Program Code: J620-002-4:2020

Program Name: FRONT-END SOFTWARE DEVELOPMENT

Title : Case Study - IMDB Web Scraping

Name: Justin Chong

IC Number: 960327-07-5097

Date : 7/7/2023

Introduction : Practising more with BeautifulSoup and Selenium when Web Scraping on IMDB's Top 1000 movies chart.

Conclusion : I am getting a lot better at Web Scraping than before with BeautifulSoup 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 [268]:

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

1. Import Data by using webscrapping

In [2]:

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

Append data found into list according to the category

In [304]:

```
driver = webdriver.Chrome('C:\\Users\ACER\Desktop\ChromeDriver\chromedriver
driver.get(url)

data = []
while True:
    soup = BeautifulSoup(driver.page_source,'html.parser')
    soup_list = soup.find_all("div", attrs={"class":"lister-item-content"})
    for tr in soup_list:
        rank = tr.find('span', attrs={'class':'lister-item-index unbold tex
        title = tr.find('a').text.rstrip()
        year = tr.find('span', attrs={'class':'lister-item-year text-muted 

        certificate_span = tr.find('span', attrs={'class':'certificate'})
        certificate = ''
        if certificate_span:
            certificate = certificate_span.text.rstrip()

        runtime = tr.find('span', attrs={'class':'runtime'}).text.rstrip()
        genre = tr.find('span', attrs={'class':'genre'}).text.rstrip().repl
        imdb_rating = tr.find('div', attrs={'class':'inline-block ratings-i

        metacritic_span = tr.find('div', attrs={'class':'inline-block ratin
        metacritic_score = ''
        if metacritic_span:
            metacritic_score = metacritic_span.find('span').text.rstrip()

        for p_tag in tr.find_all('p', attrs={'class': 'text-muted'}):
            synopsis = p_tag.text.rstrip().replace('\n', '')

        for crew in tr.find_all('p', attrs={'class':''}):
            crew_info = crew.text.rstrip().replace('\n', '')
        director_info, star_info = crew_info.split("|")
        directors = director_info.replace('Director:', '').replace('Directo
        stars = star_info.replace('Star:', '').replace('Stars:', '').lstrip

        votes_span = tr.find('span', string='Votes:')
        votes = ''
        if votes_span:
            votes = votes_span.find_next_sibling('span')['data-value']

        gross_span = tr.find('span', string='Gross:')
        gross = ''
        if gross_span:
            gross = gross_span.find_next_sibling('span')['data-value']

        top_chart_span = tr.find('span', attrs={'class':'text-muted top-cha
        top_chart_rank = ''
        if top_chart_span:
            top_chart_rank = top_chart_span.find_next_sibling('span')['data

        data.append((rank, title, year, certificate, runtime, genre, imdb_r

        button_div = tr.find('a', attrs={'class':'lister-page-next next-pag
        if button_div:
            print('next button exists')

    try:
        button = driver.find_element(By.LINK_TEXT, 'Next »')
        if button.is_displayed() and button.is_enabled():
            button.click()

    except NoSuchElementException:
        break

driver.quit()

data
```

Out[304]:

```
[('1.',
  'Spider-Man: Across the Spider-Verse',
  '(2023)',
  'PG',
  '140 min',
  'Animation, Action, Adventure',
  '8.9',
  '86',
  '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 must redefine wha
t it means to be a hero.',
  'Joaquim Dos Santos, Kemp Powers, Justin K. Thompson',
  'Shameik Moore, Hailee Steinfeld, Brian Tyree Henry, Luna Lauren Vele
z',
  '171845',
  '',
  '13'),
 ('2.',
  'Titanic',
```

Check if the data is webscrapped successfully

In [280]:

```
data
```

Out[280]:

```
[('1.',
  'Spider-Man: Across the Spider-Verse',
  '(2023)',
  'PG',
  '140 min',
  'Animation, Action, Adventure',
  '8.9',
  '86',
  '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 must redefine wha
t it means to be a hero.',
  'Joaquim Dos Santos, Kemp Powers, Justin K. Thompson',
  'Shameik Moore, Hailee Steinfeld, Brian Tyree Henry, Luna Lauren Vele
z',
  '171744',
  '',
  '12'),
 ('2.',
  'Titanic',
```

2. Building a DataFrame With pandas

Put the data into data frame with Pandas

In [291]:

```
imdb_columns = ['Ranking', 'Title', 'Year', 'Rating', 'Runtime', 'Genre', '
                'Directors', 'Stars', 'Votes', 'Gross', 'Top 250 Chart Rank
df = pd.DataFrame(data, columns = imdb_columns).set_index('Ranking')
df
```

Out[291]:

	Title	Year	Rating	Runtime	Genre	IMDB Rating	Metacritic Score	Synopsis	Di
Ranking									
1.	Spider-Man: Across the Spider-Verse	(2023)	PG	140 min	Animation, Action, Adventure	8.9	86	Miles Morales catapults across the Multiverse,...	J S F J
2.	Titanic	(1997)	PG-13	194 min	Drama, Romance	7.9	75	A seventeen-year-old aristocrat falls in love ...	Ca
3.	Avatar: The Way of Water	(2022)	PG-13	192 min	Action, Adventure, Fantasy	7.6	67	Jake Sully lives with his newfound family form...	Ca
4.	John Wick: Chapter 4	(2023)	R	169 min	Action, Crime, Thriller	7.9	78	John Wick uncovers a path to defeating The Hig...	St
5.	Indiana Jones and the Raiders of the Lost Ark	(1981)	PG	115 min	Action, Adventure	8.4	85	In 1936, archaeologist and adventurer Indiana ...	Sp
...	...	...	...	...	...	...	...	...	
996.	Vicky Donor	(2012)	Not Rated	126 min	Comedy, Romance	7.8		A man is brought in by an infertility doctor t...	
997.	Vizontele	(2001)		110 min	Comedy, Drama	8.0		Lives of residents in a small, Anatolian villa...	Er
998.	Sarfarosh	(1999)	Not Rated	174 min	Action, Drama, Thriller	8.1		After his brother is killed and father severel...	M M
999.	Airlift	(2016)	Not Rated	130 min	Action, Drama, History	7.9		When Iraq invades Kuwait in August 1990, a cal...	

	Title	Year	Rating	Runtime	Genre	IMDB Rating	Metacritic Score	Synopsis	Di
Ranking									
1,000.	Anand	(1971)	Not Rated	122 min	Drama, Musical	8.1		The story of a terminally ill man who wishes t...	Hri Mu

1000 rows × 13 columns

3. Data Cleaning

Data cleaning - remove the '()' from year

In [292]:

```
df['Year'] = df['Year'].str.strip('()')
df
```

Out[292]:

	Title	Year	Rating	Runtime	Genre	IMDB Rating	Metacritic Score	Synopsis	Dire
Ranking									
1.	Spider-Man: Across the Spider-Verse	2023	PG	140 min	Animation, Action, Adventure	8.9	86	Miles Morales catapults across the Multiverse,...	Joa Sa Po Jus
2.	Titanic	1997	PG-13	194 min	Drama, Romance	7.9	75	A seventeen-year-old aristocrat falls in love ...	J Car
3.	Avatar: The Way of Water	2022	PG-13	192 min	Action, Adventure, Fantasy	7.6	67	Jake Sully lives with his newfound family form...	J Car
4.	John Wick: Chapter 4	2023	R	169 min	Action, Crime, Thriller	7.9	78	John Wick uncovers a path to defeating The Hig...	Sta
5.	Indiana Jones and the Raiders of the Lost Ark	1981	PG	115 min	Action, Adventure	8.4	85	In 1936, archaeologist and adventurer Indiana ...	S Spie
...	...	...	...	...	...	...	...	...	
996.	Vicky Donor	2012	Not Rated	126 min	Comedy, Romance	7.8		A man is brought in by an infertility doctor t...	S
997.	Vizontele	2001		110 min	Comedy, Drama	8.0		Lives of residents in a small, Anatolian villa...	Y Erd
998.	Sarfarosh	1999	Not Rated	174 min	Action, Drama, Thriller	8.1		After his brother is killed and father severel...	Ma Ma
999.	Airlift	2016	Not Rated	130 min	Action, Drama, History	7.9		When Iraq invades Kuwait in August 1990, a cal...	M

	Title	Year	Rating	Runtime	Genre	IMDB Rating	Metacritic Score	Synopsis	Dire
Ranking									
1,000.	Anand	1971	Not Rated	122 min	Drama, Musical	8.1		The story of a terminally ill man who wishes t...	Hrish Mukh

1000 rows × 13 columns

Data cleaning - remove the min from the timemin value

In [298]:

```
df['Runtime'] = df['Runtime'].str.strip('min')
df = df.rename(columns={'Runtime': 'Runtime (min)'})
df
```

Out[298]:

	Title	Year	Rating	Runtime (min)	Genre	IMDB Rating	Metacritic Score	Synopsis	Dire
Ranking									
1.	Spider-Man: Across the Spider-Verse	2023	PG	140	Animation, Action, Adventure	8.9	86	Miles Morales catapults across the Multiverse,...	Joa Sa Po Jus
2.	Titanic	1997	PG-13	194	Drama, Romance	7.9	75	A seventeen-year-old aristocrat falls in love ...	J Car
3.	Avatar: The Way of Water	2022	PG-13	192	Action, Adventure, Fantasy	7.6	67	Jake Sully lives with his newfound family form...	J Car
4.	John Wick: Chapter 4	2023	R	169	Action, Crime, Thriller	7.9	78	John Wick uncovers a path to defeating The Hig...	Sta
5.	Indiana Jones and the Raiders of the Lost Ark	1981	PG	115	Action, Adventure	8.4	85	In 1936, archaeologist and adventurer Indiana ...	S Spie
...	...	...	...	...	...	...	...	...	
996.	Vicky Donor	2012	Not Rated	126	Comedy, Romance	7.8		A man is brought in by an infertility doctor t...	S
997.	Vizontele	2001		110	Comedy, Drama	8.0		Lives of residents in a small, Anatolian villa...	Y Erd
998.	Sarfarosh	1999	Not Rated	174	Action, Drama, Thriller	8.1		After his brother is killed and father severel...	Ma Ma
999.	Airlift	2016	Not Rated	130	Action, Drama, History	7.9		When Iraq invades Kuwait in August 1990, a cal...	M

	Title	Year	Rating	Runtime (min)	Genre	IMDB Rating	Metacritic Score	Synopsis	Dire
Ranking									
1,000.	Anand	1971	Not Rated	122	Drama, Musical	8.1		The story of a terminally ill man who wishes t...	Hrish Mukh

1000 rows × 13 columns

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

In [301]:

```
# for the gross that would have the $ and M in the value,
# I acquired the data-value of the span that matches the string "Gross:",
# which only has the numeric value, minus the $ and M

# the code is as follows:
# gross_span = tr.find('span', string='Gross:')
#     gross = ''
#     if gross_span:
#         gross = gross_span.find_next_sibling('span')['data-value']
df[['Title', 'Gross']]
```

Out[301]:

	Title	Gross
Ranking		
1.	Spider-Man: Across the Spider-Verse	
2.	Titanic	659,325,379
3.	Avatar: The Way of Water	659,682,302
4.	John Wick: Chapter 4	
5.	Indiana Jones and the Raiders of the Lost Ark	248,159,971
...	...	...
996.	Vicky Donor	169,209
997.	Vizontele	
998.	Sarfarosh	
999.	Airlift	
1,000.	Anand	

1000 rows × 2 columns

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

In [302]:

```
# for the votes that would have the "," in the value,
# I acquired the data-value of the span that matches the string "Votes:",
# which only has the numeric value, minus the ","

# The code is as follows:
# votes_span = tr.find('span', string='Votes:')
#     votes = ''
#     if votes_span:
#         votes = votes_span.find_next_sibling('span')['data-value']
df[['Title', 'Votes']]
```

Out[302]:

```
                                                   Title    Votes
Ranking
     1.                Spider-Man: Across the Spider-Verse   171744
     2.                                            Titanic  1228672
     3.                             Avatar: The Way of Water   426234
     4.                                 John Wick: Chapter 4   232802
     5.        Indiana Jones and the Raiders of the Lost Ark   998878
    ...                                                ...      ...
   996.                                        Vicky Donor    44441
   997.                                          Vizontele    37771
   998.                                          Sarfarosh    26297
   999.                                            Airlift    57942
 1,000.                                              Anand    34530

1000 rows × 2 columns
```

4. Display Cleaned and Converted Code in Pandas

```
In [303]: df
```

Out[303]:

	Title	Year	Rating	Runtime (min)	Genre	IMDB Rating	Metacritic Score	Synopsis	Dire
Ranking									
1.	Spider-Man: Across the Spider-Verse	2023	PG	140	Animation, Action, Adventure	8.9	86	Miles Morales catapults across the Multiverse,...	Joa Sa Po Jus
2.	Titanic	1997	PG-13	194	Drama, Romance	7.9	75	A seventeen-year-old aristocrat falls in love ...	J Car
3.	Avatar: The Way of Water	2022	PG-13	192	Action, Adventure, Fantasy	7.6	67	Jake Sully lives with his newfound family form...	J Car
4.	John Wick: Chapter 4	2023	R	169	Action, Crime, Thriller	7.9	78	John Wick uncovers a path to defeating The Hig...	Sta
5.	Indiana Jones and the Raiders of the Lost Ark	1981	PG	115	Action, Adventure	8.4	85	In 1936, archaeologist and adventurer Indiana ...	S Spie
...	...	...	...	...	...	...	...	...	
996.	Vicky Donor	2012	Not Rated	126	Comedy, Romance	7.8		A man is brought in by an infertility doctor t...	S
997.	Vizontele	2001		110	Comedy, Drama	8.0		Lives of residents in a small, Anatolian villa...	Y Erd
998.	Sarfarosh	1999	Not Rated	174	Action, Drama, Thriller	8.1		After his brother is killed and father severel...	Ma Ma
999.	Airlift	2016	Not Rated	130	Action, Drama, History	7.9		When Iraq invades Kuwait in August 1990, a cal...	M

	Title	Year	Rating	Runtime (min)	Genre	IMDB Rating	Metacritic Score	Synopsis	Dire
Ranking									
1,000.	Anand	1971	Not Rated	122	Drama, Musical	8.1		The story of a terminally ill man who wishes t...	Hrish Mukh

1000 rows × 13 columns

5. Saving Your Data to a CSV

In [297]:

```
df.to_csv('imdb_top_1000.csv')
```

6. Conclusion

What have you leanrt from this practice?

```
I find that most recent movies have been overtaking older movies in the top
1000 chart, as well as earning a higher gross compared to older movies.
```