Web Scraping Daily News Updates from Economic Times

This project is done by scraping the data from the “THE ECONOMICS TIMES”, the steps used to make this project is given below:-

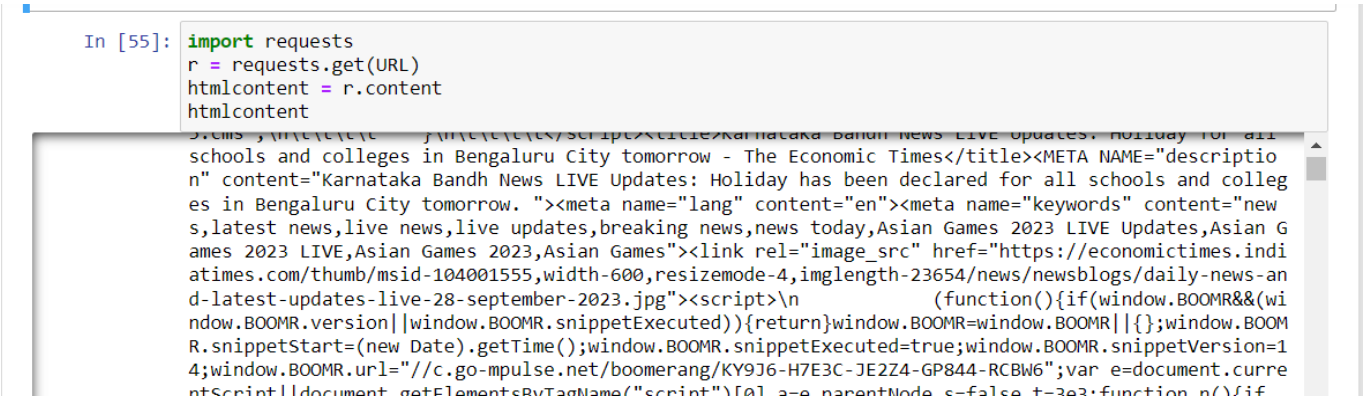
AIM:- Data extraction and analysis of news or content from a specific web page ("The Economics Times" website).

STEP1:- IMPORTING LIBRARIES

In this step, the necessary Python libraries are imported. requests is used for making HTTP requests to fetch web pages, BeautifulSoup is used to beautify which means it is used to convert the code into news headlines which can be easily readable, and pandas is used for data manipulation. Here, a variable named URL is assigned the value of the web page's URL that you want to scrape. In this case, it's a URL from the Economic Times website.

```
In [54]: import requests
         from bs4 import BeautifulSoup
         import pandas as pd
         URL ="https://economictimes.indiatimes.com/news/newsblogs/daily-news-and-latest-updates-live-28-septemb
```

STEP2 :- Now, we have successfully fetched the HTML content of the web page and stored it in the html content variable. This content can be further processed

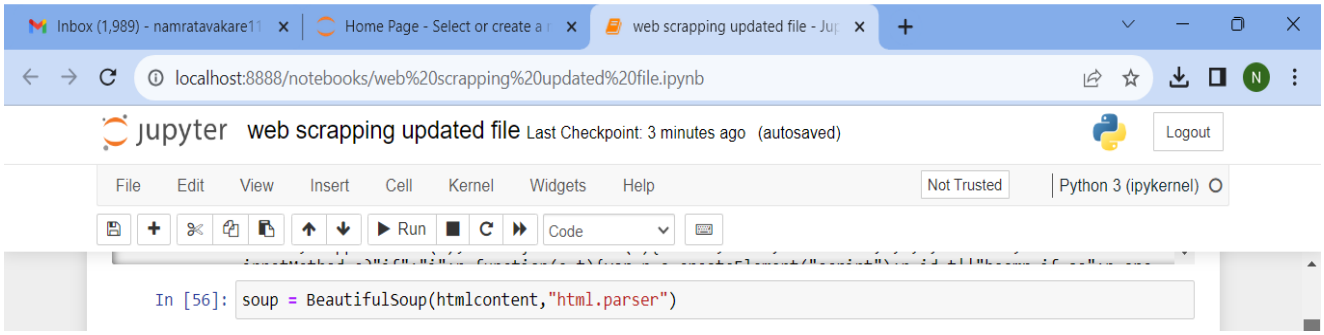


```
In [55]: import requests
         r = requests.get(URL)
         htmlcontent = r.content
         htmlcontent
```

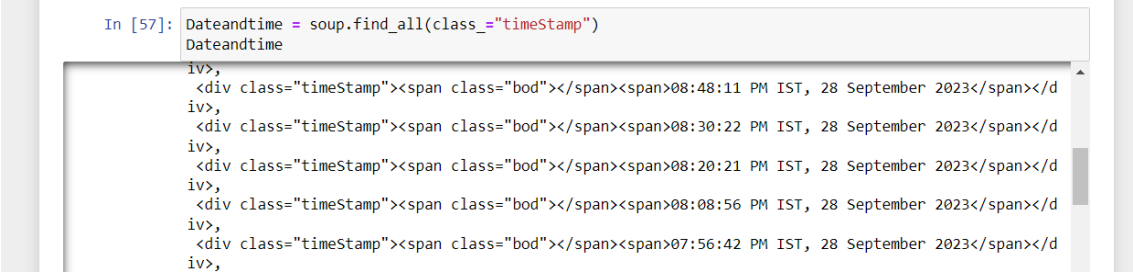
```
schools and colleges in Bengaluru City tomorrow - The Economic Times</title><META NAME="descriptio
n" content="Karnataka Bandh News LIVE Updates: Holiday has been declared for all schools and colleg
es in Bengaluru City tomorrow. "><meta name="lang" content="en"><meta name="keywords" content="new
s,latest news,live news,live updates,breaking news,news today,Asian Games 2023 LIVE Updates,Asian G
ames 2023 LIVE,Asian Games 2023,Asian Games"><link rel="image_src" href="https://economictimes.indi
atimes.com/thumb/msid-104001555,width-600,resizemode-4,imglength-23654/news/newsblogs/daily-news-an
d-latest-updates-live-28-september-2023.jpg"><script>\n            (function(){if(window.BOOMR&&(wi
ndow.BOOMR.version||window.BOOMR.snippetExecuted)){return}window.BOOMR=window.BOOMR||{};window.BOOM
R.snippetStart=(new Date).getTime();window.BOOMR.snippetExecuted=true;window.BOOMR.snippetVersion=1
4;window.BOOMR.url="//c.go-mpulse.net/boomerang/KY9J6-H7E3C-JE2Z4-GP844-RCBW6";var e=document.curre
```

and parsed using libraries like BeautifulSoup to extract specific data or information from the web page.

STEP 3 :- The code soup = BeautifulSoup(html content, "html.parser") is a crucial step in web scraping. It uses the BeautifulSoup library to parse the HTML content that we fetched earlier from the web page using the requests library.



STEP 4 :- The code Dateandtime = soup.find_all(class_="timeStamp") is used to search for and extract HTML elements that have a specific CSS class name ("timeStamp") from the soup object, which represents the parsed HTML content of the web page.



STEP 5 :- The code shown below is a Python loop that iterates through the elements stored in the Dateandtime variable, extracts the text content of each element, appends it to a list named list_Dateandtime

```
In [58]: list_Dateandtime=[]
         for i in range(0,len(Dateandtime)):
             list_Dateandtime.append(Dateandtime[i].get_text())
         print(list_Dateandtime)
         len(list_Dateandtime)
```

```
['12:49:16 AM IST, 29 September 2023', '12:03:11 AM IST, 29 September 2023', '11:42:36 PM IST, 28 Sep
tember 2023', '11:24:37 PM IST, 28 September 2023', '11:08:31 PM IST, 28 September 2023', '10:52:58 P
M IST, 28 September 2023', '10:37:31 PM IST, 28 September 2023', '10:13:10 PM IST, 28 September 202
3', '09:54:37 PM IST, 28 September 2023', '09:40:41 PM IST, 28 September 2023', '09:21:23 PM IST, 28
September 2023', '09:00:55 PM IST, 28 September 2023', '08:48:11 PM IST, 28 September 2023', '08:30:2
2 PM IST, 28 September 2023', '08:20:21 PM IST, 28 September 2023', '08:08:56 PM IST, 28 September 20
23', '07:56:42 PM IST, 28 September 2023', '07:45:15 PM IST, 28 September 2023', '07:42:04 PM IST, 28
September 2023', '07:27:24 PM IST, 28 September 2023', '07:17:44 PM IST, 28 September 2023', '07:09:5
2 PM IST, 28 September 2023', '06:59:27 PM IST, 28 September 2023', '06:45:06 PM IST, 28 September 20
```

STEP 6 :- The code latest_news = soup.find_all(class_="quote") is used to search for and extract HTML elements that have a specific class name ("quote") from the soup object, which represents the parsed HTML content of the web page.

```
In [59]: latest_news = soup.find_all(class_="quote")
         latest_news
```

```
Out[59]: [<h3 class="quote">An open and productive conversation with Think Tanks in Washington DC this mornin
         g.  Discussed transformations underway around the world and India's growing role</h3>,
          <h3 class="quote">Good to see with US Trade Representative Ambassador Katherine Tai. Spoke about our
         expanding trade and economic relationship and its broader significance</h3>,
          <h3 class="quote">Only the President is yet to come (to Rajasthan). I respect the Vice President (Ja
         gdeep Dhankhar). He visited 5 districts yesterday. Who is coming to meet him? Local BJP leaders are c
         oming to meet him. What message are they trying to convey?..</h3>,
          <h3 class="quote">We are very interested in increasing Indian exports to Russia, and we see a real p
```

STEP 7 :- 1] list_latest_news = []: This line initializes an empty list called list_latest_news. This list will be used to store the text content of the HTML elements.

2]for i in range(0, len(latest_news)):

This is a for loop that iterates through a range of numbers from 0 to the length of the latest_news collection. len(latest_news) returns the number of elements in the latest_news collection.

3]list_latest_news.append(latest_news[i].get_text()):

Inside the loop, this line extracts the text content of each HTML element in the latest_news collection using the .get_text() method and appends it to the list_latest_news list.

4]print(list_latest_news): After the loop completes, this line prints the list_latest_news list, which now contains the extracted text content from the HTML elements. These would typically be news headlines, quotes, or other textual content.

```
In [60]: list_latest_news=[]
         for i in range(0,len(latest_news)):
             list_latest_news.append(latest_news[i].get_text())
         print(list_latest_news)
         len(list_latest_news)
```

```
['An open and productive conversation with Think Tanks in Washington DC this morning.  Discussed tran
sformations underway around the world and India’s growing role', 'Good to see with US Trade Represent
ative Ambassador Katherine Tai. Spoke about our expanding trade and economic relationship and its bro
ader significance', 'Only the President is yet to come (to Rajasthan). I respect the Vice President
(Jagdeep Dhankhar). He visited 5 districts yesterday. Who is coming to meet him? Local BJP leaders ar
e coming to meet him. What message are they trying to convey?..', 'We are very interested in increasi
```

STEP 8 :- The code shown below is a loop that iterates through the elements in the list_Dateandtime list and appends them to a new list called list_Datetime. However, it includes a condition to ensure that the length of list_Datetime remains less than the length of list_latest_news.

```
In [61]: list_Datetime=[]
         for i in list_Dateandtime:
             if len(list_Datetime)<len(list_latest_news):
                 list_Datetime.append(i)
         print(list_Datetime)
```

```
['12:49:16 AM IST, 29 September 2023', '12:03:11 AM IST, 29 September 2023', '11:42:36 PM IST, 28 Sep
tember 2023', '11:24:37 PM IST, 28 September 2023', '11:08:31 PM IST, 28 September 2023', '10:52:58 P
M IST, 28 September 2023', '10:37:31 PM IST, 28 September 2023', '10:13:10 PM IST, 28 September 202
3', '09:54:37 PM IST, 28 September 2023', '09:40:41 PM IST, 28 September 2023', '09:21:23 PM IST, 28
September 2023', '09:00:55 PM IST, 28 September 2023', '08:48:11 PM IST, 28 September 2023', '08:30:2
2 PM IST, 28 September 2023', '08:20:21 PM IST, 28 September 2023', '08:08:56 PM IST, 28 September 20
```

STEP 9 :- The code shown below is designed to split date and time information stored in the list_Datetime list and then flatten it into a single list called datetime.

```
In [62]: datetime=[]
         for i in list_Datetime:
             x=i.split(',')
             for i in x:
                 datetime.append(i)
         print(datetime)
```

```
['12:49:16 AM IST', ' 29 September 2023', '12:03:11 AM IST', ' 29 September 2023', '11:42:36 PM IST',
' 28 September 2023', '11:24:37 PM IST', ' 28 September 2023', '11:08:31 PM IST', ' 28 September 202
3', '10:52:58 PM IST', ' 28 September 2023', '10:37:31 PM IST', ' 28 September 2023', '10:13:10 PM IS
T', ' 28 September 2023', '09:54:37 PM IST', ' 28 September 2023', '09:40:41 PM IST', ' 28 September
2023', '09:21:23 PM IST', ' 28 September 2023', '09:00:55 PM IST', ' 28 September 2023', '08:48:11 PM
IST', ' 28 September 2023', '08:30:22 PM IST', ' 28 September 2023', '08:20:21 PM IST', ' 28 Septembe
```

STEP 10 :- Next step is meant to separate the flattened list of date and time components from the datetime list into two separate lists: date and time. It does this by iterating through the datetime list and using the modulo operator (%) to determine whether an element at a given index is a date or a time component.

```
In [63]: date=[]
         time=[]
         for i in range(0,len(datetime)):
             if i%2==0:
                 time.append(datetime[i])
             else:
                 date.append(datetime[i])
         print(date)
         print(time)
```

```
[' 29 September 2023', ' 29 September 2023', ' 28 September 2023', ' 28 September 2023', ' 28 Septemb
er 2023', ' 28 September 2023', ' 28 September 2023', ' 28 September 2023', ' 28 September 2023', ' 2
8 September 2023', ' 28 September 2023', ' 28 September 2023', ' 28 September 2023', ' 28 September 2
023', ' 28 September 2023', ' 28 September 2023', ' 28 September 2023', ' 28 September 2023']
['12:49:16 AM IST', '12:03:11 AM IST', '11:42:36 PM IST', '11:24:37 PM IST', '11:08:31 PM IST', '10:5
```

STEP 11 :- The code we have done is using the pandas library to create a DataFrame (df) from the lists date, time, and list_latest_news.

```
In [64]: import pandas as pd
         df=pd.DataFrame({'Date':date,'time':time,"Latest_News":list_latest_news})
         df
```

Out[64]:

	Date	time	Latest_News
0	29 September 2023	12:49:16 AM IST	An open and productive conversation with Think...
1	29 September 2023	12:03:11 AM IST	Good to see with US Trade Representative Ambas...
2	28 September 2023	11:42:36 PM IST	Only the President is yet to come (to Rajastha...
3	28 September 2023	11:24:37 PM IST	We are very interested in increasing Indian ex...
4	28 September 2023	11:08:31 PM IST	India Energy Week 2024's biggest event will be...
5	28 September 2023	10:52:58 PM IST	Women are unsafe in Madhya Pradesh. Why is Shi...
6	28 September 2023	10:37:31 PM IST	More than six months have passed but nothing h...

STEP 12:- The code os.getcwd() is used to get the current working directory (CWD) in Python.

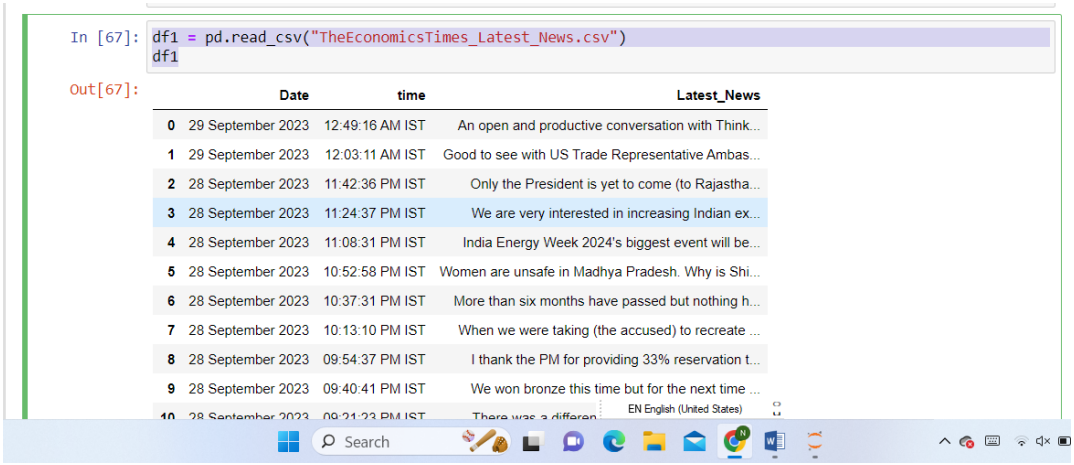
```
In [65]: import os
         os.getcwd()
```

```
Out[65]: 'C:\\Users\\USER'
```

STEP13:- The code df.to_csv("TheEconomicsTimes_Latest_News.csv", index=False) is used to save the DataFrame df as a CSV (Comma-Separated Values) file named "TheEconomicsTimes_Latest_News.csv."

```
In [66]: df.to_csv("TheEconomicsTimes_Latest_News.csv",index = False)
```

STEP 14:- The code df1 = pd.read_csv("TheEconomicsTimes_Latest_News.csv") is used to read data from a CSV (Comma-Separated Values) file named "TheEconomicsTimes_Latest_News.csv" and create a new DataFrame called df1



In [67]:
```
df1 = pd.read_csv("TheEconomicsTimes_Latest_News.csv")
df1
```

Out[67]:

	Date	time	Latest_News
0	29 September 2023	12:49:16 AM IST	An open and productive conversation with Think...
1	29 September 2023	12:03:11 AM IST	Good to see with US Trade Representative Ambas...
2	28 September 2023	11:42:36 PM IST	Only the President is yet to come (to Rajastha...
3	28 September 2023	11:24:37 PM IST	We are very interested in increasing Indian ex...
4	28 September 2023	11:08:31 PM IST	India Energy Week 2024's biggest event will be...
5	28 September 2023	10:52:58 PM IST	Women are unsafe in Madhya Pradesh. Why is Shi...
6	28 September 2023	10:37:31 PM IST	More than six months have passed but nothing h...
7	28 September 2023	10:13:10 PM IST	When we were taking (the accused) to recreate ...
8	28 September 2023	09:54:37 PM IST	I thank the PM for providing 33% reservation t...
9	28 September 2023	09:40:41 PM IST	We won bronze this time but for the next time ...
10	28 September 2023	09:21:23 PM IST	There was a differen...

CONCLUSION :- By doing the project we conclude that The data can then be easily read back into a DataFrame for further analysis or use in other data-related tasks. This type of workflow is commonly used for collecting and analyzing data from websites or external sources.

THANK YOU!