**Batch: H3 - 3 Roll No.: 16014022050**

**Experiment No. 9**

|  |
| --- |
| **Title :** To implement data scraping using Selenium/ ScraPy |

**Aim:** To perform testing using web scrapers in Python

**Expected Outcome of Experiment:**

CO2: Application of Exploratory data analysis (EDA) on Real world problems.

**Books/ Journals/ Websites referred:**

1. https://realpython.com/python-web-scraping-practical-introduction/

**Web Scrapping:**

**Web scraping** is the process of collecting and parsing raw data from the Web.

Web scraping is an automated method used to extract large amounts of data from websites. The data on the websites are unstructured.

Web scraping helps collect these unstructured data and store it in a structured form. There are different ways to scrape websites such as online Services, APIs or writing your own code.

**Why is Web Scraping Used?**

• Price Comparison: Services such as ParseHub use web scraping to collect data from online shopping websites and use it to compare the prices of products.

• Email address gathering: Many companies that use email as a medium for marketing, use web scraping to collect email ID and then send bulk emails.

• Social Media Scraping: Web scraping is used to collect data from Social Media websites such as Twitter to find out what’s trending.

• Research and Development: Web scraping is used to collect a large set of data (Statistics, General Information, Temperature, etc.) from websites, which are analyzed and used to carry out Surveys or for R&D.

• Job listings: Details regarding job openings, interviews are collected from different websites and then listed in one place so that it is easily accessible to the user.

**Why is Python Good for Web Scraping?**

• Here is the list of features of Python which makes it more suitable for web scraping.

• Ease of Use: Python is simple to code. You do not have to add semi-colons “;” or curly-braces “{}” anywhere. This makes it less messy and easy to use.

• Large Collection of Libraries: Python has a huge collection of libraries such as Numpy, Matlplotlib, Pandas etc., which provides methods and services for various purposes. Hence, it is suitable for web scraping and for further manipulation of extracted data.

• Dynamically typed: In Python, you don’t have to define datatypes for variables, you can directly use the variables wherever required. This saves time and makes your job faster.

**How Do You Scrape Data From A Website?**

• When you run the code for web scraping, a request is sent to the URL that you have mentioned. As a response to the request, the server sends the data and allows you to read the HTML or XML page. The code then, parses the HTML or XML page, finds the data and extracts it.

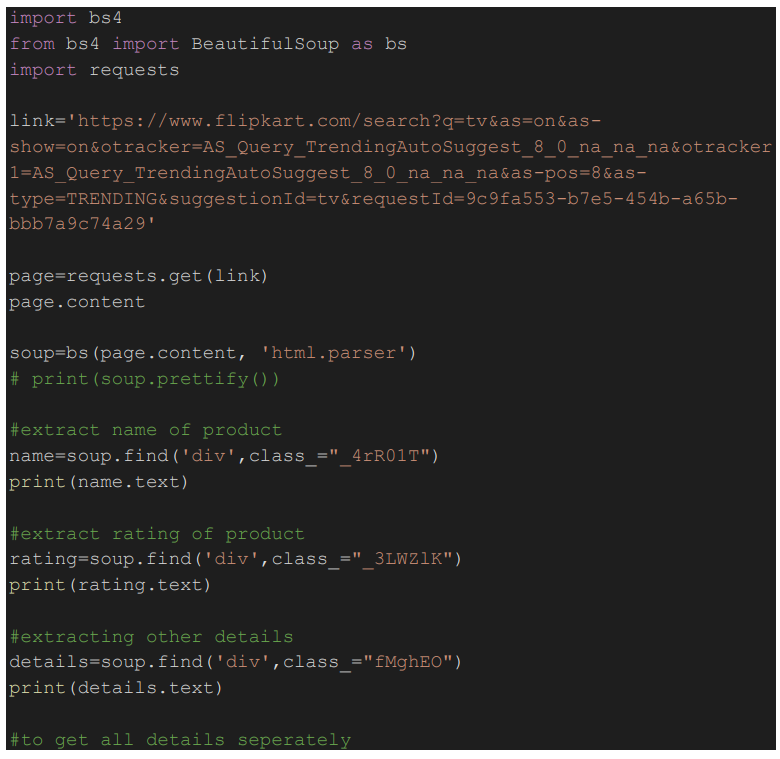
• To extract data using web scraping with python, you need to follow these basic steps:

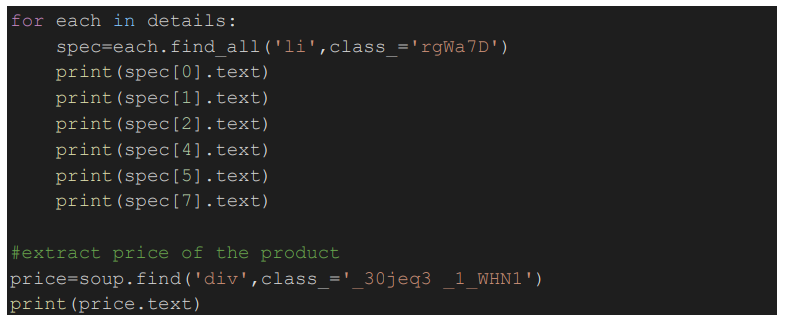
* Find the URL that you want to scrape
* Inspecting the Page
* Find the data you want to extract
* Write the code
* Run the code and extract the data
* Store the data in the required format

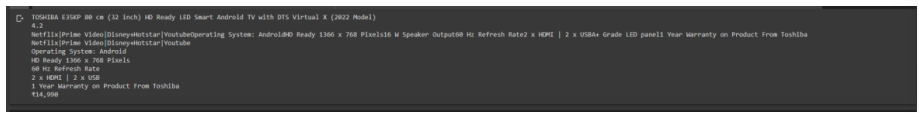
**Libraries used for Web Scraping**

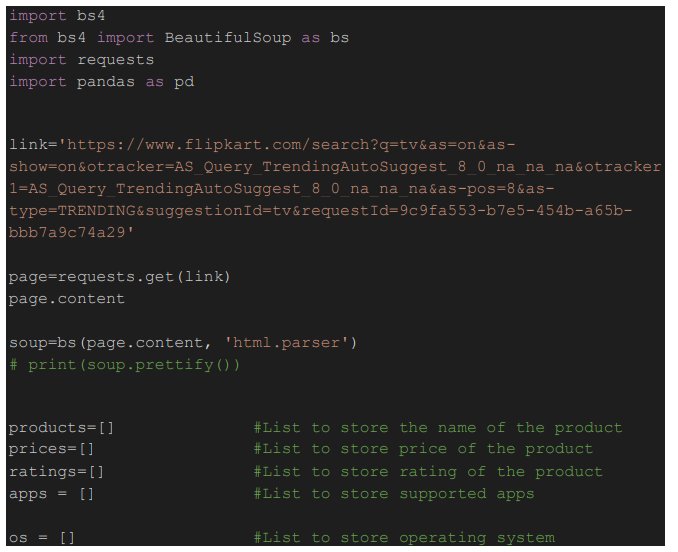
* **Selenium:** Selenium is a web testing library. It is used to automate browser activities.
* **BeautifulSoup:** Beautiful Soup is a Python package for parsing HTML and XML documents. It creates parse trees that is helpful to extract the data easily.
* **Pandas:** Pandas is a library used for data manipulation and analysis. It is used to extract the data and store it in the desired format.

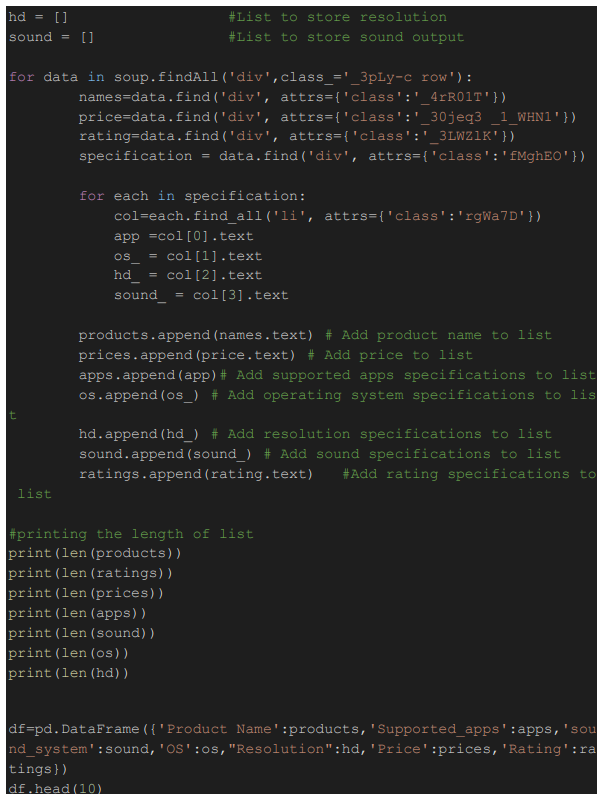
**Implementation:**

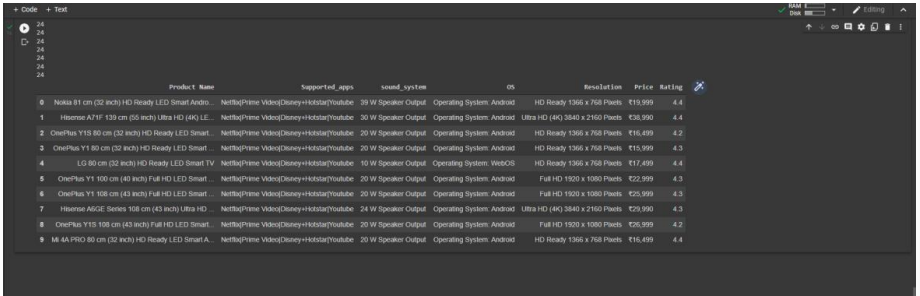






****



****

**Conclusion:**

That’s how we can scrape and store the data from the website. Here we have learned to scrape just one page, we can also perform the same on various pages and extract more data for comparison or analysis. The next step from here is to clean the data and perform analysis.

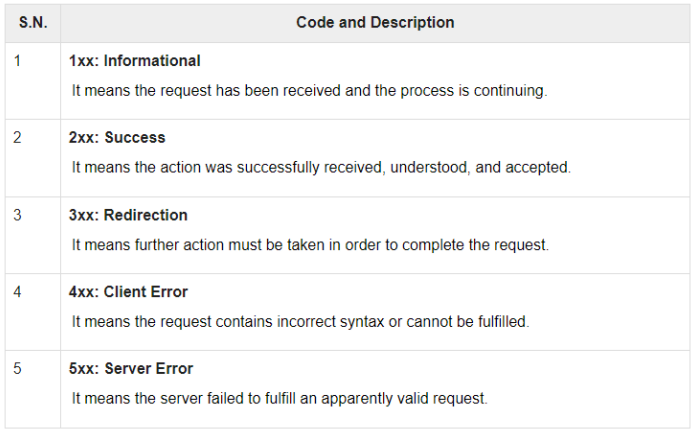
Moreover, Beautiful Soup Library is one of the easiest libraries that can be used for web scraping and is beginner-friendly but it comes with its own cons. Some of them are:

* it is not the best for scraping dynamic content i.e for the content which keeps on changing.
* it is dependent on other libraries for sending requests to a webpage.
* it does not perform the best with big web scraping projects and is a little slow compared to another library such as scrapy.

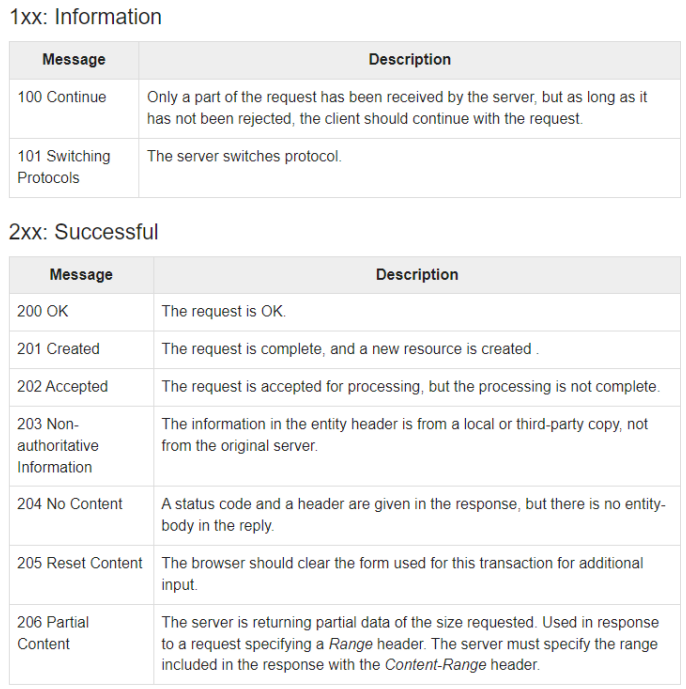
**Post lab Questions:**

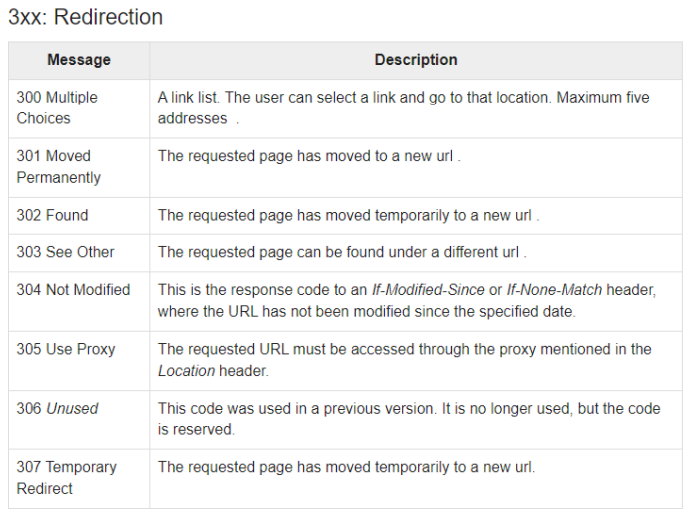
1. **What are the different HTTP response status codes?**

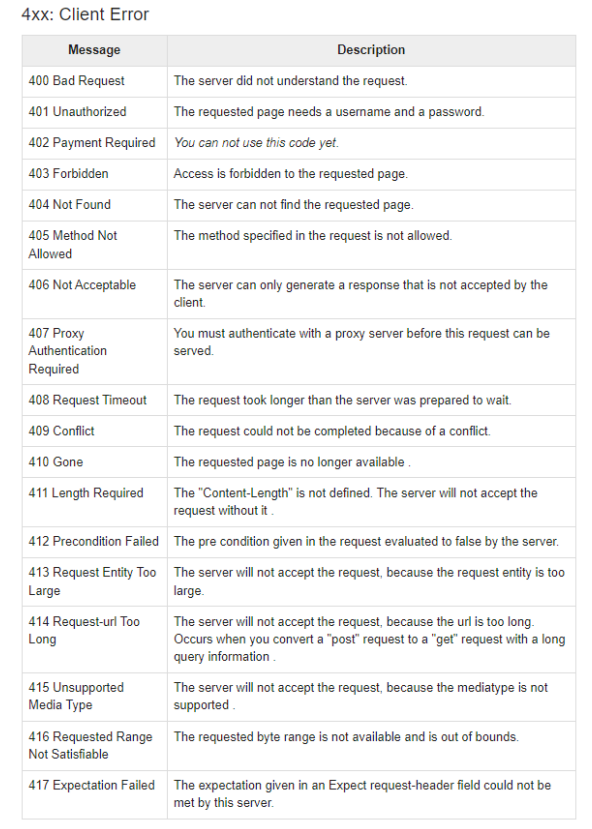
The Status-Code element in a server response, is a 3-digit integer where the first digit of the Status-Code defines the class of response and the last two digits do not have any categorization role. There are 5 values for the first digit:

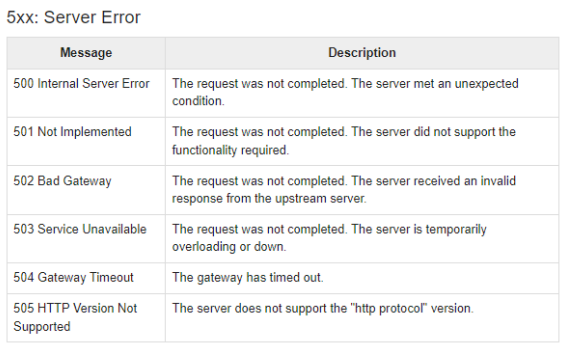
****

HTTP status codes are extensible and HTTP applications are not required to understand the meaning of all the registered status codes. Given below is a list of all the status codes.

****

****

****

****

1. **How to get the Updated Daily News using Python?**

Count number of tweets by a given Twitter account:



How to get the Updated Daily News using Python: 