Analyzing Data by Web Scraping using Python

J. Jyotsna[1], Adula Bhavani[2], Srikanth Dharavath[3], Gampa Shiva[4]
Department of Information Technology, J.B Institute of Engineering and Technology, Hyderabad

Abstract-Web scraping refers to the process of extracting structured data from HTML content. To achieve this, a web scraper is utilized to extract specific information from a desired website. This information can be obtained by using the inspect option to select the relevant class of data, which is then scraped. There are two main libraries used for web scraping: the "Request" library, which updates the HTML data, and the "Beautiful Soup" library, which scrapes data from the desired website. The data obtained can be stored in a CSV file or represented on a website or charts using the Python library, Matplotlib. In addition to acquiring data, web scraping can also be used to archive and track changes to online data that may be poorly structured or not presented in tabular form.

Keywords: Data analysis, Web Scraping, Beautiful soup.

1. INTRODUCTION

Data analysis involves extracting solutions to problems by interrogating and interpreting data. The process of analysis includes identifying problems, assessing the availability of appropriate data, and determining which methods can be used to analyze an interesting problem and communicate the results, data must be segmented into several steps, beginning with its specification, followed by assembling, organizing, cleaning, re-analyzing, and applying models and algorithms. In the end, the final outcome is achieved. Utilizing web information scraping can be a beneficial tool in this process. Scraping web information and public participation are excellent strategies for generating organic content on the web. Many individuals have utilized these methods in research and business to create content or provide feedback, which helps to increase the accuracy of business marketing and allows people to produce resources to promote and grow the business.

Web scraping is commonly known as "Screen Scraping" or "Web Data Extraction." This type of software is designed to extract all important information from various online stores and websites, compiling it into a new website. The web scraper tool is used to extract data from web hosts and is a part of applications used for web orders, data mining, web mining, online price monitoring, price comparison, competitor analysis, real estate listing collection, weather data monitoring, website change detection, review and reputation monitoring, web mashups, and web data integration. Web pages are constructed using markup languages such as HTML and XHTML and often contain a wealth of information in textual form. However, most websites are designed for human end-users and are not optimized for automated use. Thus, the toolbox that scrapes web info was made.

A web scraper can be thought of as an API used to extract data from a website. Organizations such as Amazon AWS and Google provide web scraping services, tools, and open data without any charge to users. As the paper is concentrated on data analysis using Python programming language, it is a fitting option for data-oriented applications. Python's aptitude for data analysis makes it a remarkable choice. For the analysis's objectives, Python version 3.6 will be utilized.

2. OBJECTIVE

The objective of the paper is to extract information from diverse sources using a programming technique called web scraping, utilizing the Python programming language version 3.6. A database is created to gather unorganized data from multiple sources, which is then subjected to an analytical process for further examination of its specifications, assembling, organizing, cleaning, re-analyzing, applying models and algorithms and finally providing the desired results. [5] Web scraping tool, like Beautiful Soup, is a convenient tool that is readily available for users who require easy access to data. It is an open-source web scraper user's specific need. This software is capable of extracting data through an application programming interface, or it can be used

as a general-purpose web scraper based on the user's requirements. With this tool, it's possible to extract data from e-commerce websites such as Flipkart and Google to obtain product details that are not available through the application interface. Additionally, it allows for the analysis of various aspects, including variations, comments, ratings, and other options. The main objective of this project is first, user must choose one or more websites and then they must copy the URL of that respective website, that URL will be copied into our code. By that URL we can extract the classes which are there in the X-path, choose one or more classes and paste the class into the code this is how we will extract the data in particular. Beautiful soup is the scraper which will help in extraction of the data from the website.

3. 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] Businesses have always used web scraping as a means to gain a competitive edge by resorting to activities such as undercutting a rival's pricing strategy, acquiring leads, hijacking marketing campaigns, misusing APIs, and stealing valuable data. During the early days of web scraping, the available tools were limited to manual methods such as copying and pasting visible content from websites or using Unix grep commands and regular expression matching techniques to extract data. Some developers also employed data programming and data query languages to list remote HTTP requests and parse websites. However, today web scraping has evolved into a big business with powerful tools and services available. Data extraction and analysis are widely used by digital publishers and directories, the travel industry, real estate, and e-commerce. On the other hand, data analysis has a long history, dating back to the development of relational databases (RDB) in the 1980s, which enabled clients to use SQL to retrieve data from databases. RDB and SQL made data extraction easy and spread the use of databases. Data warehouses are optimized for query response time and enable organizations to store more data and still extract it in a meaningful way, thanks to advancements in database and data warehouse technologies that made data mining possible. This led to a general commercial trend, where services started to anticipate customers' potential needs for purchasing patterns.

4. FEASIBILITY AND APPLICATION

Web scraping has become an essential tool in many fields such as e-commerce, marketing, and research, involving the extraction of data from websites. Some of its feasibility and applications include gathering data on competitors' prices, products, and marketing strategies for market research purposes. Additionally, it can be used to generate leads by collecting potential customer data, such as contact information and company details, to create targeted marketing campaigns and increase sales. Web scraping can also be utilized for sentiment analysis, allowing businesses to analyze customer feedback to gain insights into customer sentiment and improve their products and services. Moreover, web scraping enables businesses to monitor prices for products and services to adjust pricing strategies to remain competitive, and it can aggregate content from various sources, such as news websites, blogs, and social media platforms. Another use of web scraping is gathering data on competitor's website rankings, keyword usage, and backlink strategies for search engine optimization purposes. Although web scraping provides valuable insights, it is crucial to conduct it ethically and legally, while respecting website owners' terms of service and privacy policies.

5. IMPLEMENTATIONS

Python 3.6 was a major release of the Python programming language, which was launched in December 2016. It introduced several new features and improvements, such as formatted string literals, asynchronous generators, type hints, improved syntax for variable annotations, and additional built-in modules. These features have since become widely used in the Python community and have made Python 3.6 a popular choice for developers looking for a modern, powerful, and easy-to-use programming language. Beautiful Soup Scraper: Beautiful Soup is a Python library that is widely utilized for web scraping and data extraction purposes. Its collection of tools allows users to parse HTML and XML documents, navigate their structures, and extract relevant information from them. Essentially, Beautiful Soup creates a treelike structure, known as a parse tree, from the document, which can be easily searched and filtered using various functions and methods. This makes it easier to extract specific data from web pages, such as contact details, product information, or headlines. Additionally, Beautiful Soup offers support for popular parsers, such as html5lib and xml, which can handle malformed or incomplete HTML and XML documents. As a result, Beautiful Soup provides a powerful and flexible solution that simplifies the web scraping process, enabling users to extract valuable insights from online data sources.

6. METHODOLOGY

The approach employed in the project involves utilizing the capabilities of a web scraper to extract data from multiple sources, and then using python scripts to process the information. The gathered data is subsequently analyzed according to the specific needs of the customer, and stored in the company's database.[9] The beautiful soup scraper will extract the data from the X-Path of the user desired website and they will be extracted and shown in the csv file or the User interface created by user. Testing process: Our group personally tested the project by utilizing the different components that were defined earlier, and then proceeded to run it on a web browser. The extraction process yielded completely relevant data, and after conducting an analysis, the results were estimated to be accurate.

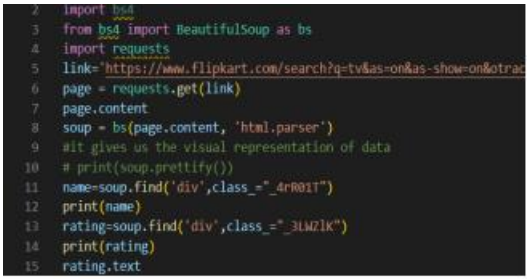

Figure.2 Code snippet for implementing scraper

Result screens: The results are presented in two distinct formats. Firstly, the user interface, which is created using Django, displays weather data scraped from Google in real-time, along with job search results from Indeed.com and global news results from the NewsAPI module. This information is continuously updated using the "request" module in Python. Secondly, the results are also presented in the form of a line graph, which shows multiple sets of scraped data from various websites, such as ESPN.in and Flipkart.com. The Flipkart data scraping results are depicted in Figure4.

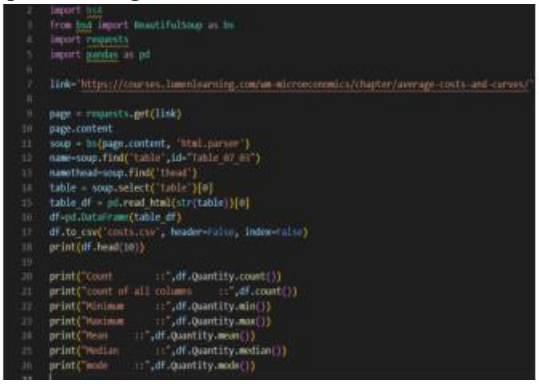

Figure.3 Code for data analysis after scraping

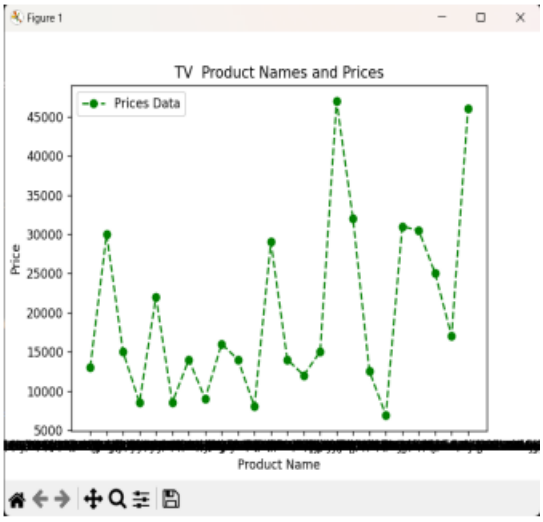

Figure.4 Line graph result of flip kart product

7. CONCLUSION

Presently, retrieving data from the hidden web poses a significant challenge due to its self-governing and diverse content. Traditional search engines are no longer effective in searching for this type of information. The primary results of this undertaking involved developing a search interface that is user friendly, an indexing system, a query processing system, and a data extraction technique that is efficient and relies on the structure of the web.

8. FUTURE SCOPE

- Improved natural language processing (NLP): NLP algorithms will be able to better extract and understand context from text on web pages, making it easier to extract relevant data.
- Greater use of machine learning (ML): ML algorithms can help automate the web scraping process by identifying patterns in data and making predictions about future data.
- More advanced web scraping tools: As the demand for web scraping increases, we can expect to see more advanced tools and platforms that make it easier for non-technical users to extract data from the web. Improved data quality techniques such as data cleaning and data validation will become more sophisticated, resulting in higher-quality data.

REFERENCE

[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 Web Scraping." Internet: https://seanholbert.wordpress.com/2011/07/15/scrappysimple-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 SIGMODWorkshop on Management of Semi Structured Data, Tucson, Arizona, May 1997."

[5] "Datahen."3 Advantages of web scraping for your enterprise" Internet: https://www.da tahen.com/3-advantages-web-scrapingenterprise/, May.17,2017""

[6] "https://en.wikipedia.org/wiki/Web_sc raping"

[7] https://www.webharvy.com/articles/ whatis-web-scraping.html

[8] http://resources.distilnetworks.com/ h/i/538 22104-is-webscraping- illegal-depends-on-what-the-meaning-of-the-word-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.pyth on.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."