DATA ANALYSIS BY WEBSCRAPING USING PYTHON

Syed Hussain Mohiuddin[1], Mohammed Atifuddin Hussain[2], Owais Mohammed Abdul Razzaq[3], Mrs. L. Vaishnavi[4]

[1,2,3]B. E Student, Department of CSE, ISL College of Engineering, India.

[4]Assistant Professor, Department of CSE, ISL College of Engineering, Hyderabad, India.

ABSTRACT:

The standard information investigations are constructed on the basis of the root and impact connection. These investigations are fashioned as an example of tiny inspection, subjective and quantitative examination, and the rationality way of producing extrapolation examination. This article compares and contrasts the deceitful ethics and practices of the Web Scraper. It also provides an explanation of how the scraper is prepared in its functionality. The method is broken down into three parts: first, the web scraper is used to get the needed links from the internet; second, the data is extracted in order to obtain the data from the source links; and third, the data is stored in a csv file. For the purpose of carrying out, the Python programming language is used. We are able to have an acceptable Scraper in our possession if we do this, combining all of them with the moral understanding of libraries and the working knowledge that we have. This will allow us to generate the output that we want. The Python programming language is the most suitable option for scraping necessary data from the website of one's choice because of the vast community and library resources that are available for Python, as well as the exquisiteness of the coding style that Python has.

Keywords – Data analysis, Web Scraping, Implementing Web Scrape.

INTRODUCTION

The technique of obtaining answers to issues via the questioning and interpretation of data is referred to as data analysis. The process of analysis includes the identification of issues, the resolution of the accessibility of appropriate data, the determination of which approach may assist in the discovery of a solution to the fascinating problem, and the communication of the outcome. In order for the data to be analyzed, it must first be separated into a number of different phases, such as beginning with its specifications, assembling, organizing, cleaning, re-analyzing, applying models and algorithms, and then arriving at the final result. Web information scraping and public support are two excellent methods that may be used to generate content on the web in a natural way. These tactics were applied by a significant number of persons in the fields of research and business for the purpose of producing content or providing critiques in order to enhance the precision of company advertising, which allows individuals to provide resources for the purpose of expanding and growing the firm. The terms "Screen Scraping" and "Web Data Extraction" are the most common ones that are associated with web scraping. The programming for the web scrubber is intended to be comprehensive for all significant data from various online retailers and mining, and it will be collected into the new website. The web scraper tool is utilized for a variety of purposes, including but not limited to the following: 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 mashup, and web data joining. For example, the scraper tool is utilized for web orders. The majority of the time, pages are created by applying content-based development languages (HTML and XHTML), and the content structure often contains a great deal of information that may be considered collaborative. Although this may be the case, the majority of internet pages are designed with human end users in mind, and not with the intention of minimizing the usage of robots. As a result, the toolkit that scrapes information from the web was developed.

LITERATURE SURVEY

Data Analysis By Web Scraping Using Python:

This paper depicts a standard data examination are based on the root and effect relationship, molded a model little assessment, abstract and quantitative assessment, the judiciousness approach of making extrapolation assessment.

The method of it is dispensed into three parts: the web scrubber draws the ideal connections from web, and afterward the information is extricated to get the information from the source joins lastly stowing that information into a csv document. Because of a gigantic local area and library assets for Python and the impeccableness of coding stylish of python language, it is most suitable one for Scraping wanted information from the ideal website

Web Scraping Using python:

Learn web scratching and creeping procedures to get to limitless information from any web source in any organization. Ideal for developers, security experts, and webmanagers acquainted with Python, this book trains essential web scratching mechanics, yet in addition digs into further developed subjects, for example, investigating crude information or utilizing scrubbers for frontend site testing

Web Scraping with Python Successfully scrape data from any website with the power of Python:

The Internet contains the most helpful arrangement of information at any point collected, generally openly open free of charge. Notwithstanding, this information isn't effectively reusable. It is implanted inside the design and style of sites and should be painstakingly separated to be valuable. Web scratching is getting progressively valuable as a way to effortlessly assemble and sort out the plenty of data accessible on the web. Utilizing a straightforward language like Python, you can creep the data out of complex sites utilizing basic program

Feasibility and Application:

It is essential to do data extraction and analysis in order to understand the underlying logic and purpose of the data. Extraction is necessary to consistently and genuinely get information before the interpretive step, rather than using it as a replacement for predicting the information's relevance. We need extraction pertaining to the articles that exist in diverse arrangements and use unique ways of reporting. The need to include the primary information elements of interest and to provide standardization. Additionally, to facilitate the identification and examination of patterns. Data analysis is crucial for gaining knowledge of data resources by specifically addressing relevant concerns. It illuminates the subject by offering surveys, planning for statistical graph creation and revamping, and other related activities. Scrapy is a web scraping framework. Scrapy is a web crawler tool that efficiently navigates websites and extracts structured data. This data may be used for several

important purposes, such as data mining, information processing, or data logging. Below is a visual representation of the scrapy structure to enhance comprehension.

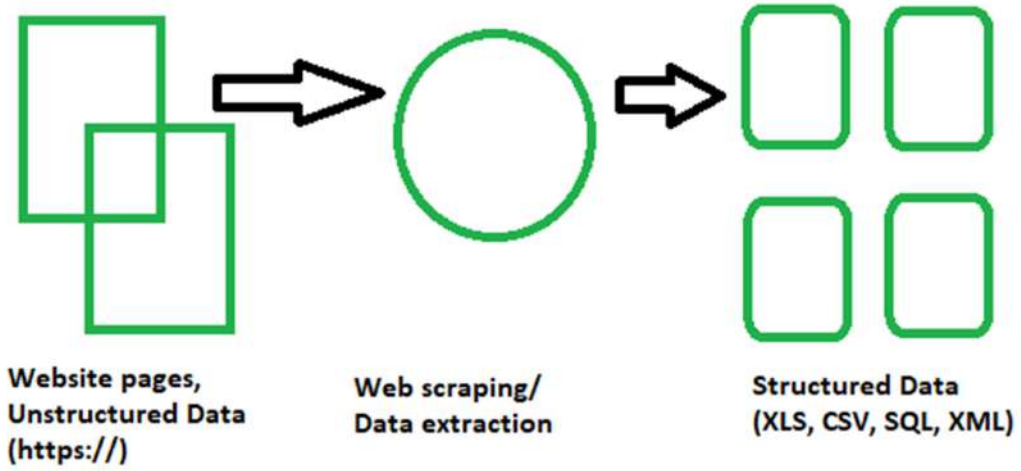


Figure.1 Framework of scraping process

Just as Scrapy was initially intended with the end goal of web scraping data from source, it can likewise be utilized to remove the data exploiting APIs or as a broadly useful web scraper. The fundamental points of interest of scrapy are that demands are booked and handled non concurrently, which implies that scrapy doesn't have to trust that a solicitation will be done and prepared, it can send another solicitation or do different things meanwhile, implying that different solicitations can prop up regardless of whether a few solicitations fizzle or a blunder occurs while doing the emphasis.

INPUT AND OUTPUT DESIGN

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1.Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3.When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

METHODOLOGY:

For the purpose of the project, the methodology that is utilized is to collect all of the data that has been extracted from a variety of sources by utilizing the vivid features of the web crawler scrapy and scripts written in the Python programming language. The data is then analyzed further in accordance with the requirements of the client, and the results are then stored in the database of the company. In addition, the web crawler scrapy, which is based on Python, could be able to assist us in retrieving the needed result so that we can complete the analytical process using a particular code and provide the desired url for the iteration to carry out in order to scrape the data from the source url. A coding system In this particular instance, the XPath approach was used to locate the specifics of each component of the Frequent Searches. The fundamental web crawling script that was utilized for the project displays the data that was crawled and placed in the database of the items that were obtained from a social networking website, namely Reddit. The code for the implementation of scrapy testing is shown in Figure 2. For the purpose of testing the project, I used the many components that were described before and made it so that it could be executed on the browser. The extraction that was carried out is determined to be entirely relevant, and the analysis that was carried out is estimated. The code for analyzing the data once it has been scraped is shown next. The outcomes When taken as a whole, the outcomes of the endeavor ultimately prove to be beneficial to comprehend. Web scrapy was used to extract the data, which was then converted into a csv file format. It turned out that the script that was built to extract the data was capable of identifying each of these sources given with a great deal of simplicity. Furthermore, the research that was carried out revealed the material that was searched the most on the website that was put through the test in the form of a percentage. In the shape of a pie chart, the result is shown in Figure 4. 7. Concluding Remarks Due to the independent and

diverse nature of hidden online content, the classic stress engine has become an inadequate method of searching for this sort of data. This has made the extraction of data from hidden web data a significant issue in today's world. Indexing, query processing, and an efficient data extraction approach based on web structure were the primary results of this project. Additionally, the user-friendly search interface was one of the important consequences. In the proceedings of the Third International Conference on Electronics Communication and Aerospace Technology (ICECA 2019), the following are presented: IEEE Conference Record # 45616; IEEE Xplore ISBN: 978-1-7281-0167-5 978-1-7281-0167-5/19/$31.00 ©2019 IEEE 453 form submission analysis and revised submission strategy. In order to completely accomplish automated integration, hidden web data need synthetic and semantic matching. In this thesis, a fully automatic and domain-dependent prototype system is provided that extracts and integrates the data that is hidden behind the search form.

TESTING AND VALIDATION

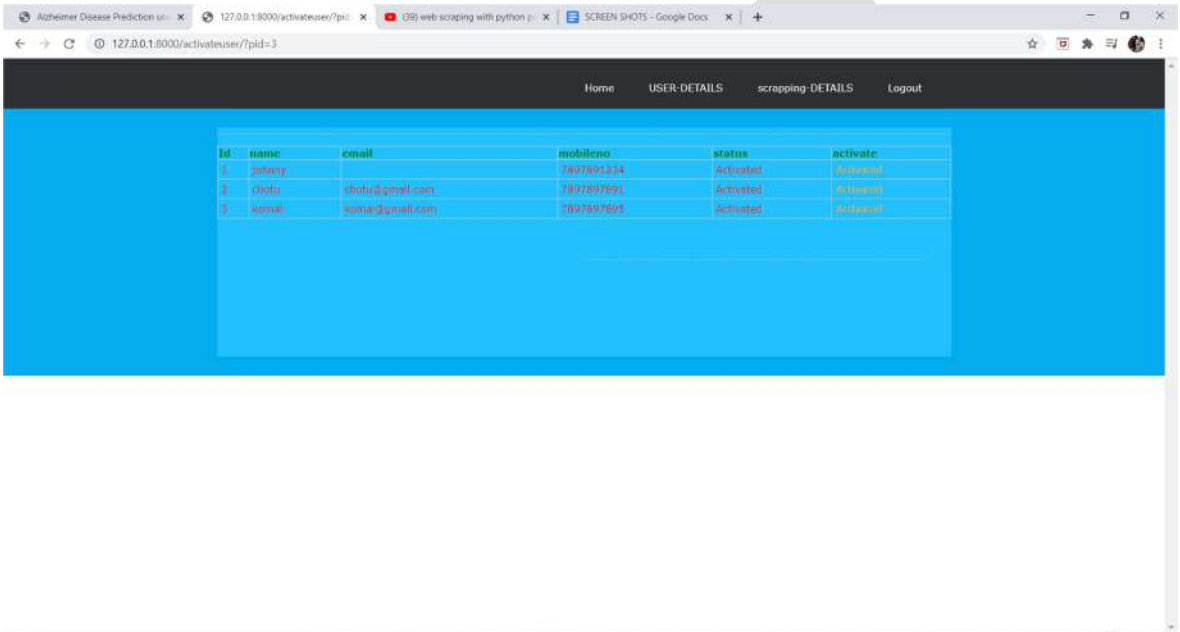
DESIGN OF TEST CASES AND SCENARIO

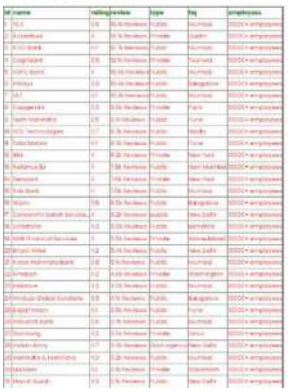
Sample Test Cases

S.no	Test Case	Excepted Result	Result	Remarks(IF Fails)
1.	User Register	If User registration successfully.	Pass	If an already user email exists then it fails.
2.	User Login	If the Username and password is correct then it will be a valid page.	Pass	Un Register Users will not logged in.
3.	Admin Add the Data	A new record will added to our dataset.	Pass	According to India metrological repository the data must be float fields otherwise its failed.
4.	admin display scraping data	all companies scraping data will display	Pass	if data is not found then it won't display.
5.	user search for scrapping company's data	list of company's data will display.	pass	if data not available in dataset data not found
6.	after clicking web scraping we will get job portal	based on job title and location we will get a job website.	Pass	no position available jobs or vacancies will not be available.

7.	Admin login	Admin can login with his login credential. If success he get his home page	Pass	Invalid login details will not be allowed here.
8.	Admin can activate the register users	Admin can activate the register user id	Pass	If user id not found then it won't login

RESULTS





CONCLUSION

Because of the independent and varied nature of the material that is concealed on the web, standard stress engines have become an inadequate method of searching for this sort of data. This has made the extraction of hidden online data a significant difficulty in the modern day. The most important results of this project were a search interface that was easy to use, indexing, query processing, and an efficient data extraction approach that was based on the structure of the website, as well as a new submission plan and a study of the form submission procedures. In order to completely accomplish automated integration, hidden web data need synthetic and semantic matching. In this thesis, a fully automatic and domain-dependent prototype system is provided that extracts and integrates the data that is hidden behind the search form.

REFERENCES

[1] ”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”

[2] ”Ghazvinian, Holbert, Viswanathan. “Simple WebScraping.”Internet:https://seanholbert.wor dpress.com/2011/07/15/scrappy-simple-webscraping/, Jun. 2015”

[3] ”Bellarosey.“Crowdsourcing-Definition.” Internet:http://crowdsourcing.typepad.com/cs/ 2006/06/crowdsourcing_a.html, Jun. 02, 2006”

[4] “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.”

[5] “Datahen."3 Advantages of web scraping for yourenterprise"Internet:https://www.datahen.c om/3-advantages-web-scraping enterprise/,May.17,2017””

[6] Ijteba Sultana, Dr. Mohd Abdul Bari ,Dr. Sanjay,” *Routing Performance Analysis of Infrastructure-less Wireless Networks with Intermediate Bottleneck Nodes*”, International Journal of Intelligent Systems and Applications in Engineering, ISSN no: 2147-6799 IJISAE,Vol 12 issue 3, 2024, Nov 2023

[7] Md. Zainlabuddin, "*Wearable sensor-based edge computing framework for cardiac arrhythmia detection and acute stroke prediction*", Journal of Sensor, Volume2023.

[8] Md. Zainlabuddin, "*Security Enhancement in Data Propagation for Wireless Network*", Journal of Sensor, ISSN: 2237-0722 Vol. 11 No. 4 (2021).

[9] Dr MD Zainlabuddin, "*CLUSTER BASED MOBILITY MANAGEMENT ALGORITHMS FOR WIRELESS MESH NETWORKS*", Journal of Research Administration, ISSN:1539-1590 | E-ISSN:2573-7104 , Vol. 5 No. 2, (2023)

[10] Vaishnavi Lakadaram, " Content Management of Website Using Full Stack Technologies”, Industrial Engineering Journal, ISSN: 0970-2555 Volume 15 Issue 11 October 2022

[11] Dr. Mohammed Abdul Bari,Arul Raj Natraj Rajgopal, Dr.P. Swetha ,” *Analysing AWSDevOps CI/CD Serverless Pipeline Lambda Function's Throughput in Relation to Other Solution*”, International

Journal of Intelligent Systems and Applications in Engineering , JISAE, ISSN:2147-6799, Nov 2023, 12(4s), 519–526

[12] Ijteba Sultana, Mohd Abdul Bari and Sanjay," *Impact of Intermediate per Nodes on the QoS Provision in Wireless Infrastructure less Networks*", Journal of Physics: Conference Series, Conf. Ser. 1998 012029 , CONSILIO Aug 2021

[13] M.A.Bari, Sunjay Kalkal, Shahanawaj Ahamad," *A Comparative Study and Performance Analysis of Routing Algorithms*", in 3rd International Conference ICCIDM, Springer - 978-981-10-3874-7_3 Dec (2016)

[14] Mohammed Rahmat Ali,: BIOMETRIC: AN e-AUTHENTICATION SYSTEM TRENDS AND FUTURE APLLICATION", International Journal of Scientific Research in Engineering (IJSRE), Volume1, Issue 7, July 2017

[15] Mohammed Rahmat Ali,: BYOD.... A systematic approach for analyzing and visualizing the type of data and information breaches with cyber security", NEUROQUANTOLOGY, Volume20, Issue 15, November 2022

[16] Mohammed Rahmat Ali, Computer Forensics -An Introduction of New Face to the Digital World, International Journal on Recent and Innovation Trends in Computing and Communication, ISSN: 2321-8169-453 – 456, Volume: 5 Issue: 7

[17] Mohammed Rahmat Ali, Digital Forensics and Artificial Intelligence ...A Study, International Journal of Innovative Science and Research Technology, ISSN:2456-2165, Volume: 5 Issue:12.

[18] Mohammed Rahmat Ali, Usage of Technology in Small and Medium Scale Business, International Journal of Advanced Research in Science & Technology (IJARST), ISSN:2581-9429, Volume: 7 Issue:1, July 2020.

[19] Mohammed Rahmat Ali, Internet of Things (IOT) Basics - An Introduction to the New Digital World, International Journal on Recent and Innovation Trends in Computing and Communication, ISSN: 2321-8169-32-36, Volume: 5 Issue: 10

[20] Mohammed Rahmat Ali, Internet of things (IOT) and information retrieval: an introduction, International Journal of Engineering and Innovative Technology (IJEIT), ISSN: 2277-3754, Volume: 7 Issue: 4, October 2017.

[21] Mohammed Rahmat Ali, How Internet of Things (IOT) Will Affect the Future - A Study, International Journal on Future Revolution in Computer Science & Communication Engineering, ISSN: 2454-424874 – 77, Volume: 3 Issue: 10, October 2017.

[22] Mohammed Rahmat Ali, ECO Friendly Advancements in computer Science Engineering and Technology, International Journal on Scientific Research in Engineering(IJSRE), Volume: 1 Issue: 1, January 2017

[23] Ijteba Sultana, Dr. Mohd Abdul Bari ,Dr. Sanjay, "*Routing Quality of Service for Multipath Manets, International Journal of Intelligent Systems and Applications in Engineering"*, JISAE, ISSN:2147-6799, 2024, 12(5s), 08–16;

[24] Mr. Pathan Ahmed Khan, Dr. M.A Bari,: Impact Of Emergence With Robotics At Educational Institution And Emerging Challenges", International Journal of Multidisciplinary Engineering in Current Research(IJMEC), ISSN: 2456-4265, Volume 6, Issue 12, December 2021,Page 43-46

[25] Shahanawaj Ahamad, Mohammed Abdul Bari, Big Data Processing Model for Smart City Design: A Systematic Review ", VOL 2021: ISSUE 08 IS SN : 0011-9342 ;Design Engineering (Toronto) Elsevier SCI Oct : 021

[26] Syed Shehriyar Ali, Mohammed Sarfaraz Shaikh, Syed Safi Uddin, Dr. Mohammed Abdul Bari, "Saas Product Comparison and Reviews Using Nlp", Journal of Engineering Science (JES), ISSN NO:0377-9254, Vol 13, Issue 05, MAY/2022

[27] Mohammed Abdul Bari, Shahanawaj Ahamad, Mohammed Rahmat Ali," Smartphone Security and Protection Practices", International Journal of Engineering and Applied Computer Science (IJEACS) ; ISBN: 9798799755577 Volume: 03, Issue: 01, December 2021 (International Journal,U K) Pages 1-6

[28] .A.Bari& Shahanawaj Ahamad, "Managing Knowledge in Development of Agile Software", in International Journal of Advanced Computer Science & Applications (IJACSA), ISSN: 2156-5570, Vol: 2, No: 4, pp: 72-76, New York, U.S.A., April 2011

[29] Imreena Ali (Ph.D), Naila Fathima, Prof. P.V.Sudha ,"Deep Learning for Large-Scale Traffic-Sign Detection and Recognition", Journal of Chemical Health Risks, ISSN:2251-6727/ JCHR (2023) 13(3), 1238-1253

[30] Imreena, Mohammed Ahmed Hussain, Mohammed Waseem Akram" An Automatic Advisor for Refactoring Software Clones Based on Machine Learning", Mathematical Statistician and Engineering ApplicationsVol. 72 No. 1 (2023)

[31] Mrs Imreena Ali Rubeena,Qudsiya Fatima Fatimunisa "Pay as You Decrypt Using FEPOD Scheme and Blockchain", Mathematical Statistician and Engineering Applications: https://doi.org/10.17762/msea.v72i1.2369 Vol. 72 No. 1 (2023)

[32] Imreena Ali , Vishnuvardhan, B.Sudhakar," Proficient Caching Intended For Virtual Machines In Cloud Computing", International Journal Of Reviews On Recent Electronics And Computer Science , ISSN 2321-5461,IJRRECS/October 2013/Volume-1/Issue-6/1481-1486

[33] Heena Yasmin, A Systematic Approach for Authentic and Integrity of Dissemination Data in Networks by Using Secure DiDrip, INTERNATIONAL JOURNAL OF PROFESSIONAL ENGINEERING STUDIES, Volume VI /Issue 5 / SEP 2016

[34] Heena Yasmin, Cyber-Attack Detection in a Network, Mathematical Statistician and Engineering Applications, ISSN:2094-0343, Vol.72 No.1(2023)

[35] Heena Yasmin, Emerging Continuous Integration Continuous Delivery (CI/CD) For Small Teams, Mathematical Statistician and Engineering Applications, ISSN:2094-0343, Vol.72 No.1(2023)

[36]