

173, Agaram Road, Selaiyur, Chennai-600073

**BHARATH INSTITUTE OF SCIENCE & TECHNOLOGY**

***Department of*** C***omputer*** S***cience &*** E***ngineering***

**WEB SCRAPING**

**TERM PAPER**

**SUB CODE: U18PRCS6P1**

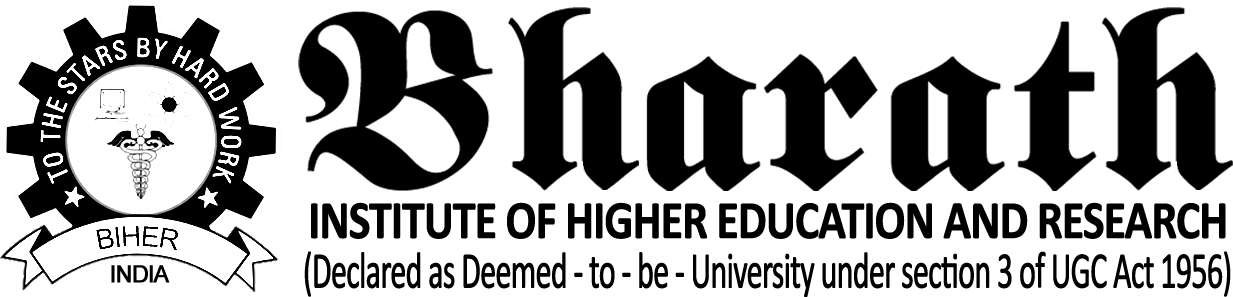
**Name: VOOTLA GNANESWAR  
 VURIBINDI VENKATA HARI PRASAD  
 YENIGALLA CHAITANYA SAI**

**Year: III Semester: VI Batch: 2019-2023 Section: CSE-T**

**U19CB096  
U19CB101  
U19CB124**

### University Register No.

**JUNE 2022**



173, Agaram Road, Selaiyur, Chennai-600073 ***Department of*** C***omputer*** S***cience &*** E***ngineering* SUB CODE: U18PRCS6P1**

**Name: VOOTLA GNANESWAR  
 VURIBINDI VENKATA HARI PRASAD  
 YENIGALLA CHAITANYA SAI**

**Year: III Semester: VI Batch: 2019-2023 Section: CSE-T**

**U19CB090  
U19CB101  
U19CB124**

### University Register No.

Certified that this is the bonafide record of work done by the above student in the **TERM PAPER (SUB CODE: U18PRCS6P1)** during the month **June 2022.**

#### Signature of Faculty-In-Charge Signature of Head of Department

*Submitted for the practical Examination held on ……………………………………*

*at Bharath Institute of Higher Education & Research.*

**Signature of Internal Examiner Signature of External Examiner**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **INDEX** | | |  | |
| **I. ABSTRACT** | | | **IV** | |
| **II. LIST OF FIGURES** | | | **V** | |
| **CHAPTERS** | | |  | |
| 1. **INTRODUCTION** | | | **1** | |
|  | | 1.1 OBJECTIVE | **1** | |
|  | | 1.2 SCOPE OF THE PROJECT | **1** | |
| 1. **LITERATURE SURVEY** | | | **2** | |
| **3. EXISTING SYSTEM** | | | **3** | |
| 1. **PROPOSED SYSTEM** | | | **4** | |
| 1. **SYSTEM DESIGN** | | | **5** | |
| 1. **REQUIREMENTS** | | | **8** | |
|  | 5.1 SOFTWARE AND HARDWARE REQUIREMENTS | | **8** | |
| **7. MODULES DESCRIPTION** | | | **9** | |
|  | 7.1 DATA ANALYSIS | | | **9** |
|  | 7.2 DATA CLEANING | | | **9** |
|  | 7.3 DATA PREPROCESSING | | | **9** |
|  | 7.4 DATA SPLITTING | | | **9** |
|  | 7.5 MODEL EVALUATION | | | **9** |
| **8. APPENDIX**  8.1 CODING  8.2 CODING IMPLEMENTATION | | | **11** | |
| **9. CONCLUSION** | | | **12** | |
| **10. FUTURE SCOPE** | | | **12** | |
| **11. REFERENCES** | | | **13** | |
| **12. PLAGIARISM** | | | **14** | |

# ABSTRACT

The main objective of web scraping is to extract information from one or many websites. Today’s search engines are equipped with specialized agents known as “web crawlers”. Then process it into simple structures such as spreadsheets, database. we develop flipkart using beautiful soup to scrape websites and store the information in excel sheets. The website flipkart to be scraped is operated upon by pandas which we create using AI which function as bots.

Web scraping replaces the need for manual data entry and more easily reveals trends among data collected. It can also aggregate information from multiple sources into one central location. While this application provides three specific examples of web crawling/scraping, it could be easily altered to better suit additional markets and/or needs.

A web crawler is a piece of code that travels the Internet and collects data from various web pages, also known as web scraping. Some web crawlers are autonomous and require no instructions once started. This project will focus on a user driven web crawler where user input will direct where the crawler goes and how the collected data is analyzed.

Web scraping is currently being used in various markets to collect and analyze information. This information can provide the ability to gain an edge over competitors such as showing the average price of their products. Web scraping has been around for over a decade but can be applied to new and emerging markets. With a few small changes in code most websites can be scraped for information

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Figure Title** | **Page No** |
| Fig 3.2 | SYSTEM ARCHITECTURE | **12** |
| Fig 5.2 | HTML PARSING | **17** |
| Fig 5.3 | DOM PARSING | **17** |
| Fig 5.4 | XPATH | **18** |
| Fig 5.6 | USAGE OF WEB SCRAPING | **19** |
| Fig 7 | DIFFERENT PROGRAMMING LIBRARIES | **21** |
| Fig 8.2 | FLOWCHART OF WEB SCRAPING | **24** |
| Fig 10.1 | CODE FOR IMPLEMENTATION OF BEAUTIFULSOUP | **26** |
| Fig 10.2 | WEBSITE (FLIPKART) USED FO SCRAPING | **27** |
| Fig 10.3 | OUTPUT IN PYCHARM | **27** |
| Fig 10.4 | CSV FILE OF SCRAPED WEBSITE | **28** |
| Fig 13.1 | PLAGIARISM REPORT - 1 | **31** |
| Fig 13.2 | PLAGIARISM REPORT - 2 | **31** |

## CHAPTER - 1

## INTRODUCTION

The World Wide Web is an interconnected network of information that consumers can access through websites. The way we share, acquire, and distribute data has changed dramatically thanks to the Internet. The amount of information available is always increasing. The quantity of content has also increased as a result of the use of the Internet as a new marketing and sales channel. Large amounts of data are provided by online retailers to characterize their products.

Access to knowledge base providers' databases is available. It is no longer practicable to manually track and record all available sources due to the disorderly growth. Web scraping took off at that point. In comparison to human data extraction, automated procedures allow for the capture of a large volume of data from the Web. Another word that has become quite essential in conjunction with Web Scraping is Meta Data. Web scraping collects a large amount of data, which enables for Meta Data analysis.

As the volume, diversity, and value of data expand, business executives must concentrate their efforts on the data that matters most. For organisations and customers, not all data is equal. Enterprises that can identify and take use of the crucial fraction of data that will produce substantial positive impact for user experience, solving complex challenges, and establishing new economies of scale will thrive throughout this data transformation. To exploit the immense potential of data, business executives should concentrate on identifying and servicing that one-of-a-kind, crucial

### 1.1 Objective

The paper's main argument is to consider a cycle of cleaning, modifying, and demonstrating data in order to obtain useful data for company dynamics. The goal of Data Analysis via Web Scratching is to extract meaningful data from information and make decisions based on the results of the inquiry.

The study has the following specific objectives:

* + - The purpose of the study is to extract data from various sources using web crawler BeautifulSoap programming.
    - The software is used to extract data via an application programming interface or as a general-purpose web crawler, depending on the needs of the customer.
    - To examine variations, comments, ratings, or anything else with an infinite number of alternatives

.

## CHAPTER - 2 LITERATURE SURVEY

Renita Crystal Pereira and Vanitha T, “Web Scraping of Social Networks,” International Journal of Innovative Research in Computer and Communication Engineering, pp. 237-240, Vol. 3, 2015. Kaushal Parikh, Dilip Singh, Dinesh Yadav and Mansingh Rathod, “Detection of web scraping using machine learning,” Open access international journal of Science and Engineering, pp.114-118, Vol. 3, 2018.Sameer Padghan, Satish Chigle and Rahul Handoo, “Web Scraping-Data Extraction Using Java Application and Visual Basics Macros,” Journal of Advances and Scholarly Researches in Allied Education, pp. 691 -695, Vol.15, 2018.

The use of social networking sites and the internet is growing by the day, such as Facebook, Twitter, LinkedIn, and others; user knowledge is also growing on the internet, which is accessible from anywhere. This also gives hackers an advantage when it comes to stealing information. From a business standpoint, social networking is critical in the development of the concept of growing revenue. It will assist consumers in achieving rapid shopping and saving time, similar to online buying. Supporting the company and earning from it, on the other hand, has advantages.

Kaushal Parikh proposed a web scraping detection with the help of machine learning It is valuable for research dependent companies. Web scraping has forever been a difficult preventive attack. Every time a company places its data on internet, it is probable that it could be copied and pasted and then utilized in the other point of view without the corporation knowing itself about it. A lot of protection mechanisms have already been in place but some of them continue to be ignored. The significance of machine learning therefore steps in. Machine learning is quite effective on pattern detection. Therefore if we succeed in making the machine understand a cadence of intruder then it will avoid these types of threats from occurring. Web scraping solutions are aimed primarily at translating complex data obtained through networks into structured data that could be stored and examined in a central database. Web scraping solutions thus have a significant impact on the result of the cause.

Sameer Padghan projected an approach where data extraction is done from web pages in assistance with web scraping easily. This method would enable the data to be scrapped from numerous websites that will minimize human intervention, save time and also enhance the quality of data relevance. It will also support the user in gathering data from the site and to save the data to their intent and use it as the individual wishes. The scraped information may be used

for database development or for research purposes and also for different similar activities. The scraping used would increase significantly and will often encroach on the framework to obtain the details. However the scraping can be stopped by using effective and safe-web scraping methods. This method should be treated as a blessing that must be used carefully for the advancement of human races.

Anand Saurkar discovered latest technique named Web Scraping. Web scraping is a quite important methodology used to produce structured data based on the unstructured data available on the internet. Scraping formed structured data, subsequently collected and evaluated in spreadsheets in central database. This research focuses on a summary of the data extraction process of web scraping, various web scraping strategies and most of the latest tools utilized to scrap web. The primary function of this methodology has been to get webbased information and integrate this into a specific repository. The authors addressed the basics of Web processing in this article. They concentrated on the Web scraping techniques. The final part of the paper presents a summary of the numerous technological resources that are available for effective web scraping in the industry.

In the field of commodity price studies, Federico Polidoro focused on the results of online scraping evaluation methodologies, with a special focus on user electronics services and items. Despite the fact that the research was completed in a short period of time, as evidenced by the following, it permitted the achievement of significant, but not conclusive, innovative efficiency outcomes. Web scraping tactics employed in the growth study will expose the researcher to a larger volume of data than is currently available in the data set, perhaps increasing the growth estimate.

## CHAPTER – 3

## EXISTING SYSTEM

The manual online data extraction technique in the existing system has two key flaws. For starters, it is incapable of accurately estimating expenditures and can rapidly inflate them. As more data is collected from each website, the expenses of data acquisition rise. Manual extraction necessitates the hiring of a large number of employees, which greatly raises the cost of labour. Second, each hand extraction has been shown to be prone to errors. Furthermore, if a business process is very complex, data cleanup can be costly and time-consuming. The faults and data cleanup processes associated with the Manual approach are illustrated in the diagram below

## CHAPTER – 4

## PROPOSED SYSTEM

Web scraping (also known as web harvesting or web data extraction) is a technique for extracting data from websites using computer software. Typically, such computer programs re- create human exploration of the World Wide Web by using either a low-level Hypertext Transfer Protocol (HTTP) or installing a full-fledged internet browser, such as Internet Explorer or Mozilla Firefox. Web scraping is synonymous with web ordering, which lists data on the internet using a web crawler and is a common strategy used by most web indexes.

Web scraping, on the other hand, focuses on converting unstructured web data, typically in HTML design, into ordered data that can be saved and analysed in a focal neighborhood data set or accounting page calculation of the displacement. To record the co-ordinates of the eyebrow, the pressure identification module analyses the parallel picture from the limit left top. The stress detection module records the co-ordinates of the eyebrow by scanning the binary image from the extreme left top.The offline displacement computation sub-modulecalculates the shifting of the eyebrow using the obtained eyebrow coordinates, and then determines the variance of the displacement.

The classifier sub-module is trained offline are employed to determine the presence of emotion. The integrated decision of individual frames eventually determines the level of stress involved. Web Scraping is a technique to extract structured data from websites. WSAPI is the

platform that enables an organization to extend their existing web based system, as well designed set of services for creating new channels, developer integration or partner integration.

# CHAPTER – 5

# SYSTEM DESIGN

## System Architecture

Fig 3.3: system architecture

Web Scraping is a strategy to separate organized information from sites. WSAPI is the stage that empowers an association to expand their current electronic framework, too planned arrangement of administrations for making new channels, designer mix or accomplice joining. It assists with offering spotless and organized information from existing sites, so the information can be easily devoured by unique frameworks. The innate plan assists engineers with fusing site changes without influencing the extraction rationale by moving them to designs. There are numerous particular reasons why organizations might need to scratch their site; one of the essential explanation being the inaccessibility of APIs.

### METHODS OF WEB SCRAPING

The methods of Web Scraping evolved together with the World Wide Web. Not all listed methods were available at the beginning. There are two examples to mention, because these are presently the most used techniques.

Since 2000 the Document Object Model (DOM) became more popular in DHTML. A broader acceptance later on allowed the HTML Parsing technique to evolve to DOM Parsing.

Second example are Application Programming Interfaces (APIs). This technique is the youngest on the list, the growth of available content APIs is dated from 2005. According to ProgrammableWeb.com the number of APIs has grown within 8 years from 0 to 10302. (Berlind, 2015)

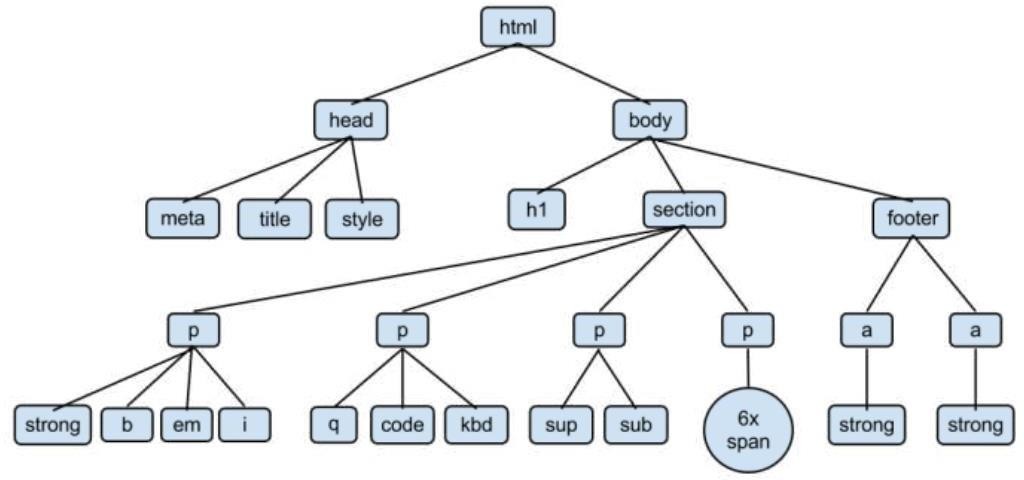
## 5.2 Manual Scraping

Manual scraping is still an option in specific situations. These situations are:

* + - When the amount of data is minimal,
    - When the data being scraped does not require a repetitive task,
    - When setting up automated scraping would take longer than the data collection itself.
    - Possibly security measures or specific characteristics of the website do not allow automated methods

## 5.3 HTML Parsing

Web sites don’t always provide their data in comfortable formats such as .csv or .json files. HTML Pages are created by the server as a response to a user’s request. At this point server software is not relevant, rather the output in the browser is important. Analysis of the HTML structure (simple page sample provided in the Figure 3) in the provided web page will show repeated elements. With a programming language script or Web Scraping tool, each page with similar pattern can be used as a source for data.



## 

Fig 5. 2: HTML parsing

* 1. **DOM PARSING**

Document Object Model (DOM) Parsing is an evolution of HTML Parsing based on developments of the language and browsers which lead to the introduction of the Document Object Model. DOM is heavily used for Cascading Stylesheets (CSS) and JavaScript. Integration of DOM revealed new possibilities for addressing some specific parts of the webpage .

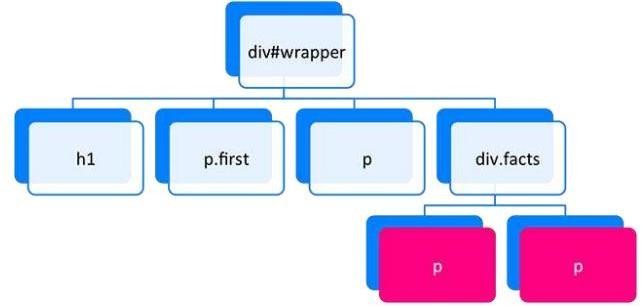


Fig 5.3: DOM parsing

## XPath

Similar addressing possibility as DOM provides XPath (XML Path Language). The name suggests a usage for XML documents. It is applicable also to HTML format. XPath requires a more precisely structured webpage than DOM and has the same possibility to address segments within the webpage. Figure 5 shows the document structure as interpreted in XPath.

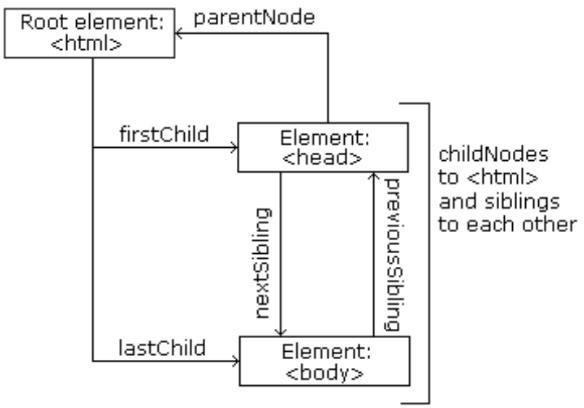


Fig 5.4:XPath

## APIs

Whilst the previous methods work to scrape human-readable outputs, Application Programming Interface(API) expects an application as a communication partner. Thus APIs are often named as machine-readable interfaces (versus human-readable). Even APIs were introduced much later than the WWW, and their growth was very fast. The world of APIs is fragmented. For a simple overview and orientation were API Directories created. Most of the available APIs are registered and described in the directory with relevant links to the sources. Two examples of such directories are - Programmable Web (https://www.programmableweb.com) and APIs (https://apis.guru/). API Directories also provide their own API, which allows users to search in their database for API Sources. A standard HTTP Request sent to an API Endpoint returns an answer from server. Each API has its own specification and options. The format of the answer can be set as option in the request. The most widely used format for API communication is JSON.

### Parse hub

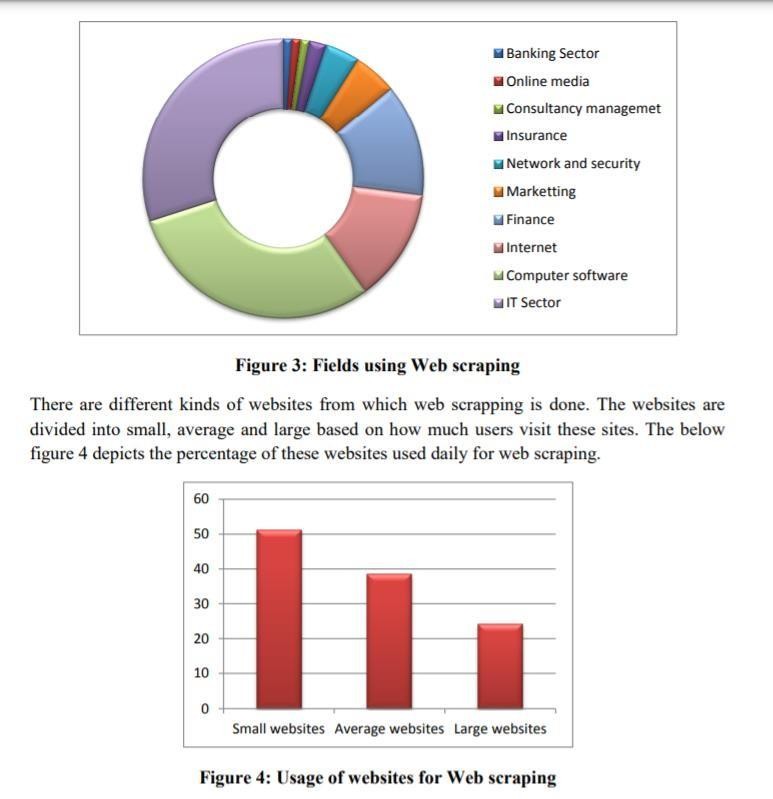
This is an internet-based scraping application designed to scrape over multiple and single URLs. The software will evaluate, capture, and convert data through sites into relevant data.

Fig 5.6: Usage of webscraping

**CHAPTER – 6**

# REQUIREMENTS

**SOFTWARE REQUIREMENTS**

PROGRAMMING LANGUAGE: PYTHON

IDE: GOOGLE COLAB

TECHNOLOGIES USED: WEB SCRAPING

PROGRAMMING LIBRARIES: BEAUTIFUL SOUP, PANDAS

**HARDWARE REQUIREMENTS**

Processor - i5-10th Gen

Speed - 1.6 GHz

RAM - 8 GB

Hard Disk - 512 GB

# CHAPTER – 7

# MODULES DESCRIPTION

# 7.1 PICK A WEBSITE AND DESCRIBE YOUR OBJECTIVE

# Browse through different sites and pick on to scrape. Identify the information you'd like to scrape from the site. Decide the format of the output CSV file. Summarize your project idea and outline your strategy in a PyCharm notebook. Use the "New" button above.

# 7.2 USE THE REQUESTS LIBRARY TO DOWNLOAD WEB PAGES

# Inspect the website's HTML source and identify the right URLs to download. Download and save web pages locally using the requests library. Create a function to automate downloading for different topics/search queries.

# 7.3 USE BEAUTIFUL SOUP TO PARSE AND EXTRACT INFORMATION

# Parse and explore the structure of downloaded web pages using Beautiful soup. Use the right properties and methods to extract the required information. Create functions to extract from the page into lists and dictionaries.

# CREATE CSV FILE(S) WITH THE EXTRACTED INFORMATION

# Create functions for the end-to-end process of downloading, parsing, and saving CSVs. Execute the function with different inputs to create a dataset of CSV files. Verify the information in the CSV files by reading them back

# CHAPTER – 8

# APPENDIX

The methodology used for the paper 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.

## Coding

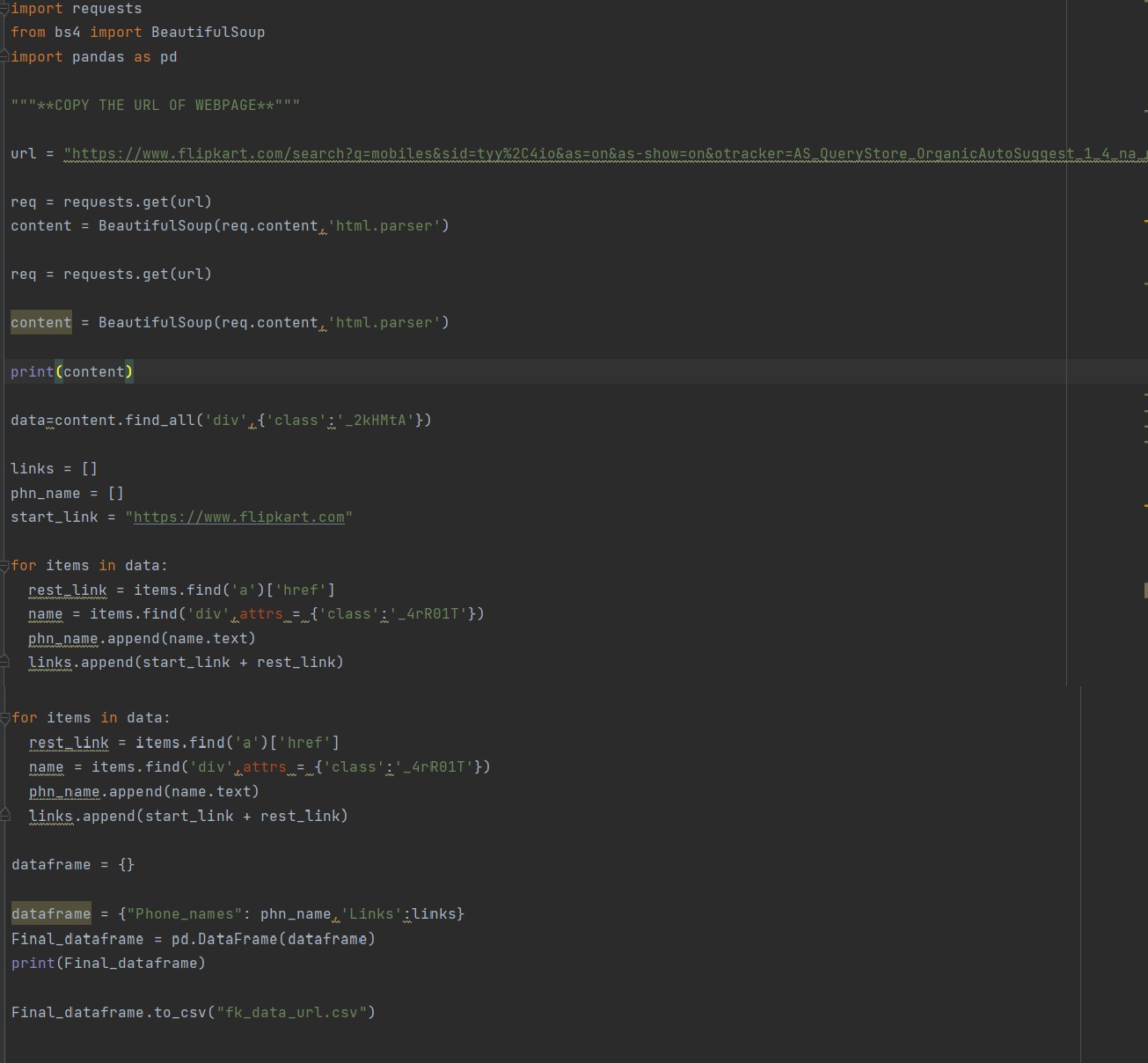
The basic web crawling script used for the project which shows the data crawled and stored in the database of the products from a social network site.

Fig 8.1: Code for Implementation of Scrapy

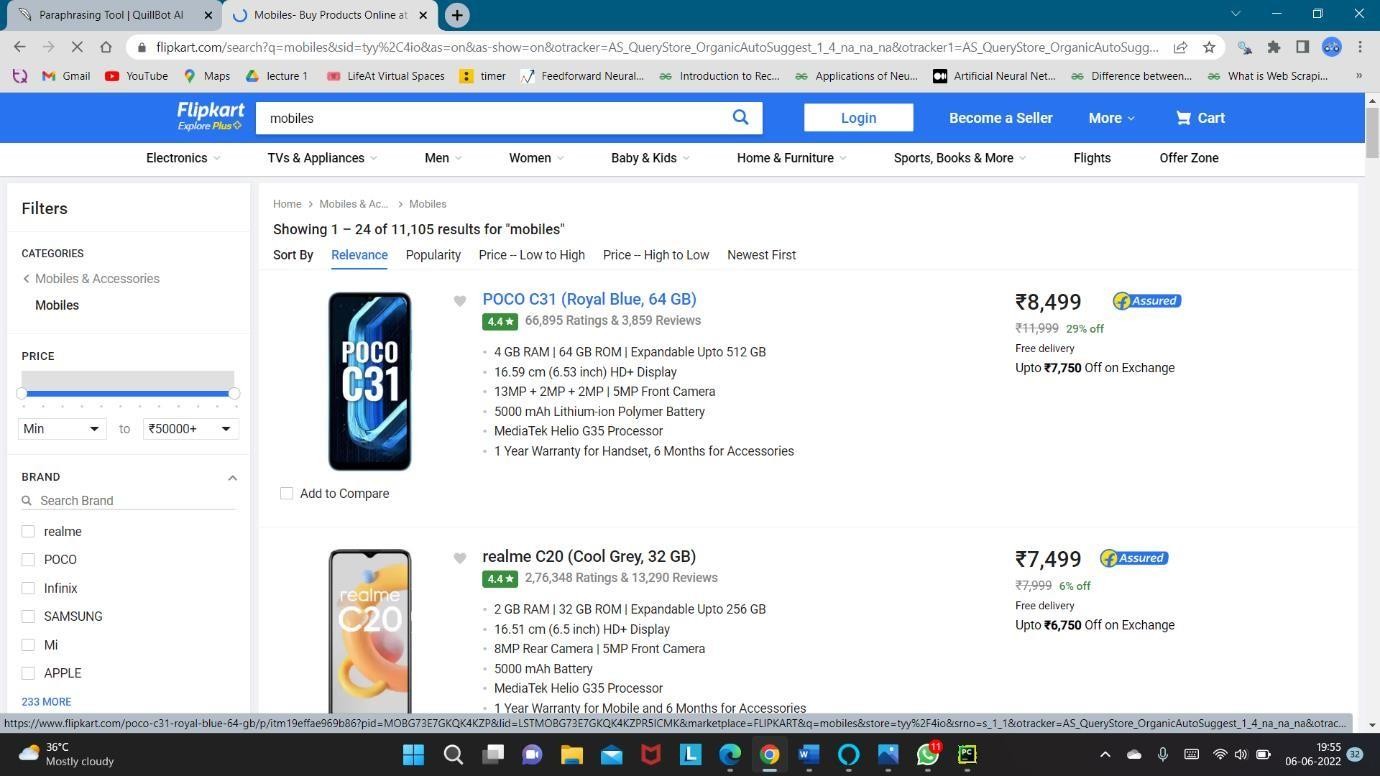
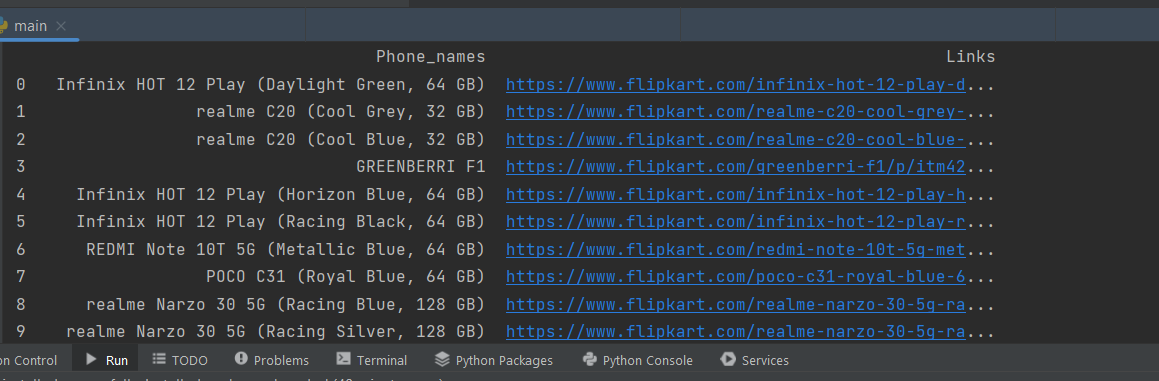


Fig 8.2: Website(Flipkart) used for scraping

The project was tested by 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.



## 

Fig 8.3: result in Pycharm

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 spreadsheet.



Fig 8.4: csv file of scraped website (flipkart)

* 1. **Code Explanation**

The code starts by importing the libraries needed for the program.

The code then creates a variable called url.

https://[www.flipkart.com/search?q=mobiles&sid=tyy%2C4io&as=on&as-](http://www.flipkart.com/search?q=mobiles&sid=tyy%2C4io&as=on&as-) show=on&otracker=AS\_QueryStore\_OrganicAutoSuggest\_1\_4\_na\_na\_na&otracker1

=AS\_QueryStore\_OrganicAutoSuggest\_1\_4\_na-na-na and it also sets the suggestion Id to mobiles%7CMobiles .

The next line of code requests data from this webpage using requests library, which returns a Scrapy object with all of the HTML content on that page as its contents.

This object is then parsed into an array of strings using html.parser function in bs4 library, and these are stored in content variable.

Next, we print out what's inside this object so that we can see what's going on: print(content)

The code is a Python script that would be used to import the data from a webpage into a pandas Data Frame.

The code above first imports the requests library and then uses it to get the content of an URL which is Flipkart's search page.

**CHAPTER 9**

**9.1 CONCLUSION**

Scraping technology for data extraction is a new and developing activity in the technology harvesting arena. Although many businesses currently use manual data extraction methods, Web scraping technologies are set to change the way data is extracted in the future. With exponential growth in this industry, the day is not far off when it will become a phenomenon, and most firms will recognise the importance of scraping innovation and how it allows them to stay well ahead of the competition. This paper provides an overview of Web scraping technology, including what it is, how it works, and the most widely used tools and technologies. Management of online media in the banking sector Insurance Security and the network Finance Marketing Computer software for the internet 0 – 10 – 20 – 30 – 40 – 50 – 60 – 70 – 80 Websites for small businesses

* 1. **FUTURE SCOPE**

Nowadays tones and tones of data generating every day through websites, applications and browsers. But processing and cleansing of these data is not so easy as generating data. So handling these data is difficult. This article will let you know the scraping data in websites, scopes, benefits and also drawbacks. Everything in this world has two sides good and bad. The betterment here is most of each sides matters.

**CHAPTER – 11**

**REFERENCES**

[1] Renita Crystal Pereira and Vanitha T, “Web Scraping of Social Networks,” International Journal of Innovative search in Computer and Communication Engineering, pp. 237-240, Vol. 3, 2015.

[2] Kaushal Parikh, Dilip Singh, Dinesh Yadav and Mansingh Rathod, “Detection of web scraping using machine learning,” Open access international journal of Science and Engineering, pp.114-118, Vol. 3, 2018.

[3] Sameer Padghan, Satish Chigle and Rahul Handoo, “Web Scraping-Data Extraction Using Java Application and Visual Basics Macros,” Journal of Advances and Scholarly Researches in Allied Education, pp. 691-695, Vol.15, 2018.

[4] Anand V. Saurkar, Kedar G. Pathare and Shweta A. Gode, “An Overview On Web Scraping Techniques And Tools,” International Journal on Future Revolution in Computer Science & Communication Engineering, pp. 363-367, Vol. 4, 2018.

[5] Federico Polidoro, Riccardo Giannini, Rosanna Lo Conte, Stefano Mosca and Francesca Rossetti, “Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation,” Statistical Journal of the IAOS, pp. 165-176, 2015.

[6] Jan Kinne and Janna Axenbeck, “Web Mining of Firm Websites: A Framework for Web Scraping and a Pilot Study for Germany,” 2019.

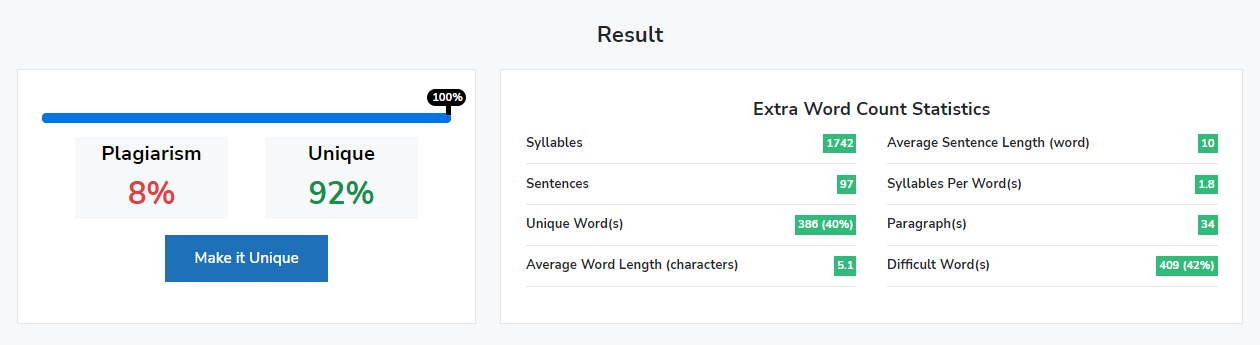
[7] Ingolf Boettcher, “Automatic data collection on the Internet,” pp. 1-9, 2015.

[8] Erin J. Farley and Lisa Pierotte, “An Emerging Data Collection Method for Criminal Justice Researchers,” Justice Research and statistics association, pp. 1-9, 2017

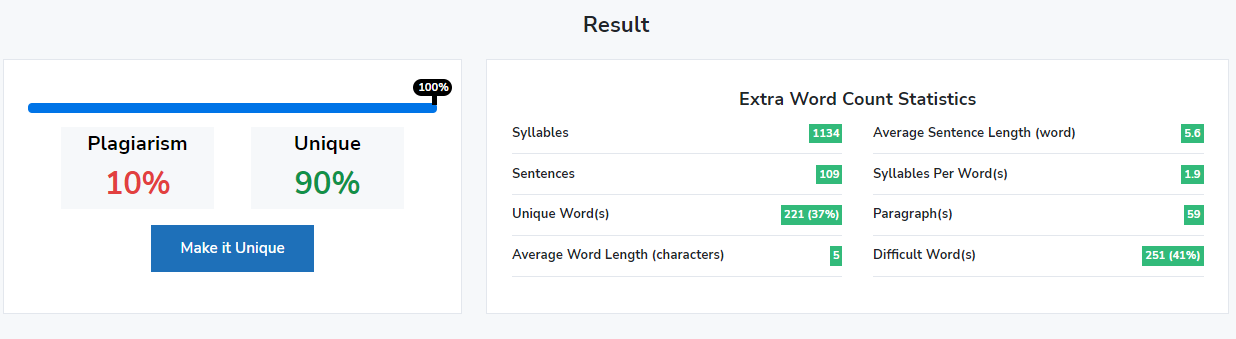
**CHAPTER – 12**

# Plagiarism Report

## Fig 11.1

****

**Fig 11.2**

****