**# 12th March 2024**

**# CSC461 – Assignment1 – IDS – Web Scraping**

**# Mohammed Ahmed Shahbaz**

**# CIIT/LHR/FA21-BSE-050**

**# Web-Scraping and Parsing using requests and BeautifulSoup library**

**Question 1:**

Import requests

From bs4 import BeautifulSoup

Import pandas as pd

# Send a GET request to the URL

url = <https://www.imdb.com/chart/top/>

response = requests.get(url)

# Parse the HTML content

Soup = BeautifulSoup(response.content, “html.parser”)

# Find the list of top 250 movies

Movies = soup.find\_all(“td”, class\_=”titleColumn”)

# Extract data for each movie

Data = []

For movie in movies:

Title = movie.find(“a”).text

Year = movie.find(“span”, class\_=”secondaryInfo”).text.strip(“()”)

Rating = movie.find\_next\_sibling(“td”, class\_=”ratingColumn”).find(“strong”).text

# Extracting duration

Movie\_page\_url = <https://www.imdb.com> + movie.find(“a”)[“href”]

Movie\_page\_response = requests.get(movie\_page\_url)

Movie\_page\_soup = BeautifulSoup(movie\_page\_response.content, “html.parser”)

Duration = movie\_page\_soup.find(“time”).text.strip()

Data.append([title, year, duration, rating])

# Convert the data into a DataFrame

Df = pd.DataFrame(data, columns=[“Title”, “Year”, “Duration”, “IMDb Rating”])

# Save the DataFrame to a CSV file

df.to\_csv("imdb\_top\_250\_movies.csv", index=False)

**Question 2:**

Import requests

From bs4 import BeautifulSoup

Headers = {

‘User-Agent’: ‘Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/58.0.3029.110 Safari/537.36’

}

Response\_1 = requests.get(<https://space-facts.com/mars/>, headers=headers)

Root\_1 = BeautifulSoup(response\_1.content, ‘html5lib’)

Div\_req\_1 = root\_1.find(‘table’, class\_=’tablepress tablepress-id-p-mars’)

Rows = div\_req\_1.find\_all (‘tr’)

For row in rows:

Columns = row.find\_all (‘td’)

Column\_1 = columns[0]

Column\_2 = columns[1]

Print(column\_1.text.strip())

Print(column\_2.text.strip())