Forward School

## Program Code: J620-002-4:2020

## Program Name: FRONT-END SOFTWARE DEVELOPMENT

## Title : Case Study - IMDB Web Scraping

**Name: Phua Yan Han**

**IC Number: 050824070059**

**Date : 7/7/23**

**Introduction : learn to combine selenium with beautiful soup**

**Conclusion : learned how to use pandas beautiful soup and selenium**

**Reference : https://medium.com/better-programming/the-only-step-by-step-guide-youll-need-to-build-a-web-scraper-with-python-e79066bd895a (https://medium.com/better-programming/the-only-step-by-step-guide-youll-need-to-build-a-web-scraper-with-python-e79066bd895a)**

In [2]:

```
import requests
from requests import get
from bs4 import BeautifulSoup
import pandas as pd
import numpy as np
import csv
from selenium import webdriver
from selenium.common.exceptions import NoSuchElementException
```

# 1. Import Data by using webscrapping

Open the URL with headless webdriver and parse the page source into html with beautifulsoup

In [3]:

```
driver = webdriver.Chrome('C:\\Users\\Asus\\Documents\\ChromeDriver\\chromedriv

url = 'https://www.imdb.com/search/title/?groups=top_1000&ref_=adv_prvt'

driver.get(url)
```

Append data found into list according to the category

In [4]:
```
keepRunning = True
data = []
#  ranking title year rating duration genre imdb score metascore description di
while keepRunning:
    soup = BeautifulSoup(driver.page_source, "html.parser")
#     try to get the table and realize there is no table and only divs so loope
    for div in soup.find_all('div', attrs={'class': 'lister-item mode-advanced'
#         realize there is only div so have to do from section to section this 
        for headerDetails in div.find_all('h3', attrs={'class':'lister-item-hea
            for header in headerDetails:
                if header.text.rstrip() != '':
                    data.append(header.text.rstrip())

#         get the dum pg duration and genre
        if div.find('span',attrs={'class':'certificate'}) != None:
            data.append(div.find('span',attrs={'class':'certificate'}).text)
        else:
            data.append('Not Rated')
        data.append(div.find('span',attrs={'class':'runtime'}).text)
        data.append(div.find('span',attrs={'class':'genre'}).text.replace('\n',
#         did the same as above and got the imdb score and meta score kill me m
        for scoreDetails in div.find_all('div', attrs={'class':'ratings-bar'}):
            data.append(scoreDetails.find('strong').text.rstrip())
            if scoreDetails.find('div',attrs={'class':'inline-block ratings-met
                data.append(scoreDetails.find('div',attrs={'class':'inline-bloc
            else:
                data.append(np.nan)
        tempArr=[]
        for cast in div.find_all('p',attrs={'class':""}):
            if cast.text!="Director" or cast.text!="Stars":
                tempArr.append(cast.text)
        tempString=''.join(tempArr)
        tempdata=tempString.split('|')
        data.append(tempdata[0].replace('\n', '').replace('Directors:', '').rep
        data.append(tempdata[1].replace('Stars:', '').replace(' \n', '').rstri
        data.append(div.find_all('p',attrs={'class':'text-muted'})[1].text.repl
        data.append(div.find('span',attrs={'name':'nv'}).text.replace('\n', '')
        if len(div.find('p',attrs={'class':'sort-num_votes-visible'}).find_all(
            if div.find('p',attrs={'class':'sort-num_votes-visible'}).find_all(
                data.append(div.find_all('span',attrs={'name':'nv'})[1].text.re
            else:
                data.append(np.nan)
        else:
            data.append(np.nan)

    try:
        button = driver.find_element_by_class_name('lister-page-next.next-page'
        button.click()
    except NoSuchElementException:
        keepRunning = False
```

Check if the data is webscrapped successfully

```
In [5]: data
```

```
Out[5]: ['1.',
 'Spider-Man: Across the Spider-Verse',
 '(2023)',
 'PG',
 '140 min',
 'Animation, Action, Adventure',
 '8.9',
 '86',
 'Joaquim Dos Santos, Kemp Powers, Justin K. Thompson',
 '   Shameik Moore,Hailee Steinfeld,Brian Tyree Henry,Luna Lauren Velez',
 'Miles Morales catapults across the Multiverse, where he encounters a team
of Spider-People charged with protecting its very existence. When the heroe
s clash on how to handle a new threat, Miles must redefine what it means to
be a hero.',
 '169,962',
 nan,
 '2.',
 'Titanic',
 '(1997)',
 'PG-13'
```

## 2. Building a DataFrame With pandas

Put the data into data frame with Pandas

In [6]:

```
num_columns = 13
num_rows = len(data) // num_columns
data_2d = [data[i*num_columns : (i+1)*num_columns] for i in range(num_rows)]
data_2d
df = pd.DataFrame(data_2d, columns=['Rank', 'Title', 'Year', 'Rating', 'Runtime
df
```

Out[6]:

| | Rank | Title | Year | Rating | Runtime | Genre | IMDB_Score | Metascore | Director | |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1. | Spider-Man: Across the Spider-Verse | (2023) | PG | 140 min | Animation, Action, Adventure | 8.9 | 86 | Joaquim Dos Santos, Kemp Powers, Justin K. Tho... | |
| **1** | 2. | Titanic | (1997) | PG-13 | 194 min | Drama, Romance | 7.9 | 75 | James Cameron | |
| **2** | 3. | Avatar: The Way of Water | (2022) | PG-13 | 192 min | Action, Adventure, Fantasy | 7.6 | 67 | James Cameron | S |
| **3** | 4. | John Wick: Chapter 4 | (2023) | R | 169 min | Action, Crime, Thriller | 7.9 | 78 | Chad Stahelski | I<br>F |
| **4** | 5. | Indiana Jones and the Raiders of the Lost Ark | (1981) | PG | 115 min | Action, Adventure | 8.4 | 85 | Steven Spielberg | |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **995** | 996. | Vicky Donor | (2012) | Not Rated | 126 min | Comedy, Romance | 7.8 | NaN | Shoojit Sircar | |
| **996** | 997. | Vizontele | (2001) | Not Rated | 110 min | Comedy, Drama | 8.0 | NaN | Yilmaz Erdogan, Ömer Faruk Sorak | |
| **997** | 998. | Sarfarosh | (1999) | Not Rated | 174 min | Action, Drama, Thriller | 8.1 | NaN | John Mathew Matthan | |
| **998** | 999. | Airlift | (2016) | Not Rated | 130 min | Action, Drama, History | 7.9 | NaN | Raja Menon | |
| **999** | 1,000. | Anand | (1971) | Not Rated | 122 min | Drama, Musical | 8.1 | NaN | Hrishikesh Mukherjee | I |

1000 rows × 13 columns

# 3. Data Cleaning

Data cleaning - remove the '()' from year

In [7]:

```
df['Year']=df['Year'].str.replace('(', '').str.replace(')', '')
df
```

```
C:\Users\Asus\AppData\Local\Temp\ipykernel_22684\2843403066.py:1: FutureWarni
ng: The default value of regex will change from True to False in a future ver
sion. In addition, single character regular expressions will *not* be treated
as literal strings when regex=True.
  df['Year']=df['Year'].str.replace('(', '').str.replace(')', '')
```

Out[7]:

| | Rank | Title | Year | Rating | Runtime | Genre | IMDB_Score | Metascore | Director | |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1. | Spider-Man: Across the Spider-Verse | 2023 | PG | 140 min | Animation, Action, Adventure | 8.9 | 86 | Joaquim Dos Santos, Kemp Powers, Justin K. Tho... | |
| **1** | 2. | Titanic | 1997 | PG-13 | 194 min | Drama, Romance | 7.9 | 75 | James Cameron | |
| **2** | 3. | Avatar: The Way of Water | 2022 | PG-13 | 192 min | Action, Adventure, Fantasy | 7.6 | 67 | James Cameron | W<br>Sal |
| **3** | 4. | John Wick: Chapter 4 | 2023 | R | 169 min | Action, Crime, Thriller | 7.9 | 78 | Chad Stahelski | R<br>Fis |
| **4** | 5. | Indiana Jones and the Raiders of the Lost Ark | 1981 | PG | 115 min | Action, Adventure | 8.4 | 85 | Steven Spielberg | F |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **995** | 996. | Vicky Donor | 2012 | Not Rated | 126 min | Comedy, Romance | 7.8 | NaN | Shoojit Sircar | |
| **996** | 997. | Vizontele | 2001 | Not Rated | 110 min | Comedy, Drama | 8.0 | NaN | Yilmaz Erdogan, Ömer Faruk Sorak | |
| **997** | 998. | Sarfarosh | 1999 | Not Rated | 174 min | Action, Drama, Thriller | 8.1 | NaN | John Mathew Matthan | K<br>I |
| **998** | 999. | Airlift | 2016 | Not Rated | 130 min | Action, Drama, History | 7.9 | NaN | Raja Menon | |
| **999** | 1,000. | Anand | 1971 | Not Rated | 122 min | Drama, Musical | 8.1 | NaN | Hrishikesh Mukherjee | K<br>Ba |

1000 rows × 13 columns

Data cleaning - remove the min from the timemin value

In [8]:

```
df['Runtime']=df['Runtime'].str.replace(' min', '')
df
```

Out[8]:

| | Rank | Title | Year | Rating | Runtime | Genre | IMDB_Score | Metascore | Director | |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1. | Spider-Man: Across the Spider-Verse | 2023 | PG | 140 | Animation, Action, Adventure | 8.9 | 86 | Joaquim Dos Santos, Kemp Powers, Justin K. Tho... | |
| **1** | 2. | Titanic | 1997 | PG-13 | 194 | Drama, Romance | 7.9 | 75 | James Cameron | |
| **2** | 3. | Avatar: The Way of Water | 2022 | PG-13 | 192 | Action, Adventure, Fantasy | 7.6 | 67 | James Cameron | V<br>Sal |
| **3** | 4. | John Wick: Chapter 4 | 2023 | R | 169 | Action, Crime, Thriller | 7.9 | 78 | Chad Stahelski | R<br>Fis |
| **4** | 5. | Indiana Jones and the Raiders of the Lost Ark | 1981 | PG | 115 | Action, Adventure | 8.4 | 85 | Steven Spielberg | I |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **995** | 996. | Vicky Donor | 2012 | Not Rated | 126 | Comedy, Romance | 7.8 | NaN | Shoojit Sircar | |
| **996** | 997. | Vizontele | 2001 | Not Rated | 110 | Comedy, Drama | 8.0 | NaN | Yilmaz Erdogan, Ömer Faruk Sorak | |
| **997** | 998. | Sarfarosh | 1999 | Not Rated | 174 | Action, Drama, Thriller | 8.1 | NaN | John Mathew Matthan | I<br>I |
| **998** | 999. | Airlift | 2016 | Not Rated | 130 | Action, Drama, History | 7.9 | NaN | Raja Menon | |
| **999** | 1,000. | Anand | 1971 | Not Rated | 122 | Drama, Musical | 8.1 | NaN | Hrishikesh Mukherjee | I<br>Ba |

1000 rows × 13 columns

Data cleaning - remove the $ and M from the data value

In [9]:

```
df['Gross(M)']=df['Gross(M)'].str.replace('$', '').str.replace('M','')
df
```

```
C:\Users\Asus\AppData\Local\Temp\ipykernel_22684\2748443626.py:1: FutureWarni
ng: The default value of regex will change from True to False in a future ver
sion. In addition, single character regular expressions will *not* be treated
as literal strings when regex=True.
  df['Gross(M)']=df['Gross(M)'].str.replace('$', '').str.replace('M','')
```

Out[9]:

| | Rank | Title | Year | Rating | Runtime | Genre | IMDB_Score | Metascore | Director | |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1. | Spider-Man: Across the Spider-Verse | 2023 | PG | 140 | Animation, Action, Adventure | 8.9 | 86 | Joaquim Dos Santos, Kemp Powers, Justin K. Tho... | |
| **1** | 2. | Titanic | 1997 | PG-13 | 194 | Drama, Romance | 7.9 | 75 | James Cameron | |
| **2** | 3. | Avatar: The Way of Water | 2022 | PG-13 | 192 | Action, Adventure, Fantasy | 7.6 | 67 | James Cameron | W<br>Sal |
| **3** | 4. | John Wick: Chapter 4 | 2023 | R | 169 | Action, Crime, Thriller | 7.9 | 78 | Chad Stahelski | R<br>Fis |
| **4** | 5. | Indiana Jones and the Raiders of the Lost Ark | 1981 | PG | 115 | Action, Adventure | 8.4 | 85 | Steven Spielberg | F |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **995** | 996. | Vicky Donor | 2012 | Not Rated | 126 | Comedy, Romance | 7.8 | NaN | Shoojit Sircar | |
| **996** | 997. | Vizontele | 2001 | Not Rated | 110 | Comedy, Drama | 8.0 | NaN | Yilmaz Erdogan, Ömer Faruk Sorak | |
| **997** | 998. | Sarfarosh | 1999 | Not Rated | 174 | Action, Drama, Thriller | 8.1 | NaN | John Mathew Matthan | K<br>I |
| **998** | 999. | Airlift | 2016 | Not Rated | 130 | Action, Drama, History | 7.9 | NaN | Raja Menon | |
| **999** | 1,000. | Anand | 1971 | Not Rated | 122 | Drama, Musical | 8.1 | NaN | Hrishikesh Mukherjee | K<br>Ba |

1000 rows × 13 columns

Data cleaning - clear the ',' from the votes value

In [10]:

```
df['Votes']=df['Votes'].str.replace(',', '')
df
```

Out[10]:

| | Rank | Title | Year | Rating | Runtime | Genre | IMDB_Score | Metascore | Director | |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1. | Spider-Man: Across the Spider-Verse | 2023 | PG | 140 | Animation, Action, Adventure | 8.9 | 86 | Joaquim Dos Santos, Kemp Powers, Justin K. Tho... | |
| **1** | 2. | Titanic | 1997 | PG-13 | 194 | Drama, Romance | 7.9 | 75 | James Cameron | |
| **2** | 3. | Avatar: The Way of Water | 2022 | PG-13 | 192 | Action, Adventure, Fantasy | 7.6 | 67 | James Cameron | V<br>Sal |
| **3** | 4. | John Wick: Chapter 4 | 2023 | R | 169 | Action, Crime, Thriller | 7.9 | 78 | Chad Stahelski | R<br>Fis |
| **4** | 5. | Indiana Jones and the Raiders of the Lost Ark | 1981 | PG | 115 | Action, Adventure | 8.4 | 85 | Steven Spielberg | F |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **995** | 996. | Vicky Donor | 2012 | Not Rated | 126 | Comedy, Romance | 7.8 | NaN | Shoojit Sircar | |
| **996** | 997. | Vizontele | 2001 | Not Rated | 110 | Comedy, Drama | 8.0 | NaN | Yilmaz Erdogan, Ömer Faruk Sorak | |
| **997** | 998. | Sarfarosh | 1999 | Not Rated | 174 | Action, Drama, Thriller | 8.1 | NaN | John Mathew Matthan | K<br>I |
| **998** | 999. | Airlift | 2016 | Not Rated | 130 | Action, Drama, History | 7.9 | NaN | Raja Menon | |
| **999** | 1,000. | Anand | 1971 | Not Rated | 122 | Drama, Musical | 8.1 | NaN | Hrishikesh Mukherjee | K<br>Ba |

1000 rows × 13 columns

# 4. Display Cleaned and Converted Code in Pandas

In [11]:

```
df
```

Out[11]:

| | Rank | Title | Year | Rating | Runtime | Genre | IMDB_Score | Metascore | Director | |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1. | Spider-Man: Across the Spider-Verse | 2023 | PG | 140 | Animation, Action, Adventure | 8.9 | 86 | Joaquim Dos Santos, Kemp Powers, Justin K. Tho... | |
| **1** | 2. | Titanic | 1997 | PG-13 | 194 | Drama, Romance | 7.9 | 75 | James Cameron | |
| **2** | 3. | Avatar: The Way of Water | 2022 | PG-13 | 192 | Action, Adventure, Fantasy | 7.6 | 67 | James Cameron | V<br>Sal |
| **3** | 4. | John Wick: Chapter 4 | 2023 | R | 169 | Action, Crime, Thriller | 7.9 | 78 | Chad Stahelski | R<br>Fis |
| **4** | 5. | Indiana Jones and the Raiders of the Lost Ark | 1981 | PG | 115 | Action, Adventure | 8.4 | 85 | Steven Spielberg | F |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **995** | 996. | Vicky Donor | 2012 | Not Rated | 126 | Comedy, Romance | 7.8 | NaN | Shoojit Sircar | |
| **996** | 997. | Vizontele | 2001 | Not Rated | 110 | Comedy, Drama | 8.0 | NaN | Yilmaz Erdogan, Ömer Faruk Sorak | |
| **997** | 998. | Sarfarosh | 1999 | Not Rated | 174 | Action, Drama, Thriller | 8.1 | NaN | John Mathew Matthan | K<br>I |
| **998** | 999. | Airlift | 2016 | Not Rated | 130 | Action, Drama, History | 7.9 | NaN | Raja Menon | |
| **999** | 1,000. | Anand | 1971 | Not Rated | 122 | Drama, Musical | 8.1 | NaN | Hrishikesh Mukherjee | K<br>B |

1000 rows × 13 columns

## 5. Saving Your Data to a CSV

In [12]:

```
df.to_csv('C://Users/Asus/Desktop/ds1.1/Data files/IMDB_Data.csv', index=False)
```

## 6. Conclusion

What have you leanrt from this practice?

In [ ]:

```
```