FLIP ROBO

ASSIGNMENT-1

WEB SCRAPING

In all the following questions, you have to use BeautifulSoup to scrape different websites and collect data as per the requirement of the question.

Every answer to the question should be in form of a python function which should take URL as the parameter. Use Jupyter Notebooks to program, upload it on your GitHub and send the link of the Jupyter notebook to your SME.

1) Write a python program to display all the header tags from **wikipedia.org** and make **data frame.**

```
Ans: def get_wikipedia_headers(url):
    response = requests.get(url)
    soup = BeautifulSoup(response.text, 'html.parser')

    headers = []
    for header in soup.find_all(['h1', 'h2', 'h3', 'h4', 'h5', 'h6']):
        headers.append(header.text.strip())

    return headers



wikipedia_url = 'https://en.wikipedia.org/wiki/Main_Page'



headers_list = get_wikipedia_headers(wikipedia_url)



df = pd.DataFrame({'Headers': headers_list})


print(df)
```

```
    output: Headers
0                  Main Page
1           Welcome to Wikipedia
2   From today's featured article
3              Did you know ...
4                  In the news
5                  On this day
6      From today's featured list
7       Today's featured picture
```

8 Other areas of Wikipedia

9 Wikipedia's sister projects

10 Wikipedia languages

2) Write s python program to display list of respected former presidents of India(i.e. Name , Term ofoffice) from https://presidentofindia.nic.in/former-presidents.htm and make **data frame.**

Ans:

```
def get_presidents(url):
    response = requests.get(url)
    soup = BeautifulSoup(response.text, 'html.parser')

    presidents_data = []
    for president in soup.select('.views-row'):
        name = president.select_one('.views-field-title').text.strip()
        term_of_office = president.select_one('.views-field-field-term-of-office').text.strip()
        presidents_data.append({'Name': name, 'Term of Office': term_of_office})

    return presidents_data



presidents_url = 'https://presidentofindia.nic.in/former-presidents.htm'


presidents_list = get_presidents(presidents_url)


df = pd.DataFrame(presidents_list)

    print(df)
```

output:

```
Empty DataFrame
Columns: []
    Index: []
```

3) Write a python program to scrape cricket rankings from **icc-cricket.com**. You have to scrape and make **data frame-**

a) Top **10 ODI teams** in men’s cricket along with the records for **matches, points and rating**.

Ans:

```
url = 'https://www.icc-cricket.com/rankings/mens/team-rankings/odi'


response = requests.get(url)
print(response.status_code, '--->',url)
print('\n')
soup= BeautifulSoup(response.content, 'lxml')



Team=[]
Matches=[]
```

```
Points=[]
Rating=[]
Country = soup.find_all('span',class_="u-hide-phablet")
for i in Country:
   Team.append(i.get_text().replace("\n",""))
   Team=Team[0:10]

match=soup.find_all('td',class_='rankings-block__banner--matches')
matchs=soup.find_all('td',class_='table-body__cell u-center-text')
mtc = match + matchs

for i in mtc:
   Matches.append(i.text)
   Matches=Matches[0:10]

pt=soup.find_all('td',class_="rankings-block__banner--points")

pts= soup.find_all('td',class_ ="table-body__cell u-center-text")
Point= pt + pts
for i in Point:
   Points.append(i.get_text().replace("\n",""))
   Points=Points[0:10]
rating = soup.find_all('td',class_="table-body__cell u-text-right rating")
for i in rating:
   Rating.append(i.get_text().replace("\n",""))
   Rating=Rating[0:10]

ODI=pd.DataFrame({})
ODI['Country']=Team
ODI['Matches']=Matches
ODI['Rating']=Rating
ODI['Points']=Points
print('\033[1m'+'ICC MENS ODI RANKING'+'\033[0m')
     ODI
```

```
Output: CC MENS ODI RANKING
     Country  Matches          Rating  Points
```

b) Top **10 ODI Batsmen** along with the records of their **team andrating.**

c) Top **10 ODI bowlers** along with the records of their **team andrating.**

4) Write a python program to scrape cricket rankings from icc-cricket.com. You have to scrape and make **data frame-**

a) Top **10 ODI teams** in women's cricket along with the records for **matches, points and rating**.

b) Top **10 women's ODI Batting** players along with the records of their **team and rating**.

c) Top **10 women's ODI all-rounder** along with the records of their **team and rating**.

5) Write a python program to scrape mentioned news details from https://www.cnbc.com/world/?region=world and make **data frame-**

i) Headline

ii) Time

iii) News Link

6) Write a python program to scrape the details of most downloaded articles from AI in last 90 days.https://www.journals.elsevier.com/artificial-intelligence/most-downloaded-articles

Scrape below mentioned details and make **data frame-**

i) Paper Title

ii) Authors

iii) Published Date

iv) Paper URL

7) Write a python program to scrape mentioned details from **dineout.co.in** and make **data frame-**

i) Restaurant name

ii) Cuisine

iii) Location

iv) Ratings

v) Image URL