tool that is free of cost for web scraping and a GUI for manually selecting HTML tags from the web page. Scraping the user requested data which is either in the form of voice or text.

IV. Working

➢ After starting the app the screen will display alert message on the screen asking for input form the user.
➢ Input should be given either by voice command or by manually typing the value
➢ Once the input has been stored (the input should be in String), the app will look for necessary data related to the input.
➢ After the processing, the screen will display the expected data on the screen and the user can hear the data as well from a programmed voice assistance.
➢ The app uses a flexible python module that makes the app useful in every means.
➢ This app can take only a single input at once and will be running continuously till we exit the app.
➢ Headphones are recommended while using the app. ➢ Make sure the microphone has been enabled while using the app.


click on app
alert message requesting for insert input
input
input should be either
voice command
manually type
text :
output

V. Result


10:56
49%
Worlds corona virus update !
As Coronavirus cases rise each day, stay updated with COVID INFO dashboard to know about the number of cases and deaths reported in the country and across the globe.
country
search
Corona cases data will be udated after every 10 minute.

VI. CONCLUSION

The main conclusion of this project is that Web scraper is a technology with potential in the field of fetching a particular field data and show it to the user. We studied web scraping at many different level. The Worldometer Coronavirus dataset can be gathered from true reports, straightforwardly from Government's correspondence channels via local media sources when considered as reliable. A group of experts and analysts who approve information from a consistently developing list of more than 5,000 sources may be feasible for this site to accumulate data in an efficient manner. The input given by the user is analysed and scraped from the website and output is produced using Google voice assistant in the form of speech. The output could be produced either in the form of text or speech. This approach is simple and straight forward to extract the number of cases in a particular state or country, the number of deaths in a particular country or state and the number of recovered patients all over the globe, country or a state

VII. ACKNOWLEDGMENT

Dr. Vinayak A. Bharadi, our Head of Department, Prof. Priyanka Bandagale, our Project Coordinator and Prof. Amar Palwankar, our Project Guide, have all been instrumental in the successful completion of our project. We are really appreciative and would like to convey our heartfelt gratitude and appreciation for all of their efforts and assistance, as well as their on-going counselling and guidance, which aided us in completing the project and gaining further information through research and study.

REFERENCES

[1] THIVAHARAN. S, SRIVATSUN. G AND SARATHAMBEKAI. S, "A SURVEY ON PYTHON LIBRARIES USED FOR SOCIAL MEDIA CONTENT SCRAPING", PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON SMART ELECTRONICS AND COMMUNICATION (ICOSEC 2020) IEEE XPLORE PART NUMBER: CFP20V90-ART; ISBN: 978-1-7281-5461-9, PP: 361-366.

[2] David Mathew Thomas and Sandeep Mathur, "Data Analysis by Web Scraping using Python", Proceedings of the Third International Conference on Electronics Communication and Aerospace Technology [ICECA 2019] IEEE Conference Record # 45616; IEEE Xplore ISBN: 978- 1-7281-0167-5, PP: 450-454.

[3] Peipei wang, Gina R. Bai and Kathryn T. Stolee, "Exploring Regular Expression Evolution", SANER 2019, Hangzhou, China Research Papers IEEE, PP: 502-513.

[4] Shreya Upadhyay, Vishal Pant, Shivansh Bhasin and Mahantesh K Pattanshetti "Articulating the Construction of a Web Scraper for Massive Data Extraction", 2017 IEEE.