Task 2: Web Scraping

**Project Objective:**

To scrape the Times of India homepage and extract the routing options, their hyperlinks, descriptions, and image URLs, saving the extracted information into an Excel sheet.

**Steps to Develop the Project:**

1. Prerequisites:

- Basic understanding of Python.

- Installation of necessary libraries: Beautiful Soup, Requests, and Pandas.

2. Setup:

- Install Beautiful Soup, Requests, and Pandas:

3. Import Libraries:

import requests

from bs4 import BeautifulSoup

import pandas as pd

4. Steps to Develop the Project:

* Step 1: Fetch the HTML Content
  + Send an HTTP request to the Times of India homepage using the requests library.

Data\_from\_website = BeautifulSoup(response.content,"lxml")

Data\_from\_website

Data\_from\_website.find\_all("a")

* Step 2: Parse the HTML Content

Identify and extract the routing options, hyperlinks, descriptions, and image URLs using Beautiful Soup.

Inspect the Times of India website using browser developer tools to identify the appropriate HTML tags and classes for extracting titles, links, descriptions, and images.

label\_name = []  
api\_link = []  
img\_url=[]  
descriptions=[]  
  
for i in Data\_from\_website.find\_all("a"):  
 label\_name.append(i.get\_text())  
 api\_link.append(i.get("href"))  
 img\_tag=i.find('img')  
 img\_scr=img\_tag.get('src') if img\_tag else None  
 img\_url.append(img\_scr)  
  
  
dis\_1=Data\_from\_website.find\_all("p")  
dis\_2=Data\_from\_website.find\_all("figcaption")  
dis\_3=Data\_from\_website.find\_all("span")  
  
for tag in dis\_1 + dis\_2 + dis\_3:  
 descriptions.append(tag.text)

* Step 3: Store Data in an Excel Sheet

Create a Pandas DataFrame and save it to an Excel file.

min\_length = min(len(label\_name), len(api\_link), len(img\_url), len(descriptions))  
  
label\_name = label\_name[:min\_length]  
api\_link = api\_link[:min\_length]  
img\_url = img\_url[:min\_length]  
descriptions = descriptions[:min\_length]  
  
df=pd.DataFrame({'Names':label\_name, "links":api\_link, "Image":img\_url, "Description":descriptions})  
df.to\_csv("times\_of\_india.csv", index=False)

**Conclusion**

This project involved web scraping the Times of India homepage to extract routing options, their hyperlinks, descriptions, and image URLs. Using Python libraries such as Beautiful Soup, Requests, and Pandas, I was able to successfully fetch, parse, and store the relevant data into an Excel sheet.