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

Data Analysis by Web Scraping using Python

David Mathew Thomas, Sandeep Mathur
Amity Institute of Information Technology
Amity University (AUUP), Sec-125, Noida
davidmathew97@gmail.com, sandeep2809@gmail.com, smathur@amity.edu

Abstract - The standard information investigation are built on the root and impact relationship, shaped an example minuscule examination, subjective and quantitative examination, the rationality approach of creating extrapolation examination. The Web Scraper's conniving ethics and procedures are juxtaposed, it explains about the working of how the scraper is premeditated. The technique of it is allocated into three fragments: the web scraper draws the desired links from web, and then the data is extracted to get the data from the source links and finally stowing that data into a csv file. The Python language is implemented for the carrying out. By doing so, linking all these with the moral knowledge of libraries and working know-how, we can have an adequate Scraper in our hand to produce the desired result. Due to an enormous community and library resources for Python and the exquisiteness of coding chic of python language, it is most appropriate one for Scraping desired data from the desired website.

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

I. Introduction

Data analysis is the method of extracting solutions to the problems via interrogation and interpretation of data. The analysis process comprises of discovering problems, resolve the accessibility of suitable data, determining which method can help in finding the solution to the interesting problem and convey the result. For the purpose of analysis, the data has to segregate into various steps further on such as starting with its specification assembling, organizing, cleaning, re-analyzing, applying models and algorithms and the final result.

Web information scraping [1] and publicly supporting are outstanding strategies for naturally creating substance on web. A considerable amount of individuals utilized these strategies in research and business for creating substance or offering criticisms to expand the exactness of business advertising that enables individuals to deliver resources in advancing and developing the business [3].

By and large, web scraping is notable for a "Screen Scraping", "Web Data Extraction". The web scrubber programming is planned for exhaustive for all noteworthy data from different online stores and mining, and collecting it into the new website. The scraper tool for the web is utilized for derived information from the web host, and as a portion of uses used for 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. [2]Pages are manufactured utilizing content-based increase dialects (HTML and XHTML), and much of the time contain a profusion of cooperative info in the content structure. Be that it may be as most website pages are anticipated for human end-users and not for minimalism of robotized use. Thus the toolbox that scrapes web info was

Authorized licensed use limited to: Instituto Federal do Piaui (IFPI). Downloaded on May 01,2023 at 18:36:25 UTC from IEEE Xplore. Restrictions apply.

made. [7]A web scraper is an API to mine data from a site. Associations like Amazon AWS and Google give web scraping tools, organizations, and open data available free of cost to end customers. As for the paper will be focused on the data analysis using python's effectiveness as a programming language, it's out to an apt choice as a single language for the data-centric application, For this, the version of Python used will be Python 3.6 for the analysis.

II. Objective

The point of the paper is to remove the information from different sources with the assistance of programming known as the web crawler Scrapy utilizing the programming language Python adaptation 3.6.The Database is created which collects all the unstructured data from various sources and then analyzes them by the analytic process of its specifications, assembling, organizing, cleaning, re-analyzing, applying models and algorithms and finally providing the desired results. [5]Web scraping software's such as Scrapy is available for whenever ease of access needed by the user and also it's an open-source web-crawling framework for the collection of any data as per user's needs. The software is used to extract data using an application programming interface or as a general-purpose web crawler required by the desired customer. We are also able to scrape the data of E-commerce sites such as Flipkart, Amazon, etc. so as to find out the product details which aren't available with the application interface and to analyze the variation, comments, ratings or anything else with innumerable options.

III. Literature Review

To know how the data extraction process has evolved has so much one must understand the techniques involved in this method of web scraping is important scraping has been around nearly as long as the web. [6]The impact behind business web scraping has dependably been to pick up a simple business advantage and incorporate things like undermining a contender's special valuing, taking leads, commandeering promoting efforts, diverting APIs, and the inside and out robbery of and information.

The primary aggregators and examination motors seemed hot on the impact points of the web based business blast and worked generally unchallenged until the legitimate difficulties of the mid-2000s. Early scraping apparatuses were really fundamental - physically reordering anything unmistakable from the site. When software engineers got included, scraping graduated to the Unix grep order or customary articulation coordinating procedures posting remote HTTP demands utilizing attachment programming, and parsing site utilizing information programming and parsing site utilizing information inquiry dialects. Today, in any case, it's an altogether different story: web scraping is huge business with powerful devices and administrations to coordinate.

Extraction and Analysis of information are generally utilized by the Digital distributers and catalogs, Travel, Real home, and E-trade. Then again, examination and figuring come path back with the advances in accumulation components and the innovation of Real Databases: The data had been seen and dealt with as data to be set up for data examination. [13]The pivotal turning point was the nearness of RDB (Relational Database) amid the 1980s which empowered customers to create Sequel (SQL) to recoup data from the database. For customers, the advantage of RDB and SQL is to have the ability to separate their data on intrigue. It made the methodology to get data basic and spread database use. Information Warehouse: The distinction from regular social databases is that information stockrooms are generally streamlined for reaction time to inquiries. The improvement of data mining as made possible appreciation to database and data stockroom progressions, which engage associations to store more data and still separate it in a reasonable manner. [4]A general commercial pattern developed, where administrations began to "foresee" client's potential needs

Authorized licensed use limited to: Instituto Federal do Piaui (IFPI). Downloaded on May 01,2023 at 18:36:25 UTC from IEEE Xplore. Restrictions apply.

dependent on examination of the chronicled obtaining designs.

IV. Feasibility and Application

As we know to find out the logic behind the purpose of data, data extraction and analysis is a must. The need for extraction is required in order to consistently promotion authentically acquire the information according to before the interpretative stage, not as to foresee the significance of information as a substitute for extraction. We need extraction with respect to the articles are in various configurations and utilize distinctive styles of announcing. The need to feature the principle information components of intrigue and to give institutionalization. Also to aid pattern recognition and analysis. As for data analysis is essential for awareness of data resources by focusing on the applicable issues.

It throws a light on by providing with the surveys, planning for statistical graphs designing and redesigning etc.

Scrapy

Though the web crawler, scrapy is an application system for gliding the sites and removing organized information which can be utilized for a wide scope of priority applications, similar to fact quarrying, data concocting or recording the data. The framework of scrapy depicted below for better understanding.

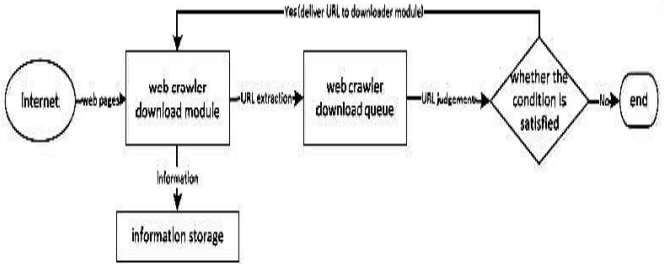

Figure.1 Framework of scraping process

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.

V. Implementations

Python 3.6 is scheduled to be the last major version in the 2x series before it moves into an extended maintenance period. It contains a large number of the highlights that were discharged in python 3.6. This ad libbed form incorporates the accompanying highlights, for example, an arranged word reference type, new unit highlights including test skipping, new state strategies, and attest techniques, and test disclosure, a match quicker to the module, programmed numbering of fields in the str. group() technique. Buoy up upgrades back ported from 3.x, tile support for T kinter. A dark port of the memory see object from 3.x, set literals set and word reference understandings, lexicon sees, new sentence structure for settled with proclamations and the sys config module.

Use of Scrapy

Scrapy is an application framework for crawling locales and expelling composed data which can be used for a wide extent of supportive applications, like data mining, information taking care of or real reported. [12]Despite the way that Scrapy was at first expected for web scraping, it can in like manner be used to evacuate data using APIs, (for instance, Amazon AWS) or as an all-around valuable web crawler. Sketchy is written in Python. Let's take an example on Wiki related to one such issue "A simple online photo gallery may offer three options to users, as specified through HTTP GET parameters in the URL. If there exist four ways to sort images, three choices of thumbnail size, two file formats, and an option to disable user provided contents, then the same set of content can be accessed with 48 different URLs, all of which may be linked on the site. This carefully worked-out combination creates a

Authorized licensed use limited to: Instituto Federal do Piaui (IFPI). Downloaded on May 01,2023 at 18:36:25 UTC from IEEE Xplore. Restrictions apply.

problem for crawlers, as they must sort through endless combinations of relatively minor scripted changes in order to recover unique content."

VI. Methodology

The methodology used for the project is to gather all the data extracted from various sources by using the vivid features of the web crawler scrapy using the scripts written in python language and further analyze it as per the requirements of the customer where the data is stored in the company's database. [9]The web crawler scrapy which is python based also may help us retrieve the desired result as we analysis process by specific code and give the desired url for the iteration to perform for scrapping the data from the source url.

Coding

The basic web crawling script used for the project which shows the data crawled and stored in the database of the products from a social network site, in this case, Reddit by the XPath method involved to find the details of each element of the Frequent Searches.

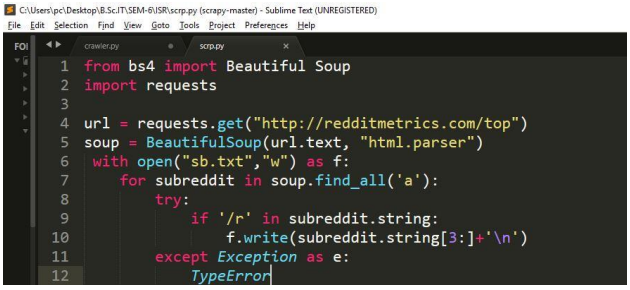

Figure.2 Code for Implementation of Scrapy

Testing

The project was tested by me using the various components as defined earlier and made to run on the browser. The extraction done turns out to be completely relevant and the analysis made is estimated.

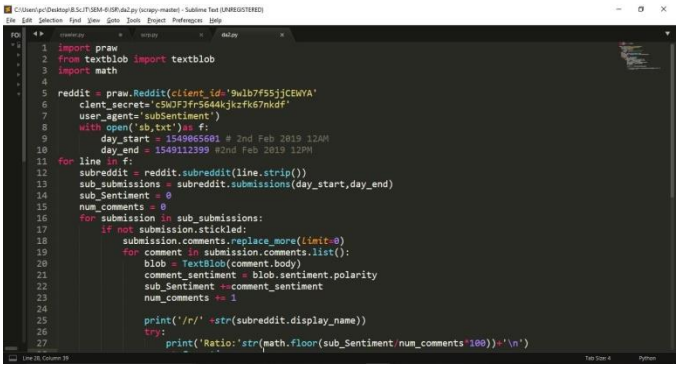

Figure.3 Code for analyzing the data after scraping

Results

The overall results of the project turn out to be helpful to understand. The Web scrapy extracted the data and made into csv file format. The script which was written to extract the data turned out to be both of finding each of these sources provided with great ease. Moreover, the analysis done has shown the most searched content in the site taken for test in the percentage format.

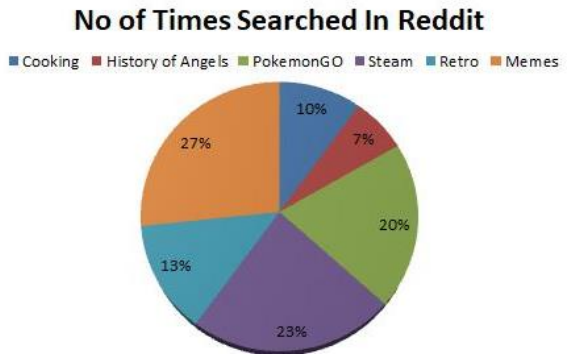

Figure.4 The result in the form of pie chart

VII. Conclusion

The extraction of data hidden web data is a major challenge nowadays because of autonomous and heterogeneous nature of hidden web content traditional stress engine has now become an ineffective way to search this kind of data. The main outcomes of this project were user friendly search interface, indexing, query processing, and effective data extraction technique based on web structure,

Authorized licensed use limited to: Instituto Federal do Piaui (IFPI). Downloaded on May 01,2023 at 18:36:25 UTC from IEEE Xplore. Restrictions apply.

form submission analysis and new submission plan. Hidden web data need synthetic and semantic matching to fully achieve automatic integration in this thesis fully automatic and domain dependent prototype system is proposed that extract and integrate the data lying behind the search form.

VIII. Future Scope

The difficulties that came ahead are the non-uniform structure that is the web is a dynamic space with irregularity in information organizations and structure. There are no standards to be pursued while building web nearness. Because of this Lack of consistency, gathering information in a machine-meaningful arrangement can be troublesome principle issue gets intensified with the scale when you need organized information process a k an information extraction this spot test when a ton of subtleties are to be required to infiltrate to a particular plan from a large number of their sources this can be overwhelmed by the besides improvement in the help and condition arrangement of the Components utilized. Indeed, even with every one of the confinement's web information still present use openings in the event that we realize how to put it at the correct use.

IX. 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.wordpress.com/2011/07/15/scrappy-simple-web-scraping/, 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.com/3-advantages-web-scraping-enterprise/,May.17,2017""
6. "https://en.wikipedia.org/wiki/Web_scraping"
7."https://www.webharvy.com/articles/what-is-web-scraping.html"
8."http://resources.distilnetworks.com/h/i/53822104-is-webscraping- illegal-depends-on-what-the-meaning-of-theword-is-is/181642"
9."https://www.quora.com/What-is-the-legality-of-web-scraping"
10. https://en.wikipedia.org/wiki/Web_crawler
11."Kolari, Pand Joshi A. ,"Web mining : research and practice , Computing in Science & Engineering", IEEE Transactions on Knowledge and Data Engineering, vol. 6, no. 2,Vol. 6 , No. 4 , 2004"
12."Pythonversion3.6,http://www.python.org."
13."Kengtel,W:Wagner,M.Proteins1999,37,334-345."
14. "BrightPLanet.com Deep web White Paper. http://www.completeplanet.com/Tutorials/DeepWeb/index.asp."

Authorized licensed use limited to: Instituto Federal do Piaui (IFPI). Downloaded on May 01,2023 at 18:36:25 UTC from IEEE Xplore. Restrictions apply.