# WEB SCRAPING USING PYTHON

**A REPORT on project baSED LEARNING**

**(semester -II)**

*Submitted by*

**NAME OF THE CANDIDATE(S)**

10755 Yash Sonavane

10756 Piyush Wajage

10761 Shashank Bejgamwar

10763 Viraag Borundiya

10766 Jatin Yadav

10769 Bhavesh Patil

**FIRST YEAR ENGINEERING**



# Society for Computer Technology and Research’s

# PUNE INSTITUTE OF COMPUTER TECHNOLOGY

# DHANKAWADI, PUNE – 43

**A.Y. 2021-22**

**Abstract**

The Internet is large collection of data in various domains which are taken from

different sources. It is origin of large information and data source about

Everything present and happening in the world. Data that is present on internet is not

well-structured and mannered way, that’s why it takes much time for users to search

relevant information that they are searching for because of unstructured data. There is

problem of getting useful data in less time, so for this Web Scraping is an effective

alternative. Web Scarping means to extract useful data according to recommendations

from different sites and presenting them in structured way. Then the useful data is

reused for analytical purpose. We will see which tools and coding is needed for Web

scraping in this project. Web scraping helps unstructured data to be get converted to

structured data by extraction of useful data from sites.

**ACKNOWLEDGEMENT**

We have taken efforts for this project, this was possible only because of support provided by Faculties, Seniors, YouTube, and teammates. Their support helped us to solve various issues faced by us during the project. We would also like to express sincere thanks to all of them.

We are thankful to PUNE INSTITUTE OF COMPUTER TECHNOLOGY for providing us opportunity to work on a real life daily based application project. My appreciation and thanks goes to our fellow mates who inspired us to work on issues faced by college students.

Mentioning special thanks to our PBL Co-Ordinator and mentor Dr. Shivaji Mundhe for his valuable guidance providing effective suggestions whenever needed, also special gratitude to free Codecamp, Geekforgeeks, StackOverflow, therealpython, for Coding and Scraping skills which were needed in the project.

Place: Name of Student (in Capital) & Sign

**TABLE OF CONTENT**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Title** | **Page No.** |
| **1** | **Introduction** |  |
| **2** | **Literature** |  |
| **3** | **Methodology** |  |
| **4** | **Setup and Working** |  |
| **5** | **Tools** |  |
| **6** | **Result /Application** |  |
| **7** | **Conclusion** |  |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure No** | **Title** | **Page No** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Nomenclature**

|  |
| --- |
| API APPLICATION PROGRAMMING INTERFACE |
| CSV COMMA SEPARATED VALUES |
| HTML HYPERTEXT MARKUP LANGUAGE  URL UNIFORM RESOURCE LOCATOR  VS Code VISUAL STUDIO CODE  OS OPERATING SYSTEM |
| XML EXTENSIBLE MARKUP LANGUAGE |

**Chapter 1**

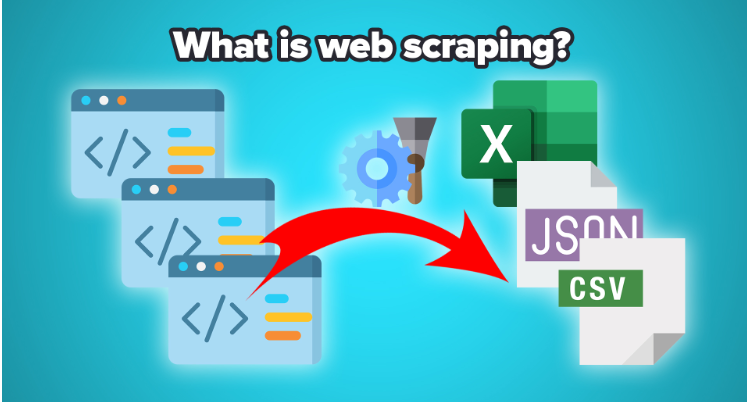
**INTRODUCTION TO WEB SCRAPING**

**Web Scraping with Python**

## Imagine that you have to pull a huge amount of data from websites and you want to do it quickly without wasting hours on it. How can you do it without having to go to each website and get data? Web Scraping helps us to overcome these problems. Web Scraping makes this task easier and faster.

## ****What is Web Scraping?****

Web scraping is the default method used to extract large amounts of data from websites. Web scraping helps to collect this unstructured data and to store it in an orderly fashion. These include using online services, particular API’s or even creating your code for web scraping from scratch. Many large websites, like Google, Twitter, Facebook, etc. have API’s that allow you to access their data in a structured format. This is the best option, but there are other sites that don’t allow users to access large amounts of data in a structured form



**How does web scraping work?**

Web Scraping can extract all the data available on particular sites or the specific data that a user wants. Ideally, it’s best if you specify the data you want so that it can only extract that data quickly.

For example, you might want to scrape a Flipkart page for the types of Mobile phones or Laptops available, but you might only want the data about the prices of different mobiles, and laptops and not the customer opinion.

**Is Web Scraping Legal?**

Now let’s talk about **whether web scraping is legal or not**. Some sites do just allow it when used legally. Web scraping is a tool that can be used in the right way or wrong way.

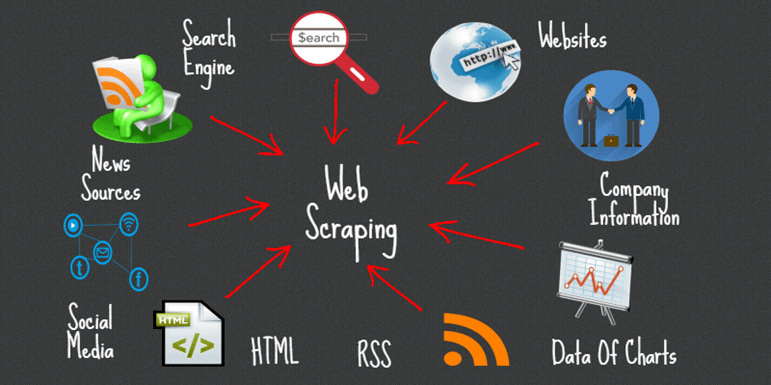
Web scrapping becomes illegal if someone tries to scrap non-public data such as data available on government websites. Nonpublic data is not reachable to everyone; if you try to extract such data then it is a violation of the legal term.

There are several tools available to scrap data from websites, such as:

* Scrapping-bot
* Scrapper API

Just remember, not everyone wants you pulling data from their web servers. Always check a website’s Terms of Use before you start scraping, and be respectful about how you time your web requests so that you don’t flood a server with traffic.

**Chapter 2**



**LITERATURE**

**(A summary taken from past research papers and references)**

**Paper 1: Data Analysis by Web Scraping Using Python:**

This paper demonstrates standardized data tests based on root-and-effect relationships, consisting of a small-scale experimental model, abstract and mathematical experiments, and a clever method of performing extrapolation tests. Its method is distributed in three parts: the web filter pulls the appropriate connection to the web, after which the information is extracted to get the information from the source joins, and finally puts that information in a CSV document. Due to the large local area and facilities of the Python library and the ease of typing in the Python language style, it is appropriate for the information required by Scraping from the relevant website.

**Paper 2:Web Scraping Using python:**

Learn the techniques of scratching and crawling to get unlimited information from any web resource in any organization. Ideal for engineers, security professionals, and Python-familiar webmasters, this book trains important scraping tools on the web, but also digs into advanced topics, for example, investigating crude information or using scams to check the foreground.

**Paper 3: Web Scraping with Python: Successfully scrape data from any website with the power of Python:**

The Internet contains very useful information formats for any site that is collected, which is usually freely opened for free. However, this information can no longer be used successfully. It is embedded within the structure and style of the spaces and must be carefully separated to be valuable. Scratching the web is becoming increasingly important as a way to easily integrate and optimize the amount of data accessible on the web. Using a specific language like Python, you can extract information from complex sites using a basic program.

**Paper 4 : ISR paper David on web scraping using python**

To know how the data extraction process has evolved with much one needs to understand the techniques involved in this web scraping method it is important to scrub it has been around for as long as the web. The effect of business web scraping honestly has been to get easy business profits and integrate things like undermining a competitor's special value, tracking, constructive promotion efforts, diversion of APIs, and internal and external hijackings and information.

**Paper 5 - An Introduction to Web Scraping for Research**

It’s important to remember that because web scraping involves the collection of data produced by others, it’s necessary to consider all the potential privacy and security implications involved in a project. Prior to your project, ensure you understand what constitutes sensitive data on campus and reach out to both your IT and IRB about your project so you have a data management plan prior to collecting any websites.

**Chapter 3**

**METHODOLOGY**

1. Pick a website and describe your objective

* Identify the information you want to scrape from the website. Decide the format of the output in CSV file.

1. Use the requests library to download web pages

* Inspect the website's HTML source and identify the right URLs to scrape.
* Save web pages locally using the requests library.

1. Use Beautiful Soup to parse and extract information

* Parse and explore the structure of 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.

1. Create CSV file(s) with the extracted information

* Create functions for the end-to-end process of parsing, downloading, and saving CSV files.
* Verify the information in the CSV files by reading them back using Pandas.

**Chapter 4**

**SETUP AND WORKING**

**Platform used:**

Visual Studio Code, also commonly referred to as VS Code, is a source-code editor made by Microsoft for Windows, Linux and macOS. Features include support for debugging, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded Git.

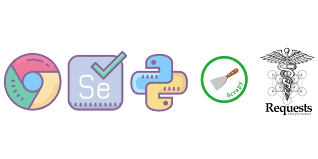
**Initial setup:**

**(a) Creating a Virtual Environment**

The creation of a separate Virtual Environment for the project is essential as it allows the program to install and use the Python libraries without creating any conflicts with other programs using the same libraries. It is required to create the Virtual Environment in the same folder as of the entire Project.

**(b) Installing Python Libraries**

The Python libraries may be installed either by using the Command Prompt or by using the PyCharm IDE Terminal. We use the “pip” command for the installation of required libraries.



In this project, we used various Python libraries:

**Beautiful Soup:**

This library is used to parse the html documents and defines the html tags. It creates a parse tree for parsed pages, which extract data through html. It receiving data from html, xml and other markup languages.

Suppose you have found some web pages that display data relevant to your research, such as date or address information but this does not provide a way to download the data directly.

Beautiful soup helps you pull special context from a webpage, remove html tags and save information.

**Requests:**

This is the most basic python library. It is used for http requests. This allows us to send http requests such as GET, POST etc. in websites so that we can retrieve data from the online websites. This library is simple and it can be used easily.

**(c) Programming the Web Scraper**

We use VS Code to create a Python program inside the Project folder. In the program, we import the installed python libraries first. The “webdriver” interface of Selenium library is used to implement the browser automation code. Request library code makes request to the web servers and BeautifulSoup library code is used to identify the HTML tags and parse the data from the web server. Pandas library code is used to convert the parsed data into desirable format.

**Working**

The Python program contains a function, which takes a word as an argument. The URL of a certain website is used to navigate to the desired web page by using Selenium “webdriver” interface. We use the “headless” option in the webdriver to hide the browser automation process. The search bar of the web page serves as a navigation tool for changing the web page contents. It can be used by putting the HTML tag for the search bar in selenium code. We can send the desired search elements by using the “Keys” module in Selenium. The word taken as the argument is then sent through Keys to the search bar. This word makes the website navigate to the desired web page. The current address of web page is taken as the source and all the required contents are then extracted using BeautifulSoup. BeautifulSoup uses tags to identify the specific elements required for extraction. A Python dictionary element can be used to facilitate the proper storage and access of data elements. The dictionary element allows the data to be accessed in a structured manner using a key associated to the data element.

During the execution of the Python program, an input is passed to the Scraper function that takes it as an argument and the data from the web page is extracted after function execution. We can now convert the extracted data into desired format using Pandas library.

After the program execution, a data file is created in the previously determined format that can be accessed to view the stored information.

**Chapter 5**

**TOOLS**

* **VS Code**: VS Code is also known as Visual studio code which is a source-code editor by Microsoft for Windows, Linux and mac OS. It has features like support for debugging, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded Git. It provides a virtual environment to code in various coding languages like HTML, C++, JAVA, JAVASCRIPT, PYTHON etc.



* **PIP tool**: PIP stands for Package installer for Python. It is effective package system which is used to install also manage various software systems. It connects to a special online package system which is known as python package index.
* **Selenium**: It is an automation-testing tool for automating web browsers. It is an open-source suite for different web browsers and platforms. Selenium supports different languages like c, java, python etc. It runs automation script smoothly. Selenium provides python binding APIs, allowing it to access web drivers such as chrome, Firefox etc. It provides flexibility and is free of cost.

It performs actions like:

• Element clicking

• Page refreshing

• Go to website link

* **Pandas:** Pandas is a python library or a python package that provides an expressive data structure with high performance. It is a fast, flexible, and most powerful library widely used for data analysis, data manipulation, and data storing. It is an open-source for data analysis and data manipulation. Pandas provide a heterogeneous data type in a data frame. Pandas data frame contains two-dimensional data structure with row and column axis and by default, it creates indexing begins with 0. It stores the data in CSV files, JSON, and Excel Sheets.

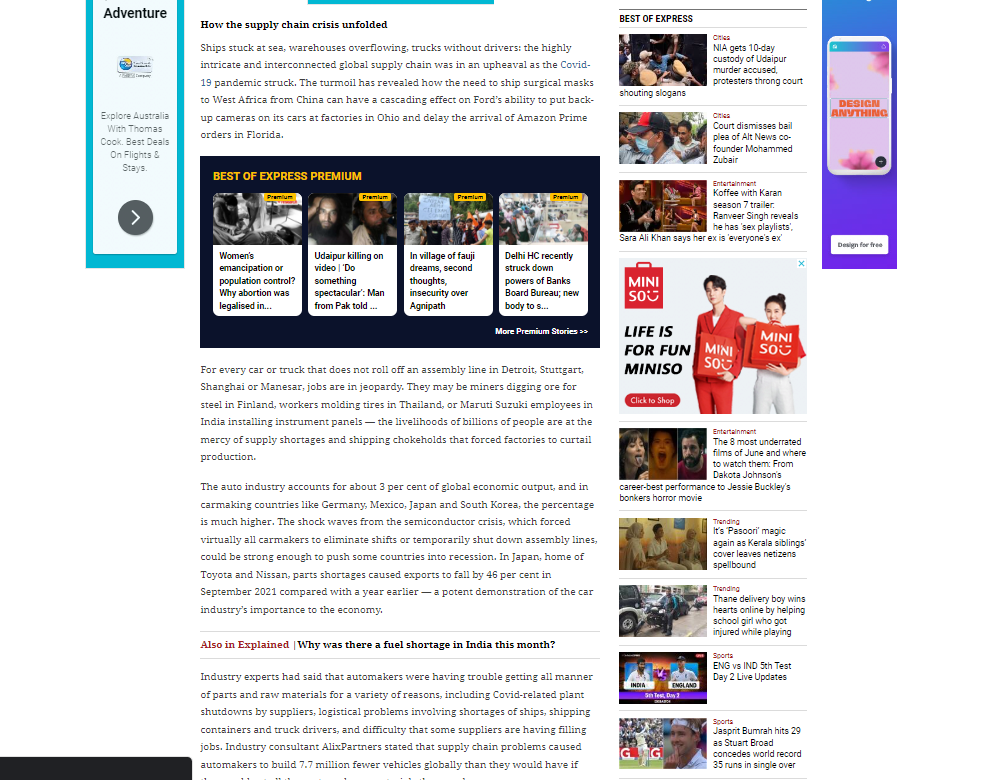
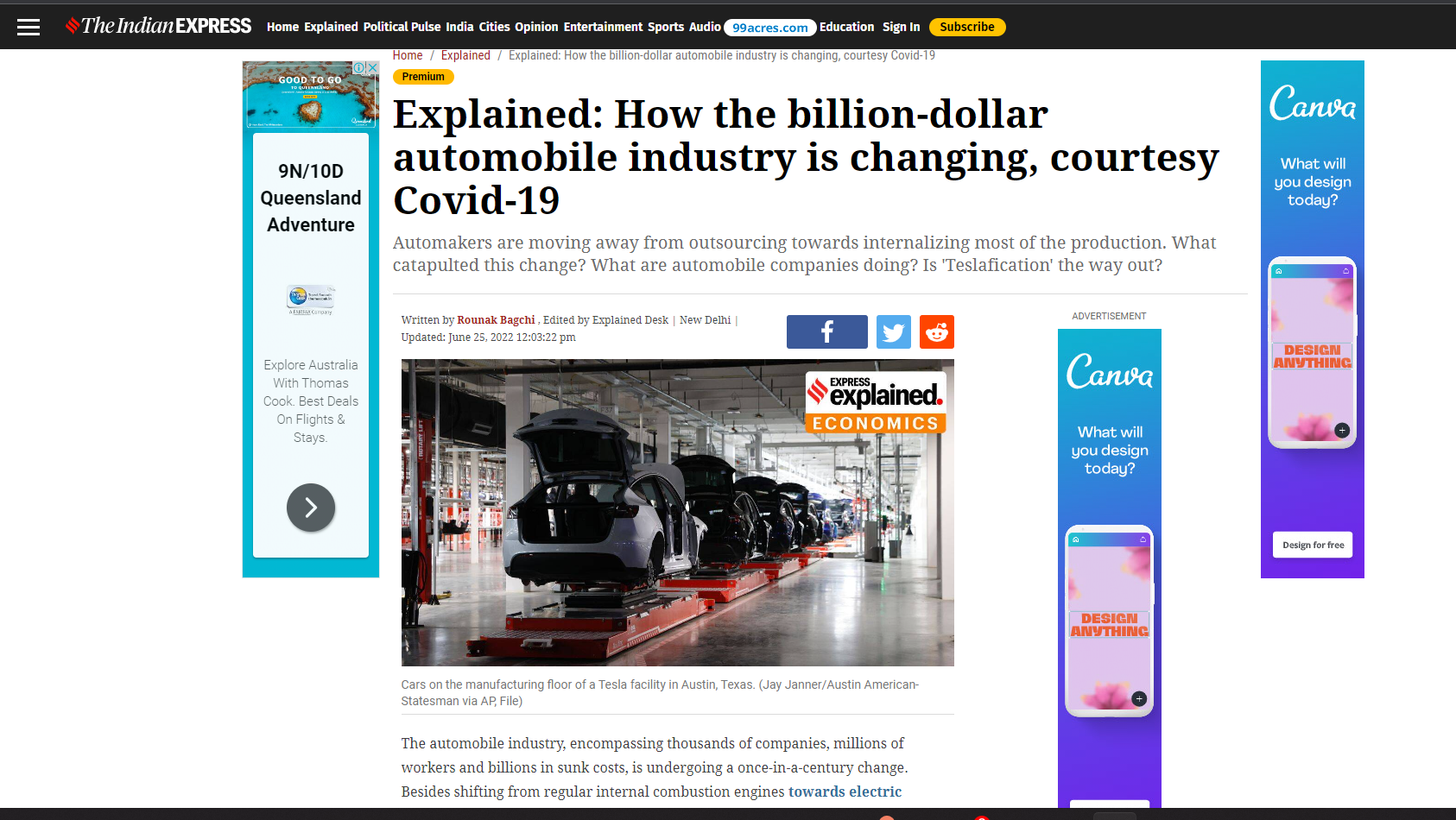


**Chapter 6**

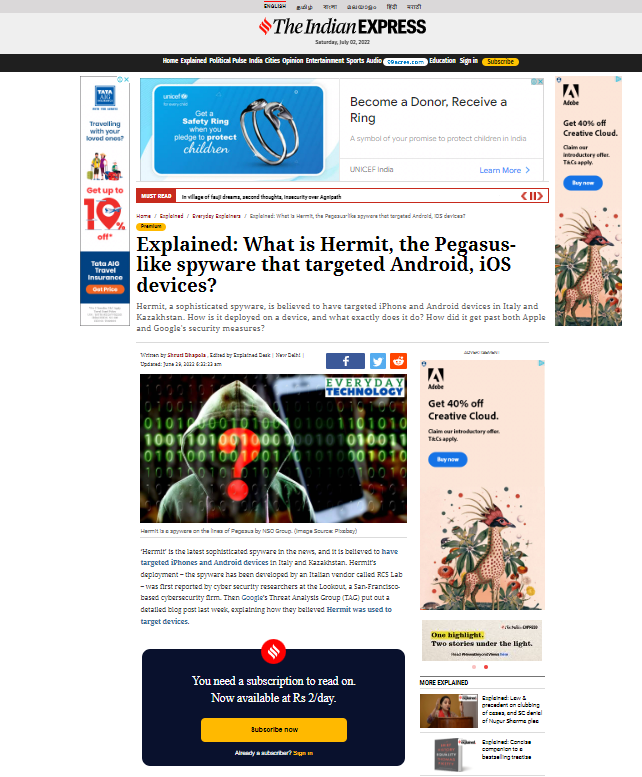
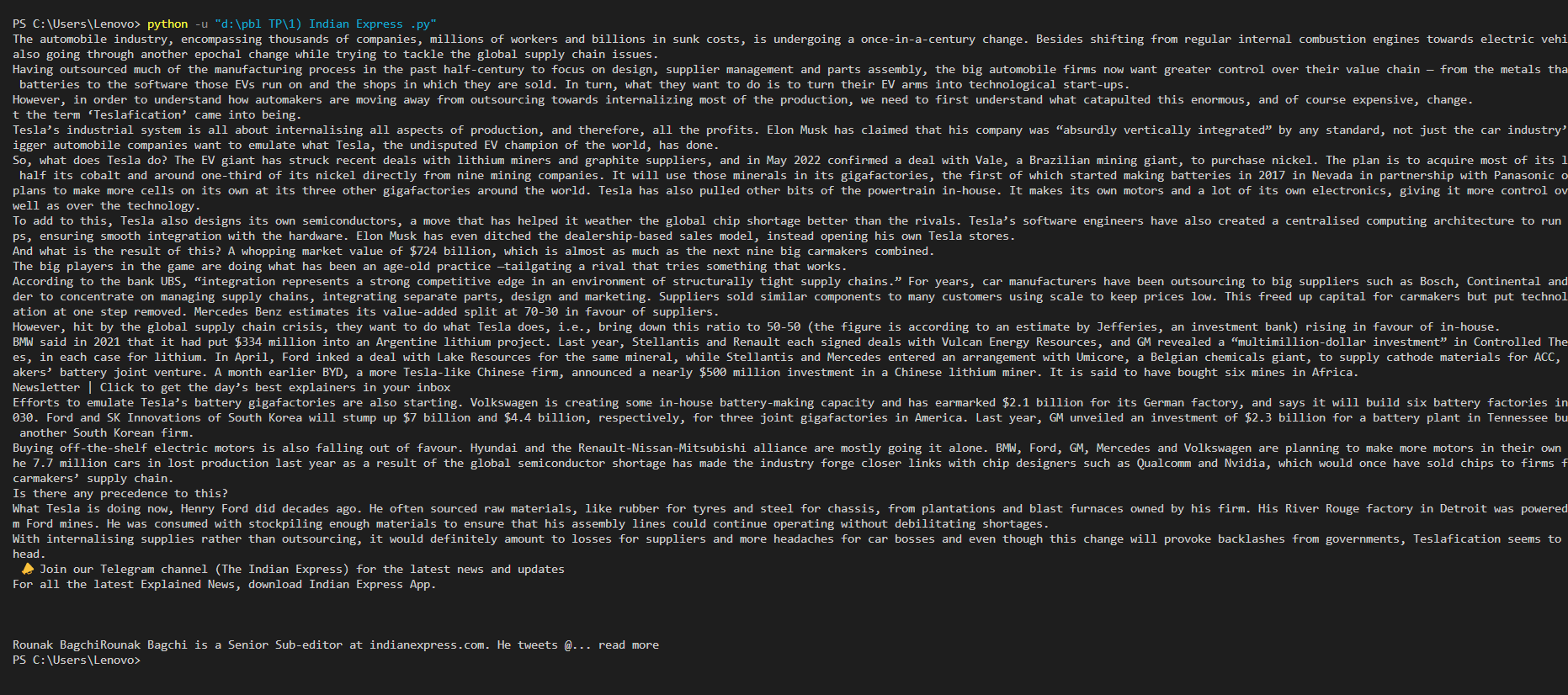
**Result / Applications**

1. **News article Scraping from a news website (Indian Express):**

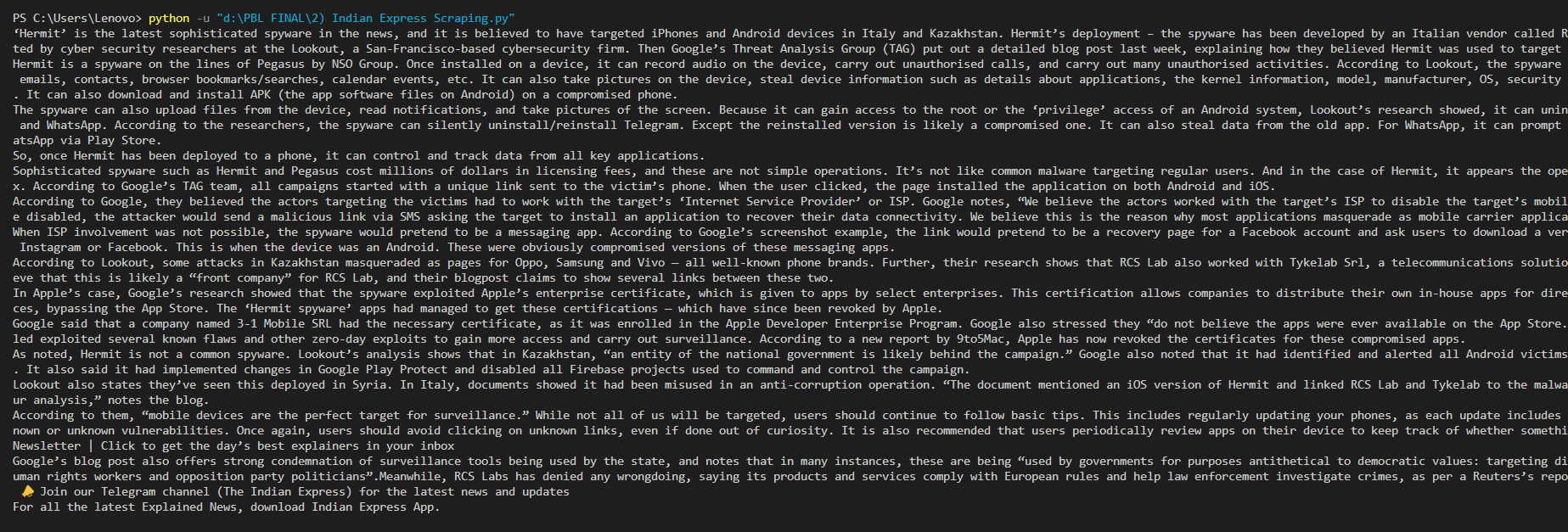
When we want to read a news article on the website and want to collect useful information from it. Without doing it manually or without doing the copy and paste process again and again this application is really helpful for scraping the information. Also, this application gives news articles without taking any subscription to the news website.



The above images show news articles on the Website with lots of ads and pop-ups. This can be overcome by web scraping.



Also sometimes required to take a subscription to a news website. This problem is also solved by web scraping.



1. **To get all videos, audios and images links from a particular website:**

While surfing on a website, some videos, audios and images on a website are required for a different purpose. These video links, audio links, and image links can be scraped or harvested by the web scraping technique.

1. Images Links:

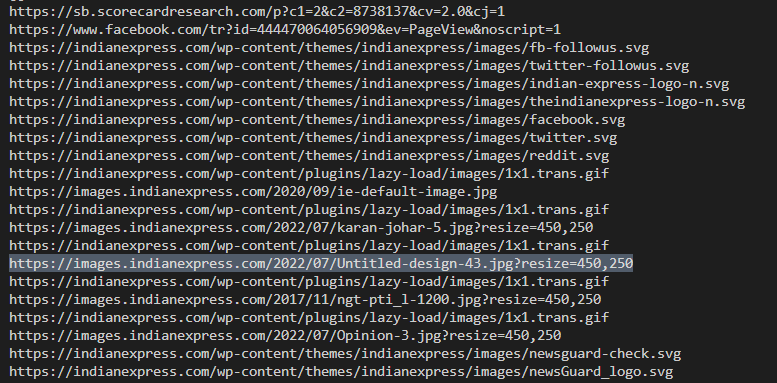
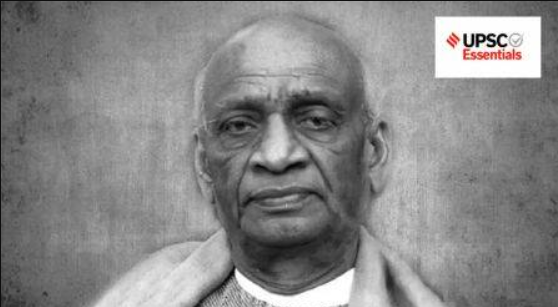
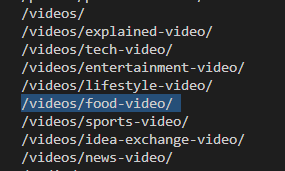


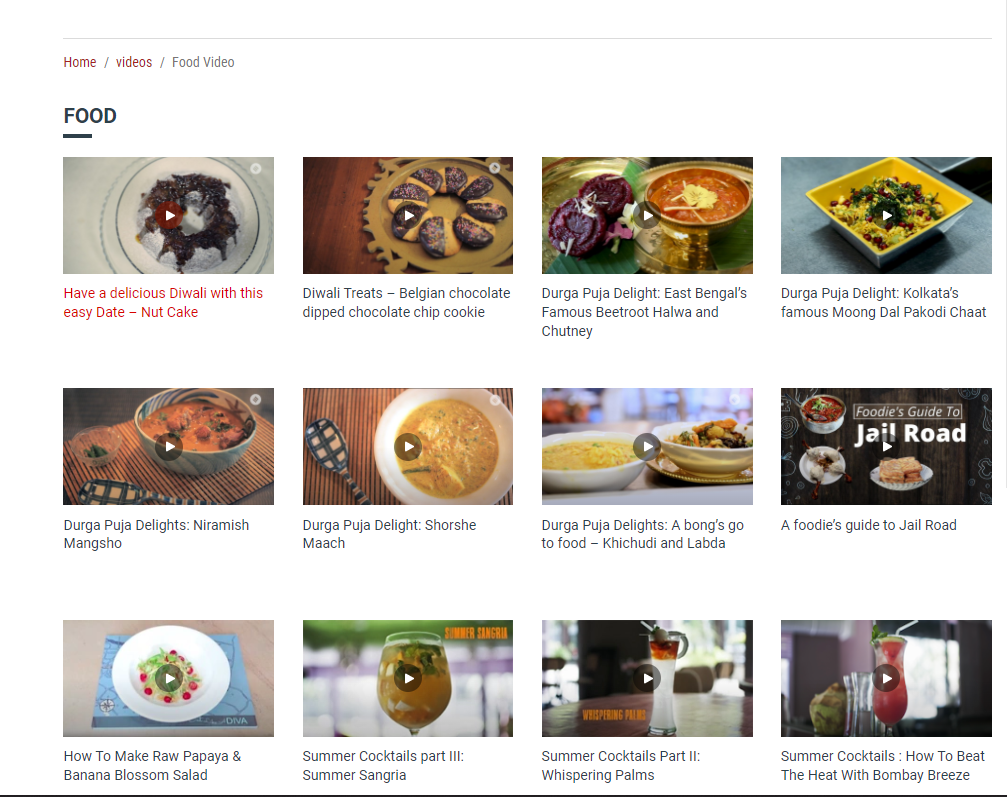
Image of above highlighted link available on website:



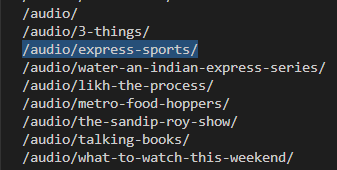
1. Videos Links:



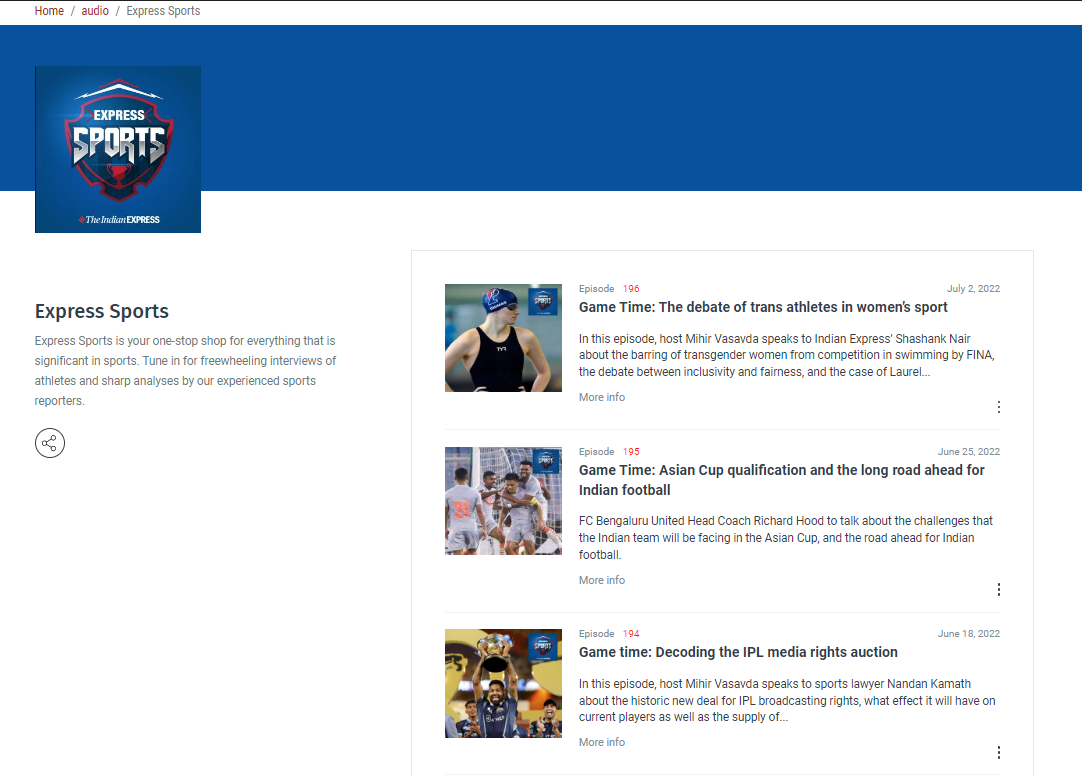
Videos stored in highlighted link available on website:



1. Audios Links:



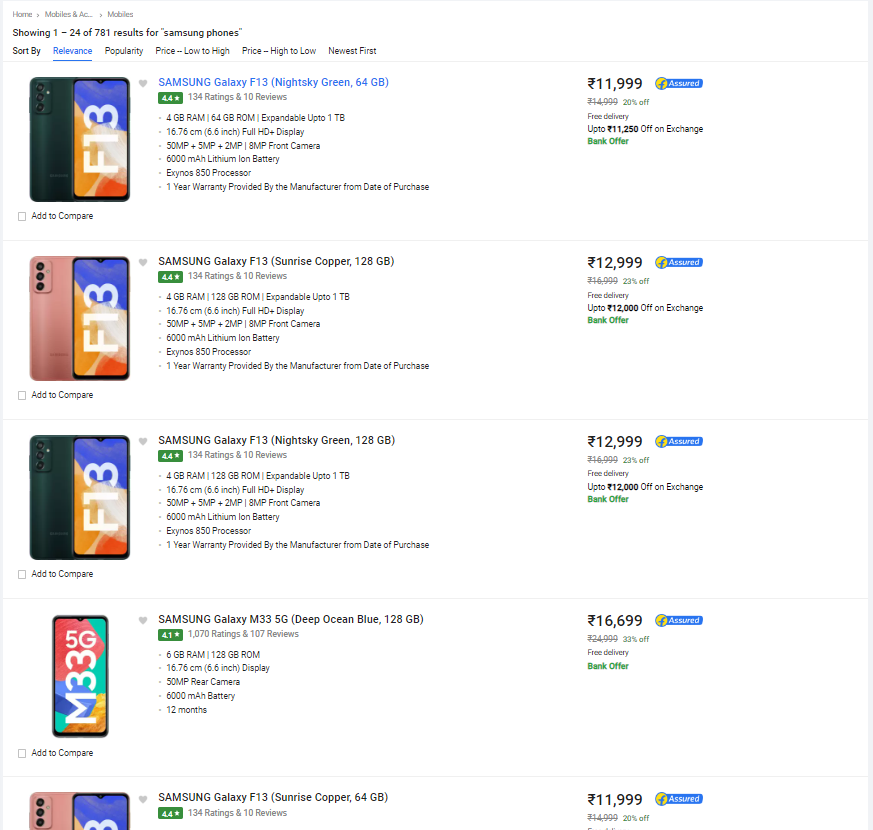
Audios stored in highlighted link available on website:



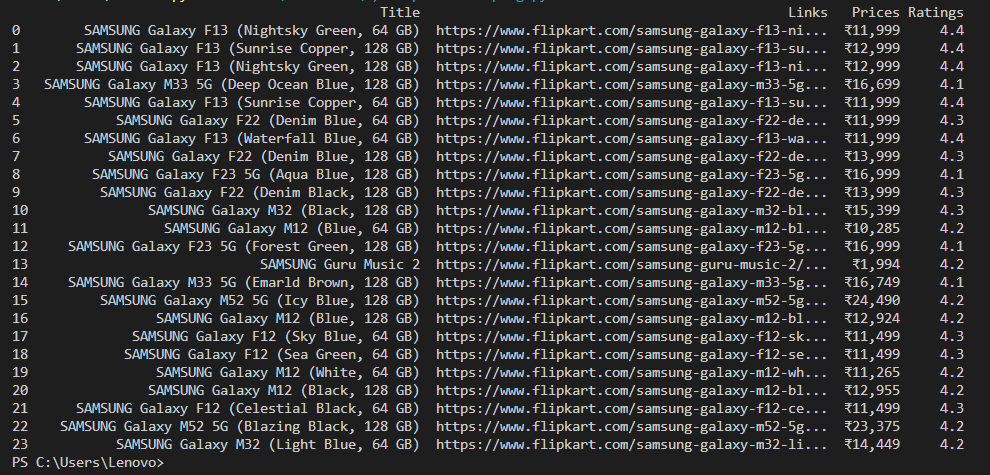
1. **To scrap information about the name, price, ratings, and links of particular products.:**

Now, most people are buying products online without going to any particular market physically. They want to save their time and money in traveling so they try to purchase the products from home.

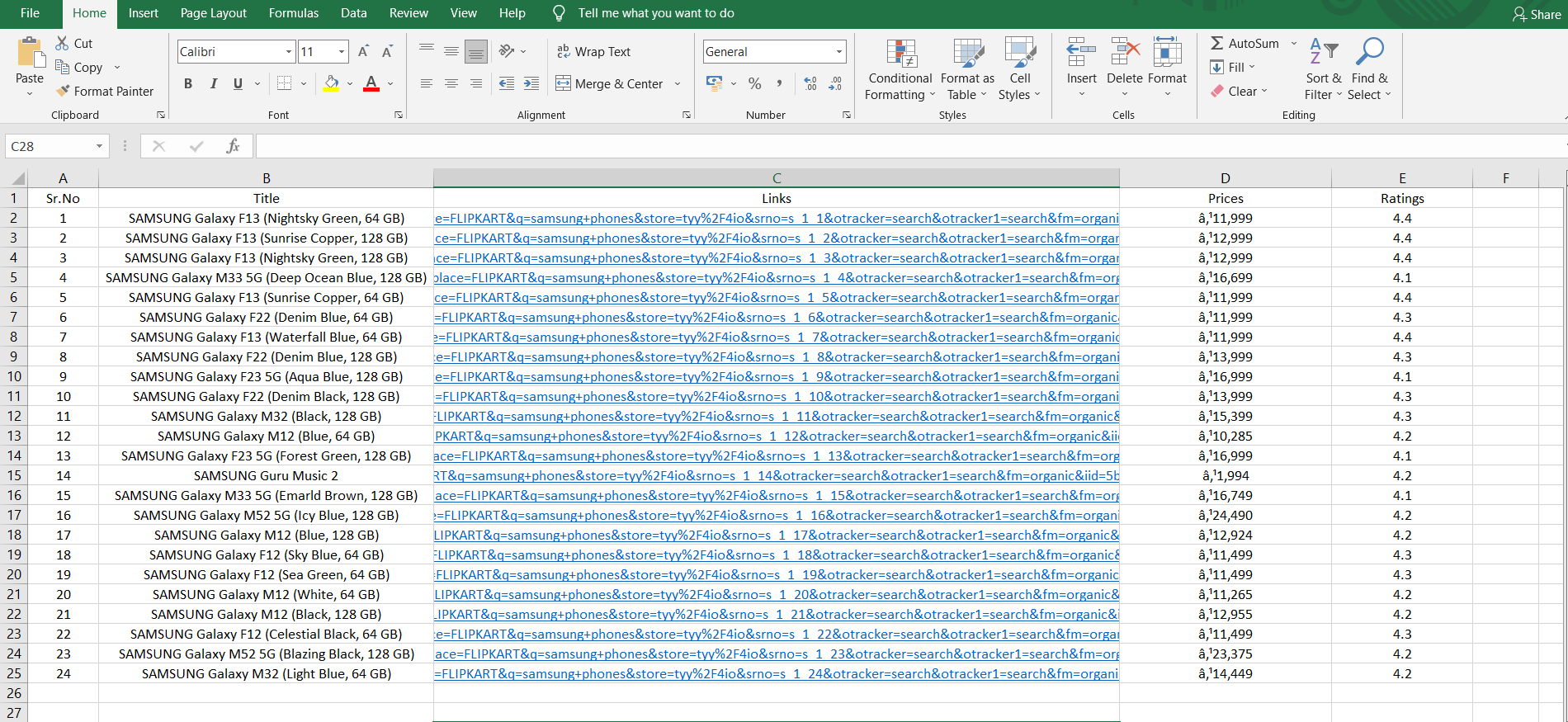
But while searching the products online on Flipkart, Amazon, or any e-commerce website, a lot of items related to products show on-page. Instead of going and briefing every item, names, prices, ratings, and links will be scraped by the web scraping technique.



Output shown in terminal:



Output saves in CSV file:



Similarly, data of various electronics products can be extracted like laptops and mobiles.

1. **Automated price comparison of products using web scraping technique.**

In process of comparing the prices of a product on various e-commerce websites like Amazon, Flipkart, and Croma manually and frequently, a lot of time, as well as energy, is consumed. in order to do it in an efficient way, Automated web scraping can be performed using the selenium library in python.

1. On Amazon:
2. On Flipkart:
3. On Croma:

**Output displayed in terminal:**



**Chapter 7**

**CONCLUSION**

Extracting data through scraping technology is a new evolving activity in the technology harvesting arena. Though many companies are still using manual process of extracting data but Web Scraping solutions will transform the traditional method of extracting data. The day is not that far with exponential growth throughout this field when it can become a phenomenon and most companies will understand the value of scraping innovation and how it enables them remain ahead in the race dramatically. This paper presents the survey of Web scraping technology incorporating what it is, how it works, the popular tools and technologies of web scraping, the websites used for this technology and the top most fields which are making use of this technology.

So far, we have explored many tools and methods for web scraping using python. now it is time to implement what we have learned.

we can apply our knowledge for different purposes, 1st we have a program that will help you read article from any news website or copy the article without any annoying advertisement or subscription remainder.

2nd is the program that will help you extract electronic product details from an E-commerce website (Flipkart) and save it in csv file format that will help you decide the product that you want to buy

and last, 3rd one is based on selenium library that is a illustration of automated web scarping, this program help you to get prices of a same product from different websites so you don’t have to visit several websites for same product and will help you to save a lot of time.

In this project we have demonstrated just a bit of web scraping,

and so let you know that we haven’t even scratched the surface of web scarping is just a first step so that you will get data and second one is to use it to predict and analyze it using machine learning and neural network.

**REFERENCES**

1. Bo Zhao, Web Scraping, College of Earth, Ocean and Atmospheric Sciences, Oregon State University, Corvallis, USA
2. https://realpython.com/python-web-scraping-practical-introduction/
3. https://www.geeksforgeeks.org/python-web-scraping-tutorial/
4. Oxylabs - YouTube channel – What is web scraping and its uses. Detailed explanation.
5. Scraping robot website – 11 web scraping Idea.
6. All the photos are taken from google.