Analyzing Data through Web Scraping

Dharamvir[1],Dr. MS. Shashidhara[2]
Tarun DuttSharma[3],Tej Singh [4],Shrijaykesarwani[5] ,Hariom Mishra[6]
[1]Asst.Professor,[2]Professor and Head
[3,4,5,6]Final Year MCA
[1,2,3,4,5,6] Department of MCA, The Oxford College of Engineering , Bengaluru , Karnataka ,
INDIA-560068
Corresponding Author:[1]dhiruniit@gmail.com

Abstract - The Basic Data Mining Techniques are based on the standard relationships, shaped on an example over insignificant experiments and quantitative practicals, therational approach toward exploration.The standard goal is to find a solution to business, scientific or social problems. Data extraction or scraping is a method that is to extract data from a web by itself, which can be achieved very easily and in an instant. The Scraper's overlook and techniques are just posed in more advert ways to visualize the Analyzed Data, it explains about how the web scraping isplanned. The proper method of which is presented in simple steps: the scraper fetches the required links from the web, and then the information is retrieved to get it from the source links and conclusively stores and manipulates that dataset using a CSV data file created from a dataset. Python programming is used for the carrying out this manipulation of data.ForWhich we have to use various Python modules to create or build our new datasets and perform visualizationsusing matplotlib, Pandas, NUMPY, Beautiful Soup, Regular expression module, and also python urlib.By doing so, We will be using a basic request from python urlib with the help of an SSL error ignoring fetching the page and then passing it to the amazing Beautiful Soup, which is then going to parse HTML and output a pretty text dump.

Keywords – Beautiful Soap, Data Mining,Web Scrapping.

1. Introduction

Data Mining and Visualization is the method of extracting knowledge and data to achieve the given objectives and goals viamanipulationof data. The analysis of data consists of finding and identifying problems and anomalies, resolving the problemsand providing accessibility of relevant data, and identifying what methods can help in discovering the best solution to the given real-world problems and delivering their respective outputs. For the specific goal for Analysis,we have to segregate data into different format divisions further such as to start with specifications organizing, re-assembling, cleaning, re-analysis Implementing machine learningmodels and high-end algorithms, and finally reassessing the final resulted outcome.

Website Data scraping andsupporting are amazing techniques for naturally and carefully creating analyzed articles and content on the internet. Many individuals already utilize these techniques or strategies in doing research and businesses profiting from offering reviews to increase their perfection for advertising that allows league individuals to perfectly deliver resources that help them in growing and developing their profits concerning their organizations.

As we know, web scraping is often identified as,
"Web Data Extraction". By The use of python programming and with the help of two python libraries named Requests and Beautiful Soup. The tool used for web Scraping is used for deriving accurate and valuable information for analysis of data, and as a part of that tool is used for web data Mining, online goal or motivation of observing changes and value correlations between the element survey of scratching. We useBeautiful Soup. By using this we will be able to store the HTML page as a string inside the HTML variable. For a document to parse, firstly it has to go through Beautiful soup Constructor now after that we get a new object "soup" which stores documents or represents them as a data structure that is nested. In Beautiful soup, HTML is accessed as a tree structure with methods used to parse HTML. By merging all these with the gained information in form of the method of using libraries and working Techniques, we can propose a relevant Scraper to output a better and more understandable output. Due to the presence huge community and libraries for Python and the impeccable coding of python language, for Scraping or extraction of required data from the web

2. Objective

The proper goal or objective of this research paper is to understand the extraction of the information and data from different resources on the internet with the usage of the web crawlingtool Scrappy and Beautiful soapby usingthe language Python programming. Theinstance of a new database is created of which major part is still a collection of different the unstructured articles from a variety of resources and then analyzing them by the passing through the principals of data analysis specifications, an example is organizing and cleaning, reassembling, and then re-analyzing, application of machine learning models and learning algos and providing outputs with the desired outcomes. Web data extraction can be done with or without the help of tools .such as with the help of Scrapy which is the tool used for extraction of data whenever a complex method is not needed by the scrappers .it is aneasily available web-crawling API developed for the extraction of any data according to the user's requirements. This tool is utilized in the extraction of data by implementing an API. which is requested by Clients. There is also a data scraping method using Beautiful soap which is the most interesting library present in python. We can use this python module in various ways if you are familiar with python programming you can easily work on scrapping data with this module .let's take the example of products that sometimes cannot be seen on the application interface. Data such as variations, ratings, reviews, or any of the data has a wide range of availability.

3. Literature Review

To properly understand how data extraction or scarpinghas evolved, one has to understand the principles and techniques that are part of this method of web scraping or can be said methods that are important for performing web data extraction are known to the world since the existence of wb.To have access to and gather the information on the internet there is already many practical application. These are some confounding real-world applications: the Global I.T giant Google uses text or information stored on the internet to improve or train its Google to translate. The major profiting primary Organizations seemed to rise on the impact point of booming web-based businesses and gained hand until and up to the 2000s.

With advances in storage components and innovations in Real Databases, validation and computation are back again. The data was displayed and processed as data set up for data validation. [13] An important turning point was the advent of RDB (relational database) in the 1980s. This allows customers to create Sequel (SQL) to recover data from the database. For customers, the advantage of RDB and SQL is that they can segregate their data by conspiracy. This created a methodology for basically retrieving data and spreading the use of databases. Information Warehouse: Unlike regular social databases, information warehouses are usually optimized for query response time. Improved data mining allows us to assess the progress of databases and datasets. This requires the association to store more data and sensibly separate them. [4] In response to the recorded procurement design review, a general commercial pattern was developed in which the government began to "predict" the potential needs of its customers

4.Feasibility and Application

The main goal of the existence of data and information is for the extraction and then analysis which is part

Preparing appropriate articles and surveys based on those articles The need for the extraction of data is to Identify the intentions and summary for the information extracted from the web. it also provides different configurations to announce in the distinctive styles. there is also a need of featuring the primary information parts of the intrigue to maintain the pattern matching and recognition for data analysis. and as for the importance of data analysis, it is essential for alerts and warning towards malicious data resources, it helps in various ways such as

Creating surveys, planning, and blueprints, graph designing, map building, and creating different models based on a machine is quite impossible without a data source

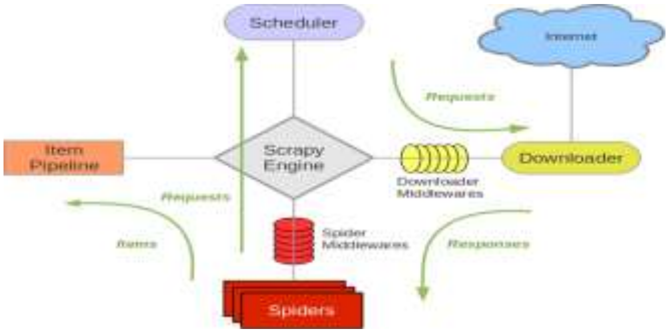

Figure 1. Architecture of scrapy

Scrapy is Api developed for surfing the sites and extracting information needed for a wide variety of purposes such as recording the and storing factual data the architecture of scrapy is given above As scrapy was the tool which was intended only for data extraction but now this is also used to remove the API's that are exploiting the data. the most important feature of scrapy is that the demands are handled non- concurrently.

5. IMPLEMENTATION

The fundamental of implementing the web scraping is given usually in two most popular ways that both can be done using python which is easy to use and has an abundance of libraries to utilize those two methods of implementation are

Use of scrapy tool

This is a tool that is the complete package for fetching data from the web and processing them and storing it in the databases. it usually is easy to use and can be installed in python using "pip" and then using scrapy shell commands to fetch the HTML data and various other commands you can also use conda to install scrapy using "**conda install –c conda –forge scrapy**" you can execute commands through scrapy shell to fetch data from web using " fetch" command with the URL can give you desired data the shell returns a response the user can check response using "view " command you can view HTML script using "print " command in the shell using scrapy software developers can check their theories. scrapy can also provide a Pipeline for items that can aid you to write your methods into a spider which can be used in various such as validation.

The image part with relationship ID rId11 was not found in the file.

A seconduse is the implementation ofBeautiful soup

which is also a quite famous library of python and widely used using this library requires the use use of another important library of python which is "REQUESTS" this library is usually used to parse HTML and create an object for accessing HTML content this library usually treats HTML like a structured tree were accessing a tag can be done easily using various python techniques there are different platforms where you can perform data extraction using beautiful soup depending on the different requirements needed to run python program.

Table 1.1 **use of Beautiful soap example of Stock prices**

Symbol	**Name**	**High**	**Low**	**Close**	**Volume**	**URL**
AAB	Aberdeen International Inc	0.135	0.130	0.130	146215	http://eoddata.com//stockquote/TSX/AAB.htm
AAV	Advantage Oil & Gas Ltd	6.630	6.060	6.610	2106233	http://eoddata.com//stockquote/TSX/AAV.htm
ABCT	ABC Technologies Holdings Inc	5.790	5.550	5.790	4029	http://eoddata.com//stockquote/TSX/ABCT.htm
ABCT.RT	ABC Technologies Holdings Inc Rights	0.010	0.005	0.005	58102	http://eoddata.com//stockquote/TSX/ABCT.RT.htm
ABOUT	Absolute Software Corp	11.650	11.100	11.130	186919	http://eoddata.com//stockquote/TSX/ABST.htm
AX.PR.E	Artis REIT Pref Ser E	24.350	24.250	24.300	3300	http://eoddata.com//stockquote/TSX/AX.PR.E.htm
AX.PR.I	Artis REIT Pref Series I	25.660	25.400	25.660	1269	http://eoddata.com//stockquote/TSX/AX.PR.I.htm
AX.UN	Artis Real Estate Investment Trust Units	13.130	12.930	13.020	308738	http://eoddata.com//stockquote/TSX/AX.UN.htm
AUX	Alexco Resource Corp	1.950	1.820	1.930	162738	http://eoddata.com//stockquote/TSX/AXU.htm
AYA	Aya Gold and Silver Inc	10.550	9.860	10.360	543086	http://eoddata.com//stockquote/TSX/AYA.htm

6. Conclusion

According to the points above discussed the extraction of web information comes with a few difficulties which include the extraction of heterogeneous data and autonomous data that are hidden or scattered around the web which is the reason why old methods of data extraction have become ineffective. The goal which this research set to achieve is to provide an interface that is user-friendly and interactive towards the scattered and hidden data on the web the interface should be using the techniques that are adapted towards the new web structures .in this research or thesis full automated system is proposed that ease the extraction of the data from complex structures of the web.

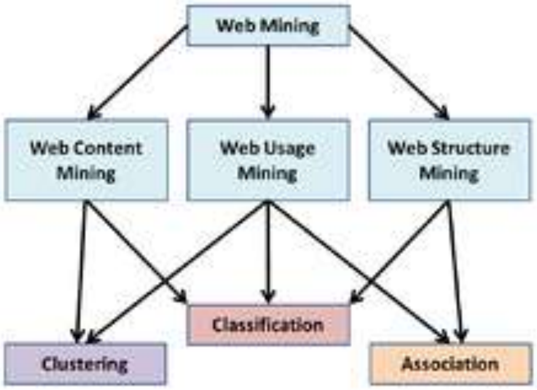

Figure 1.3Architecture of web Mining

7. Future scope

The problems that can occur shortly may be due to the complex and unpredictable structure of web information that is stored and scattered. As there are no certain principles or norms to follow and the web is a dynamic space that changes more frequently. And due to the lack of this consistency, there may be difficulties in getting an organized information process this issue can get more intense as the increase in size or scale of data. even then there are still openings and solutions to reach the particular confinements of data.

8. References

1. S. Aphiwong siphon and P. Chongstitvatana, "Methods of Web scraping ".JETIR Vol1, issue 4 July 2018.

2. 2. Kk. Agarwal, "Beautiful soup a Scraping Technology," IJRTE, vol. 07, no. 06, pp. 824-817, May 2018.

3. Rs. Agarwal, "Analysis of Detection of THROUGH Web scraping," ICRTAC, pp. 312-393, 2017.

4. 4. F. Islam, "Bengali Fake News Detection," ICRRC – 2K20, IEEE, pp. 281-287, 01 October 2020. 5. S. D. Samantaray, "," IJSR, vol. 08, no. 01, pp. 1126-1132, January 2019. 6. I. Vogel, "web scraping through Tools," IEEE, pp. 599-6512, 21 July 2019