

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.
- 2) Write a python program to display IMDB's Top rated 100 movies' data (i.e. name, rating, year of release) and make data frame.
- 3) Write a python program to display IMDB's Top rated 100 Indian movies' data (i.e. name, rating, year of release) and make data frame.
- 4) 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>

- 5) Write a python program to scrape cricket rankings from icc-cricket.com. You have to scrape:

- 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.- Top 10 ODI teams in men's cricket along with the records for matches, points and rating.

- 6) Write a python program to scrape cricket rankings from icc-cricket.com. You have to scrape:

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

7) Write a python program to scrape mentioned news details from

<https://www.cnbc.com/world/?region=world> :

i) Headline

ii) Time

iii) News Link

Ans.- iii) News Link

8) 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 :

i) Paper Title

ii) Authors

iii) Published Date

iv) Paper URL

Ans.- i) Paper Title

9) Write a python program to scrape mentioned details from dineout.co.in :

i) Restaurant name

ii) Cuisine

iii) Location

iv) Ratings

v) Image URL

Ans.- i) Restaurant name

10) Write a python program to scrape the details of top publications from Google Scholar from

https://scholar.google.com/citations?view_op=top_venues&hl=en

i) Rank

ii) Publication

iii) h5-index

iv) h5-median

Ans.- ii) Publication