

ASSIGNMENT-1 WEB SCRAPING

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

```
Ans. from urllib.request import urlopen
from bs4 import BeautifulSoup
html = urlopen('https://en.wikipedia.org/wiki/Main_Page')
bs = BeautifulSoup(html, "html.parser")
titles = bs.find_all(['h1', 'h2', 'h3', 'h4', 'h5', 'h6'])
print('List all the header tags :', *titles, sep='\n\n')
```

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

```
Ans. #https://presidentofindia.nic.in/former-presidents.htm
# via: https://analytics.india.gov/

import requests

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

j = requests.get(url).json()

print("(i.e. Name , Term of office)-")

print("Active Users Right Now:")

print(j['data'][0]['active_visitors'])
```

- 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.
- b) Top 10 ODI Batsmen along with the records of their team and rating.
- c) Top 10 ODI bowlers along with the records of their team and rating.

```
Ans. first_data=soup.find("div",attrs={"data-cricket-
scope":"odi"}).find("div",class_="rankings-block__top-
player").get_text(strip=True,separator=" ").split(" ")
```

```
other_data=soup.find("div",attrs={"data-cricket-scope":"odi"}).find_all("tr",class_="table-
body")
```

```
final_lst=[]
```

```
final_lst.append(first_data)
```

for i in data:

```
    split_lst=i.get_text(strip=True,separator=" ").split(" ")
```

```
    final_lst.append(split_lst)
```

- 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.

Ans.

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

Headline

iTime

News Link

Ans. import requests

```
from bs4 import BeautifulSoup
```

```
import json
```

```
def print_headlines(response_text):
```

```
    soup = BeautifulSoup(response_text, 'lxml')
```

```
    headlines = soup.find_all(attrs={"itemprop": "headline"})
```

```
    for headline in headlines:
```

```
        print(headline.text)
```

```
url = 'https://inshorts.com/en/read'
response = requests.get(url)
print_headlines(response.text)
```

- 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>

Paper Title

Authors

Published Date iv) Paper URL

- 7) Write a python program to scrape mentioned details from [dineout.co.in](https://www.dineout.co.in) and make data frame

Restaurant name

Cuisine

Location

Ratings

Image UR