**WEB SCRAPING**

**WORKSHEET – 5**

# In all the following questions, use Jupyter Notebooks to program, upload it on your GitHub and send the links to your SME:

1. Write a python program to display all the header tags from **‘en.wikipedia.org/wiki/Main\_Page’.**

**Program:**

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')

1. Write a python program to display IMDB’s Top rated 100 movies’ data(i.e. Name, IMDB rating, duration, director, star cast - 10 main characters)

Program:

from bs4 import BeautifulSoup

import requests

import re

# Download IMDB's Top 250 data

url = 'http://www.imdb.com/chart/top'

response = requests.get(url)

soup = BeautifulSoup(response.text, 'lxml')

movies = soup.select('td.titleColumn')

links = [a.attrs.get('href') for a in soup.select('td.titleColumn a')]

crew = [a.attrs.get('title') for a in soup.select('td.titleColumn a')]

ratings = [b.attrs.get('data-value') for b in soup.select('td.posterColumn span[name=ir]')]

votes = [b.attrs.get('data-value') for b in soup.select('td.ratingColumn strong')]

imdb = []

# Store each item into dictionary (data), then put those into a list (imdb)

for index in range(0, len(movies)):

# Seperate movie into: 'place', 'title', 'year'

movie\_string = movies[index].get\_text()

movie = (' '.join(movie\_string.split()).replace('.', ''))

movie\_title = movie[len(str(index))+1:-7]

year = re.search('\((.\*?)\)', movie\_string).group(1)

place = movie[:len(str(index))-(len(movie))]

data = {"movie\_title": movie\_title,

"year": year,

"place": place,

"star\_cast": crew[index],

"rating": ratings[index],

"vote": votes[index],

"link": links[index]}

imdb.append(data)

for item in imdb:

print(item['place'], '-', item['movie\_title'], '('+item['year']+') -', 'Starring:', item['star\_cast'])

1. Write a python program to find geospatial coordinates (i.e. latitudes and longitudes) of a city name entered by user from **‘**[**www.google.co.in/maps**](http://www.google.co.in/maps)**’.**
2. Write a python program to scrap book name, author name, genre and book review of 5 books from

# [‘www.bookpage.com](http://www.bookpage.com/)’

1. Write a python program to scrap cricket rankings from [**‘www.icc-cricket.com**](http://www.icc-cricket.com/)**’**. You have to scrap:
   1. Top 10 ODI teams.
   2. Top 10 ODI Batsmen
   3. Top 10 ODI bowlers