Forward
School

Program Code: J620-002-4:2020

Program Name: FRONT-END SOFTWARE DEVELOPMENT

Title : Case Study - IMDB Web Scraping

Name: Ooi Caaron

IC Number: 990701-07-5837

Date : 7/7/23

Introduction : First question is the hardest part when scrapping the data from the web and put it to seperate columns

Conclusion : Still need to practice more and do revision

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 [1]:

```
import requests
from requests import get
from bs4 import BeautifulSoup
import pandas as pd
import numpy as np
from selenium import webdriver
```

1. Import Data by using webscrapping

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

In [10]:

```
import time
url = 'https://www.imdb.com/search/title/?groups=top_1000&ref_=adv_prvt'
driver = webdriver.Chrome('C://GoogleDriver//chromedriver')
driver.get(url)
soup = BeautifulSoup(driver.page_source, 'html.parser')
```

Append data found into list according to the category

In [11]:

```
data = []
for page in range(1, 21):
    soup = BeautifulSoup(driver.page_source, 'html.parser')
    lists = soup.find_all("div", class_="lister-item-content")
    for list in lists:
        row_data = {
            'Rank': '',
            'Title': '',
            'Year': '',
            'PG': '',
            'Runtime': '',
            'Genre': '',
            'IMDB_Rating': '',
            'Metascore': '',
            'Director': '',
            'Stars': '',
            'Description': '',
            'Votes': '',
            'Gross': ''
        }
        rank = list.find("span", attrs={"class":"lister-item-index"}).text.strip()
        title = list.find("a").text.strip()
        year = list.find("span", attrs={"class":"lister-item-year"}).text.strip()
        pg = list.find("span", attrs={"class":"certificate"})
        if pg is not None:
            pg = pg.text.strip()
        else:
            pg = "Not Rated"
        runtime = list.find("span", attrs={"class":"runtime"}).text.strip()
        genre = list.find("span", attrs={"class":"genre"}).text.strip()
        imdb_rating = list.find("strong").text.strip()
        metascore = list.find("span", attrs={"class":"metascore"})
        if metascore is not None:
            metascore = metascore.text.strip()
        else:
            metascore = '0'
        description = list.findAll("p", attrs={"class":"text-muted"})[1].text.strip()
        director = list.find_all("a", href=True)
        director = [tag.text for tag in director if tag.get("href") and "adv_li_dr_" in
        stars = list.find_all("a", href=True)
        stars = [tag.text for tag in stars if tag.get("href") and "adv_li_st_" in tag.ge
        votes = list.findAll("span", attrs={"name":"nv"})[0].text
        gross = list.findAll("span", attrs={"name": "nv"})
        gross = gross[1].text.strip() if len(gross) > 1 else 0
        row_data['Rank'] = rank
        row_data['Title'] = title
        row_data['Year'] = year
        row_data['PG'] = pg
        row_data['Runtime'] = runtime
        row_data['Genre'] = genre
        row_data['IMDB_Rating'] = imdb_rating
        row_data['Metascore'] = metascore
        row_data['Director'] = director
        row_data['Stars'] = stars
        row_data['Description'] = description
        row_data['Votes'] = votes
        row_data['Gross'] = gross
        data.append(row_data)
    if page == 1:
        next_button = driver.find_element_by_xpath('/html/body/div[2]/div/div[2]/div[3]/
        next_button.click()
    else:
        if page < 20:
            next_button = driver.find_element_by_xpath('/html/body/div[2]/div/div[2]/div
            next_button.click()
#     time.sleep(2)

data
```

Out[11]:

```
[{'Rank': '1.',
  'Title': 'Spider-Man: Across the Spider-Verse',
  'Year': '(2023)',
  'PG': 'PG',
  'Runtime': '140 min',
  'Genre': 'Animation, Action, Adventure',
  'IMDB_Rating': '8.9',
  'Metascore': '86',
  'Director': ['Joaquim Dos Santos', 'Kemp Powers', 'Justin K. Thompso
n'],
  'Stars': ['Shameik Moore',
   'Hailee Steinfeld',
   'Brian Tyree Henry',
   'Luna Lauren Velez'],
  'Description': 'Miles Morales catapults across the Multiverse, where
he encounters a team of Spider-People charged with protecting its very
existence. When the heroes clash on how to handle a new threat, Miles m
ust redefine what it means to be a hero.',
```

Check if the data is webscrapped successfully

In [12]:

```
data
```

Out[12]:

```
[{'Rank': '1.',
  'Title': 'Spider-Man: Across the Spider-Verse',
  'Year': '(2023)',
  'PG': 'PG',
  'Runtime': '140 min',
  'Genre': 'Animation, Action, Adventure',
  'IMDB_Rating': '8.9',
  'Metascore': '86',
  'Director': ['Joaquim Dos Santos', 'Kemp Powers', 'Justin K. Thompso
n'],
  'Stars': ['Shameik Moore',
   'Hailee Steinfeld',
   'Brian Tyree Henry',
   'Luna Lauren Velez'],
  'Description': 'Miles Morales catapults across the Multiverse, where
he encounters a team of Spider-People charged with protecting its very
existence. When the heroes clash on how to handle a new threat, Miles m
ust redefine what it means to be a hero.',
```

2. Building a DataFrame With pandas

Put the data into data frame with Pandas

In [13]:

```
df = pd.DataFrame(data, columns=['Rank', 'Title', 'Year', 'PG','Runtime','Genre','IMDB_R
                                 'Metascore','Director','Stars','Description','Votes','G
df
```

Out[13]:

	Rank	Title	Year	PG	Runtime	Genre	IMDB_Rating	Metascore	Directo
0	1.	Spider-Man: Across the Spider-Verse	(2023)	PG	140 min	Animation, Action, Adventure	8.9	86	[Joaqui Do Santo Kem Power Justin I Th
1	2.	Titanic	(1997)	PG-13	194 min	Drama, Romance	7.9	75	[Jame Camero
2	3.	Avatar: The Way of Water	(2022)	PG-13	192 min	Action, Adventure, Fantasy	7.6	67	[Jame Camero
3	4.	John Wick: Chapter 4	(2023)	R	169 min	Action, Crime, Thriller	7.9	78	[Cha Stahelsk
4	5.	Indiana Jones and the Raiders of the Lost Ark	(1981)	PG	115 min	Action, Adventure	8.4	85	[Steve Spielberg
...	...	...	...	...	...	...	...	...	
995	996.	Vicky Donor	(2012)	Not Rated	126 min	Comedy, Romance	7.8	0	[Shoo Sirca
996	997.	Vizontele	(2001)	Not Rated	110 min	Comedy, Drama	8.0	0	[Yilma Erdoga Öm Faru Sora
997	998.	Sarfarosh	(1999)	Not Rated	174 min	Action, Drama, Thriller	8.1	0	[Joh Mathe Mattha
998	999.	Airlift	(2016)	Not Rated	130 min	Action, Drama, History	7.9	0	[Ra Meno
999	1,000.	Anand	(1971)	Not Rated	122 min	Drama, Musical	8.1	0	[Hrishikes Mukherje

1000 rows × 13 columns

3. Data Cleaning

Data cleaning - remove the '()' from year

In [14]:

```
df['Year'] = df['Year'].str.replace(r"\(|\)", "", regex=True)
df
```

Out[14]:

	Rank	Title	Year	PG	Runtime	Genre	IMDB_Rating	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. Th...
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]
3	4.	John Wick: Chapter 4	2023	R	169 min	Action, Crime, Thriller	7.9	78	[Chad Stahelski]
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	0	[Shoojit Sircar]
996	997.	Vizontele	2001	Not Rated	110 min	Comedy, Drama	8.0	0	[Yilmaz Erdogan, Ömer Faruk Sorak]
997	998.	Sarfarosh	1999	Not Rated	174 min	Action, Drama, Thriller	8.1	0	[John Mathew Matthan]
998	999.	Airlift	2016	Not Rated	130 min	Action, Drama, History	7.9	0	[Raja Menon]
999	1,000.	Anand	1971	Not Rated	122 min	Drama, Musical	8.1	0	[Hrishikesh Mukherjee]

1000 rows × 13 columns

Data cleaning - remove the min from the timemin value

In [15]:

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

df
```

Out[15]:

	Rank	Title	Year	PG	Runtime	Genre	IMDB_Rating	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. Th...
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]
3	4.	John Wick: Chapter 4	2023	R	169	Action, Crime, Thriller	7.9	78	[Chad Stahelski]
4	5.	Indiana Jones and the Raiders of the Lost Ark	1981	PG	115	Action, Adventure	8.4	85	[Steven Spielberg]
...	...	...	...	...	...	...	...	...	...
995	996.	Vicky Donor	2012	Not Rated	126	Comedy, Romance	7.8	0	[Shoojit Sircar]
996	997.	Vizontele	2001	Not Rated	110	Comedy, Drama	8.0	0	[Yilmaz Erdogan, Ömer Faruk Sorak]
997	998.	Sarfarosh	1999	Not Rated	174	Action, Drama, Thriller	8.1	0	[John Mathew Matthan]
998	999.	Airlift	2016	Not Rated	130	Action, Drama, History	7.9	0	[Raja Menon]
999	1,000.	Anand	1971	Not Rated	122	Drama, Musical	8.1	0	[Hrishikesh Mukherjee]

1000 rows × 13 columns

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

In [16]:

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

Out[16]:

	Rank	Title	Year	PG	Runtime	Genre	IMDB_Rating	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. Th...
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]
3	4.	John Wick: Chapter 4	2023	R	169	Action, Crime, Thriller	7.9	78	[Chad Stahelski]
4	5.	Indiana Jones and the Raiders of the Lost Ark	1981	PG	115	Action, Adventure	8.4	85	[Steven Spielberg]
...	...	...	...	...	...	...	...	...	...
995	996.	Vicky Donor	2012	Not Rated	126	Comedy, Romance	7.8	0	[Shoojit Sircar]
996	997.	Vizontele	2001	Not Rated	110	Comedy, Drama	8.0	0	[Yilmaz Erdogan, Ömer Faruk Sorak]
997	998.	Sarfarosh	1999	Not Rated	174	Action, Drama, Thriller	8.1	0	[John Mathew Matthan]
998	999.	Airlift	2016	Not Rated	130	Action, Drama, History	7.9	0	[Raja Menon]
999	1,000.	Anand	1971	Not Rated	122	Drama, Musical	8.1	0	[Hrishikesh Mukherjee]

1000 rows × 13 columns

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

In [17]:

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

Out[17]:

	Rank	Title	Year	PG	Runtime	Genre	IMDB_Rating	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. Th...
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]
3	4.	John Wick: Chapter 4	2023	R	169	Action, Crime, Thriller	7.9	78	[Chad Stahelski]
4	5.	Indiana Jones and the Raiders of the Lost Ark	1981	PG	115	Action, Adventure	8.4	85	[Steven Spielberg]
...	...	...	...	...	...	...	...	...	...
995	996.	Vicky Donor	2012	Not Rated	126	Comedy, Romance	7.8	0	[Shoojit Sircar]
996	997.	Vizontele	2001	Not Rated	110	Comedy, Drama	8.0	0	[Yilmaz Erdogan, Ömer Faruk Sorak]
997	998.	Sarfarosh	1999	Not Rated	174	Action, Drama, Thriller	8.1	0	[John Mathew Matthan]
998	999.	Airlift	2016	Not Rated	130	Action, Drama, History	7.9	0	[Raja Menon]
999	1,000.	Anand	1971	Not Rated	122	Drama, Musical	8.1	0	[Hrishikesh Mukherjee]

```
1000 rows × 13 columns
```

4. Display Cleaned and Converted Code in Pandas

In [21]:

```
print(df)
```

```
       Rank                                          Title  Year         P
G  \
0        1.            Spider-Man: Across the Spider-Verse  2023         P
G
1        2.                                        Titanic  1997      PG-1
3
2        3.                       Avatar: The Way of Water  2022      PG-1
3
3        4.                           John Wick: Chapter 4  2023
R
4        5.  Indiana Jones and the Raiders of the Lost Ark  1981         P
G
..      ...                                            ...   ...
...
995    996.                                    Vicky Donor  2012  Not Rate
d
996    997.                                      Vizontele  2001  Not Rate
d
997    998.                                      Sarfarosh  1999  Not Rate
d
998    999.                                        Airlift  2016  Not Rate
d
999  1,000.                                          Anand  1971  Not Rate
d

    Runtime                          Genre IMDB_Rating Metascore  \
0       140  Animation, Action, Adventure         8.9        86
1       194                Drama, Romance         7.9        75
2       192    Action, Adventure, Fantasy         7.6        67
3       169       Action, Crime, Thriller         7.9        78
4       115             Action, Adventure         8.4        85
..      ...                           ...         ...       ...
995     126               Comedy, Romance         7.8         0
996     110                 Comedy, Drama         8.0         0
997     174       Action, Drama, Thriller         8.1         0
998     130        Action, Drama, History         7.9         0
999     122                Drama, Musical         8.1         0

                                              Director  \
0    [Joaquim Dos Santos, Kemp Powers, Justin K. Th...
1                                      [James Cameron]
2                                      [James Cameron]
3                                     [Chad Stahelski]
4                                   [Steven Spielberg]
..                                                 ...
995                                   [Shoojit Sircar]
996                 [Yilmaz Erdogan, Ömer Faruk Sorak]
997                              [John Mathew Matthan]
998                                      [Raja Menon]
999                             [Hrishikesh Mukherjee]

                                                 Stars  \
0    [Shameik Moore, Hailee Steinfeld, Brian Tyree ...
1    [Leonardo DiCaprio, Kate Winslet, Billy Zane, ...
2    [Sam Worthington, Zoe Saldana, Sigourney Weave...
3    [Keanu Reeves, Laurence Fishburne, George Geor...
4    [Harrison Ford, Karen Allen, Paul Freeman, Joh...
..                                                 ...
995  [Ayushmann Khurrana, Yami Gautam, Annu Kapoor,...
996  [Yilmaz Erdogan, Demet Akbag, Altan Erkekli, C...
997  [Ali Khan, Akhilendra Mishra, Makrand Deshpand...
```

```
998  [Akshay Kumar, Nimrat Kaur, Kumud Mishra, Prak...
999  [Rajesh Khanna, Amitabh Bachchan, Sumita Sanya...

                                           Description    Votes   Gross
0    Miles Morales catapults across the Multiverse,...   171148     #12
1    A seventeen-year-old aristocrat falls in love ...  1228477  659.33
2    Jake Sully lives with his newfound family form...   425970  659.68
3    John Wick uncovers a path to defeating The Hig...   232446     NaN
4    In 1936, archaeologist and adventurer Indiana ...   998751  248.16
..                                                 ...      ...     ...
995  A man is brought in by an infertility doctor t...    44440    0.17
996  Lives of residents in a small, Anatolian villa...    37770     NaN
997  After his brother is killed and father severel...    26295     NaN
998  When Iraq invades Kuwait in August 1990, a cal...    57938     NaN
999  The story of a terminally ill man who wishes t...    34528     NaN

[1000 rows x 13 columns]
```

5. Saving Your Data to a CSV

In [20]:

```
df.to_csv('imdb.csv',index=False)
```

6. Conclusion

What have you leanrt from this practice?

In [ ]:

```
# First question is the hardest part when scrapping the data from the web and put it to
```