2023 International Conference on Sustainable Emer in Innovations in En ineerin and Technolo (ICSEIET)

Web Scraping: Huge Data collection from Web

Chandradeep Bhatt[1]
CSE Department,
Graphic Era Hill University
Dehradun, India
bhattchandradeep@gmail.com

Gaitri[2]
CSE Department,
Graphic Era Hill University
Dehradun, India
gaitribisht943@gmail.com

Devendra Kumar[3]
Department of Computer Science
ABES Engineering College
Ghaziabad, India
devendra.arya@gmail.com

Rahul Chauhan[4]
Computer Science and Engineering
Graphic Era Hill University
Dehradun, India
chauhan14853@gmail.com

Ashish Vishvakarma[5]
Department of Information technology
Institute of Technology & Management
Uttarakhand, India
technology212@gmail.com

Teekam Singh[6]
Computer Science and Engineering
Graphic Era Deemed to be University
Dehradun, India
teekamsingh.iitr@gmail.com

***Abstract*— Web Scraping also known as data scraping means extracting the data from a particular website or web page. Extracting the data is useful for data analysis, collecting business data, real time data etc. Webpages contain a lot of information or data which is in unstructured format and cannot be used directly for data analysis and other purpose. So, after extracting the data it is converted into structured format. Extracting the data from web pages is useful but it is also illegal. We can scrap the data which is available publicly but in a limited request. Some websites may explicitly prohibit or limit scraping activities, so it is essential to respect their policies. Before scraping a website, review their terms of service, privacy policy, and any other relevant policies. Without owner's permission we cannot extract the private data from their websites. Selecting the website and conducting web scraping responsibly, we can collect desired data that can be used for various purposes, such as market research, price comparison, trend analysis, or gaining insights into consumer behavior.**

Keywords*— *Web scrapping, Python, Data Analysis, websites.

I. INTRODUCTION

Web Scraping is used for extracting useful and valuable data from a website for analysis, research or business intelligence. There are many programming languages which is used for web scraping. Python is one of these and mostly used for this. It has simple and clean syntax and it also provide many libraries like BeautifulSoup and Scrapy, which simplify parsing HTML documents and extracting data. Python enables the organizations to search the web, collect relevant data and convert it into designs without any difficulty.

Web scraping is useful for converting unstructured data into structured form and storing it into files like CSV and spreadsheets. However, it is important to consider legal and ethical implications, and follow proper data extraction procedures. This paper helps how to utilize Python tools for web scraping, which will enhance data analysis capabilities and helps in decision-making.

E-commerce website like Filpkart is well for its items and feedback. With the help of Python and its libraries one can scrap the product information like price, customer reviews and other information that will helps competitors to better understand customer preferences, market trends etc. In today's market, with the rapid growth of e-commerce, understanding consumer and business behavior has become increasingly crucial for firms to maintain their competitiveness.

Web scraping using python, with particular focus on the famous e-commerce website Filpkart, is mentioned here.

Web scraping has several applications in many industries. Web scraping has a variety of purposes, such as:

a) Data gathering and analysis: Web scraping extract data from websites, enabling companies and researchers to gather big or redundant data for analysis.

b) Training of machine learning and artificial intelligence models and algorithms: Web scraping is used to gather data for these models and algorithms. Researchers may build robust models that can forecast, categorize, or carry out tasks in acceptable language by merging different data.

c) Market research and competition analysis: Web scraping offers information on the market, rivalry tactics, and consumer preferences.

d) Automation: Instead of visiting multiple websites and copying information manually, one can write a web scraping program to scrap the data which saves time and increase efficiency.

e) Real-time data: Web scraping allow to extract and collect real-time data from websites which is useful for monitoring prices, social media trends, stock availability to stay up-to-date.

f) Lead Generation: Web scraping may be used to harvest contact details from websites, assisting organizations in lead generation and client database building.

g) Price comparison and monitoring: Online retailers may use the website to keep an eye on the prices of their rivals and modify their pricing strategy as necessary.

h) Sentiment Analysis and Brand Monitoring: Web scraping may be used to harvest data from social media and review websites in order to analyze consumer behavior or user reviews of a certain brand or product.

i) Material collection and news monitoring: Web scraping is used to collect news, blog articles, and other online material from a number of sources. This data may be used by news organizations to arrange material, track breaking news, and examine new patterns.

j) Content aggregation: Web scraping enables to collect content from different websites and display it on your own platform. This is useful for creating news aggregators, price comparison websites, or any other platform that requires data from multiple sources.

II. LITERRATURE SURVEY

Web scraping is widely used by a large number of individuals to gather data from multiple websites. Numerous

979-8-3503-2919-3/23/$31.00 ©2023 IEEE

2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET) | 979-8-3503-2919-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSEIET58677.2023.10303037

studies and resources have focused on web scraping using Python over the years. Those resources not only emphasize the benefits of online but also has its limitations and concerns. These includes certain restriction, legal and ethical issues, and the impact of advancements website and evolving technology.

John Smith, 2017, has created a web scraping library which simplifies the process of scraping data. It provides many functionalities like handling HTML parsing tasks and common scraping challenges. But due to captcha handling it creates conflicts with website owners. Ryan Mitchell, 2018, focused on various scraping techniques, utilizing APIs, and handling dynamic content. Mitchell explored different parsing library to extract desired data. Later, he faces challenges with large scale data scraping tasks. Jane Doe, 2019, mainly focuses on practices for handling user authentication, rate limiting and ensuring data quality. She also focuses on user privacy and data usage from social media platform. She highlights on respecting privacy rights, necessary permissions, and adhering to legal and ethical guidelines when scraping data from these platforms. Sarah Johnson, 2020, focuses on ethical issues and providing guidelines for safe data extracting. She promotes ethical scraping practices and provides guidelines for scraping data safely. Michael Anderson, 2021, compares different techniques for extracting data that rely on JavaScript.

Traditional scraping techniques may not be sufficient for dynamic content. So, he focuses on how to scrape such JavaScript driven websites. But problems arises when using scraping libraries together with JavaScript frameworks. These problems can result in inconsistencies and compatibility issues.

Table 1. Popular work of web scraping

S. No	Year	Author	Methods	Drawbacks
1	2017	John Smith	Python libraries like BeautifulSoup and Scrapy. HTML parsing and data extraction techniques.	Absence of a standardized protocol and conflicts with website owners caused by CAPTCHA handling.
2	2018	Ryan Mitchell	Scraping techniques such as HTML parsing, APIs, dynamic content.	Large-scale data scraping tasks was challenging to perform efficiently
3	2019	Jane Doe	Methods for handling user authentication, rate limiting, and data quality.	Ethical concerns regarding user privacy and data usage from social media platforms.
4	2020	Sarah Johnson	Ethical issues in web scraping and provides guidelines for scraping practices safely.	Challenges in interpreting and adhering to website-specific terms of service and scraping policies.
5	2021	Michael Anderson	Comparison between different techniques for web scraping websites that heavily rely on JavaScript to generate content.	Problems arises when using scraping libraries together with JavaScript frameworks. These problems can result in inconsistencies and compatibility issues.

III. METHODOLOGY AND DATASET

The goal of web scraping is to collect all data and information from various websites using technology such as beautiful soup, auto scraper, scrapy or selenium etc. Collecting the data and storing it into csv or excel files. With the help of python libraries, scraping the data from filpkart website. Firstly, pinpoint the relevant data that need to be extracted such as price, review, ratings etc. Then HTML source code of filpkart is parsed using python library beautiful soup. Requests library is used to send HTTP request to website's server.

Python's libraries that are used:

a. BeautifulSoup: It simplifies the process of web scrapping and parsing HTML and XML documents. It provides a powerful and user-friendly interface for extracting data from web pages. The library also supports various parsing libraries, giving flexibility in choosing the most suitable one according to our needs.

b. Requests: The library is used to sends the HTTP request to website's server and offers crucial features for managing cookies, headers, and sessions.

c. Pandas: It is open-source python library used for data analysis, data manipulation, and data processing.

The following steps involved as:

Step 1. Obtaining webpage HTML source code from the Filpkart website. The requests library will send the HTTP request to the website's server and then the response is retrieved. It will get the the web page's HTML content, which will be used for later processing.

Step 2. The HTML material that was downloaded from the filpkart website is parsed by the BeautifulSoup library.

Step 3. The useful data such as the product name, price, rating, and customer reviews, are located using the HTML structure and the inspection capabilities offered by web browsers.

Step 4. Using the BeautifulSoup library, the identified data components are extracted from the parsed HTML. The library offers ways to locate particular tags and get the text or properties related to those tags. The proper files are used to store the extracted data for further processing.

Step 5. CSV (Comma Separated Values) or Excel files are used for storing retrieved data. Functions to save data frames as CSV or Excel files are available in the Pandas library.

Step 6. Care is taken to respect website terms of services and legal obligations when scraping the web. To avoid excessive

or disruptive scraping, the system respects the robots.txt file on the website, and limit the number of requests done per second or minute.

IV. PROPOSED WORK

In order to scrape the data from web pages, we should respect their terms of services and privacy and should avoid from sending the server an excessive number of requests. Python is used to collect the desired data from filpkart for future analysis and its practical use involved scraping information from Filpkart. Python has many libraries such as beautifulsoup, requests, pandas which helps in scraping the data and managing the data. It involves sending the http request and parsing the html of that website.

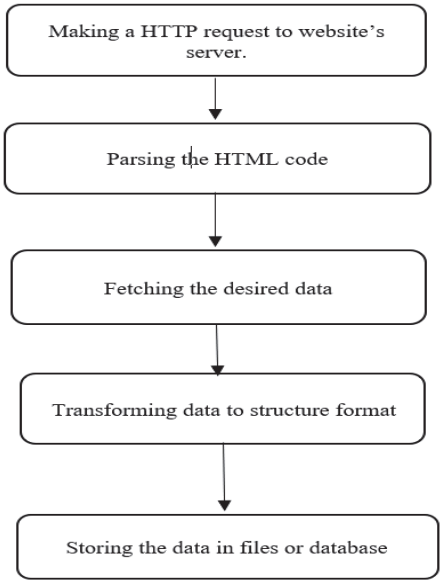

Fig. 1. Flowchart of the work

Given below is the brief description of every methodology used-

1. Find the url:- Get the url of the website you want to scrap.

For example, if we want to scrap the Flipkart website to scrap the name, price, ratings or description of laptops then the url for this page is https://www.flipkart.com/search?q=laptop&sid=6bo%2Cb5g&as=on&as-show=on&otracker=AS_QueryStore_OrganicAutoSuggest_1_3_na_na_na&otracker1=AS_QueryStore_OrganicAutoSuggest_1_3_na_na_na&as-pos=1&as-type=RECENT&suggestionId=laptop%7CLaptops&requestId=5399ebda-a795-40ad-a159-cb163e128f78&as-searchtext=lap

2. Inspect the website's page: The data is nested in tags. So, find the tag under which the data is present which we want to extract.

3. Extract the data: Extract the data you want to scrap like price, name, rating etc. by using beautifulsoup.

4. Storing the data: After extracting the required data, store it in CSV format or in excel.

V. RESULT

a) This report collects the useful data from the websites such as product's price, rating or customer reviews

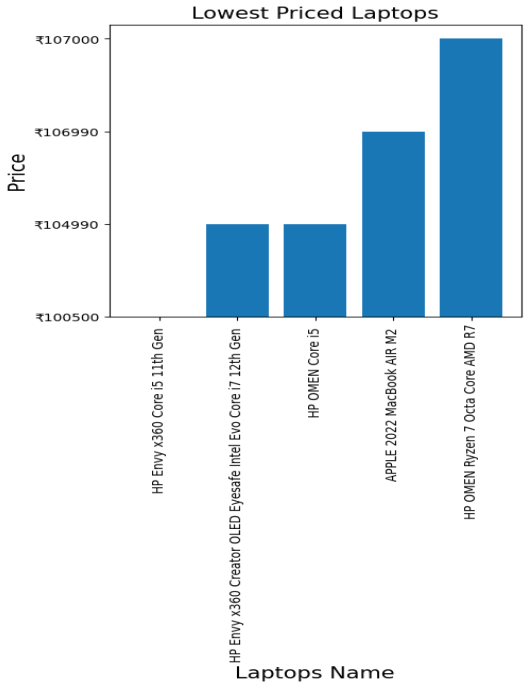

Fig. 2. The bar graph shows the prices of the laptops which is scraped from e-commerce website Flipkart. X-axis represents the product name and the Y-axis represents the prices of laptop.

b) Scraping the availability status of a product helps to track whether the item is in stock or out of stock.

c) Scraped data can be used to generate reports, visualize trends or make prediction.

	Product Name	Prices	Discount:	Rating & Review
2	Primebook 4G Andro	16,990	32% off	650 Ratings&217 Reviews
3	Primebook 4G Andro	18,990	24% off	131 Ratings & 51 Reviews
4	ASUS Vivobook 15 (	34,990	38% off	2,167 Ratings & 206 Revie
5	APPLE 2020 Macbo	79,990	19% off	9,937 Ratings& 865 Revie
6	HP 14s Intel Core i3	38,990	21% off	4,590 Ratings&396 Reviev
7	ASUS TUF Gaming A	54,990	27% off	476 Ratings&50 Reviews
8	ASUS Vivobook 15 (	44,990	36% off	1,110 Ratings&97 Review
9	Lenovo Legion 5 Ry	1,04,990	35% off	26 Ratings&5 Reviews
10	HP 247 G8 Athlon D	23,890	28% off	50 Ratings&4 Reviews
11	acer Aspire 7 Ryzen	49,999	41% off	5,903 Ratings&700 Reviev
12	HP 15s Intel Core i5	55,990	20% off	1,616 Ratings&114 Reviev
13	Lenovo IdeaPad Gar	68,990	30% off	171 Ratings&20 Reviews
14	realme Book (Slim) (	35,990	34% off	13,515 Ratings&1,956 Rev
15	HP 14s Intel Core i3	36,490	22% off	2,693 Ratings&231 Reviev
16	Lenovo ThinkBook 1	32,990	43% off	349 Ratings&65 Reviews
17	Lenovo Legion Ryze	84,990	29% off	9 Ratings&0 Reviews
18	MSI Bravo 15 Ryzer	59,990	17% off	65 Ratings&14 Reviews
19	HP 15s Intel Core i3	38,990	22% off	703 Ratings& 60 Reviews
20	Lenovo IdeaPad 3 R	43,975	45% off	1,905 Ratings&202 Reviev
21	Lenovo IdeaPad Ryz	72,990	37% off	75 Ratings&8 Reviews
22	ASUS Vivobook K15	47,990	40% off	1,278 Ratings&133 Reviev
23	ASUS TUF Gaming I	54,990	26% off	5,429 Ratings&587 Reviev
24	ASUS ROG Strix G1	84,990	26% off	311 Ratings&31 Reviews
25	Lenovo IdeaPad Slim	33,990	44% off	403 Ratings&36 Reviews

Fig 3. The output result shows the laptop's details such as name, prices, discount and rating and reviews.

VI. CONCLUSION

This Paper showed that using python for web scraping is an efficient way to collect data from Filpkart. We can gather important data like items, price, ratings and review using web scraping techniques.

The findings of this study highlight the significance of websites for gathering large amounts of data from e-commerce companies like Filpkart. The data may be utilized for a variety of things, including market research, pricing monitoring, rivalry analysis, and client analysis.

This paper also highlights the challenges of web scraping, including dealing with anti-scraping mechanisms, maintaining ethical practices, and ensuring data privacy. It is important to follow terms of services and guidelines of a websites to avoid legal issues. This study report shows the scraping of websites content become easier using python.

REFERENCES

[1] W. Jiahao, "Web Scraping using Python: Step by step guide" ResearchGate publications, 2019

[2] A. Kisseljov, "Following ICT bachelor programme offerings (studyinfo. fi) using web scraping with Python, 2023.

[3] B. Massimino, "Accessing online data: Web-crawling and information-scraping techniques to automate the assembly of research data. Journal of Business Logistics, 37(1), 34–42, 2016.

[4] M. Rose, "Web Scraping with Python and BeautifulSoup. Towards Data Science", 2020.

[5] R. Lawson, "Web scraping with Python. Packt Publishing Ltd, 2015.

[6] L. Wu, A. S. Mattila, C. Y. Wang & L. Hanks, "The impact of power on service customers' willingness to post online reviews. Journal of Service Research, 19(2), 224–238, 2015.

[7] A. V. Anand, K. G. Saurkar, S. A. Gode "An overview of web scraping techniques and tools" International Journal on Future Revolution in Computer Science and Communication Engineering, April 2018.

[8] S. O'Reilly, "Nominative fair use and Internet aggregators: Copyright and trademark challenges posed by bots," web crawlers and screen-scraping technologies. Loyola Consumer Law Review, 19, 273, 2006.

[9] R. N. Landers, R. C. Brusso, K. J. Cavanaugh & A. B. Collmus, "A primer on theory-driven web scraping: Automatic extraction of big data from the internet for use in psychological research," Psychological Methods, 21(4), 475–492, 2016.

[10] S. Vanden Broucke & B. Baesens, "Practical web scraping for data science," Apress, 2018.

[11] D. Doran & S. S. Gokhale, "Web robot detection techniques: Overview and limitations,"Data Mining and Knowledge Discovery, 22(1), 183–210, 2011.

[12] R. Lawson & J. Sharp, "Web Scraping with Python: Collecting More Data from the Modern Web," O'Reilly Media, 2015.

[13] A. Chinnathambi, "Web Scraping with Python and Selenium," Packt Publishing, 2018.

[14] W. McKinney & M. Wes, "Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython," O'Reilly Media, 2012.

[15] R. Vording, "Harvesting unstructured data in heterogenous business environments; exploring modern web scraping technologies, 2021