Web Scraping IMDb

A tutorial to scrape movie information from IMDb

1: Import Necessary Libraries

In [ ]:

```
from bs4 import BeautifulSoup as bs
import requests #to access website
import pandas as pd
```

2: Load the webpage

In [ ]:

```
r = requests.get("https://www.imdb.com/chart/top/")

# Convert to a beautiful soup object
soup = bs(r.content)

# Print out HTML
contents = soup.prettify()
print(contents[:100])
```

```
<!DOCTYPE html>
<html xmlns:fb="http://www.facebook.com/2008/fbml" xmlns:og="http://ogp.me/ns#">
 <h
```

3: Creating empty list

In [ ]:

```
movie_title = []
movie_year = []
movie_rating = []
```

4: Extract HTML tag contents

In [ ]:

```
imdb_table = soup.find(class_="chart full-width")
```

In [ ]:

```
movie_titlecolumn = imdb_table.find_all(class_="titleColumn")
```

In [ ]:

```
movie_ratingscolumn = imdb_table.find_all(class_="ratingColumn imdbRating")
```

In [ ]:

```
for row in movie_titlecolumn:
    title = row.a.text # tag content extraction
    movie_title.append(title)
movie_title
```

Out[ ]:

```
['The Shawshank Redemption',
 'The Godfather',
```

```
'The Godfather: Part II',
'The Dark Knight',
'12 Angry Men',
"Schindler's List",
'The Lord of the Rings: The Return of the King',
'Pulp Fiction',
'The Good, the Bad and the Ugly',
'The Lord of the Rings: The Fellowship of the Ring',
'Fight Club',
'Forrest Gump',
'Inception',
'The Lord of the Rings: The Two Towers',
'Star Wars: Episode V - The Empire Strikes Back',
'The Matrix',
'Goodfellas',
"One Flew Over the Cuckoo's Nest",
'Seven Samurai',
'Se7en',
'The Silence of the Lambs',
'City of God',
'Life Is Beautiful',
"It's a Wonderful Life",
'Star Wars: Episode IV - A New Hope',
'Saving Private Ryan',
'Spirited Away',
'Interstellar',
'The Green Mile',
'Parasite',
'Léon: The Professional',
'Hara-Kiri',
'The Pianist',
'The Usual Suspects',
'Terminator 2: Judgment Day',
'Back to the Future',
'Psycho',
'Modern Times',
'The Lion King',
'American History X',
'City Lights',
'Grave of the Fireflies',
'Gladiator',
'Whiplash',
'The Departed',
'The Intouchables',
'The Prestige',
'Casablanca',
'Once Upon a Time in the West',
'Rear Window',
'Cinema Paradiso',
'Alien',
'Apocalypse Now',
'Memento',
'Indiana Jones and the Raiders of the Lost Ark',
'The Great Dictator',
'The Lives of Others',
'Django Unchained',
'Paths of Glory',
'Sunset Blvd.',
'WALL·E',
'The Shining',
'Avengers: Infinity War',
'Witness for the Prosecution',
'Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb',
'Spider-Man: Into the Spider-Verse',
'Joker',
'Princess Mononoke',
'Oldboy',
'Your Name.',
'Once Upon a Time in America',
'The Dark Knight Rises',
'Aliens',
'Coco',
```

```
'Hamilton',
'Capharnaüm',
'Das Boot',
'Avengers: Endgame',
'High and Low',
'American Beauty',
'Toy Story',
'3 Idiots',
'Amadeus',
'Braveheart',
'Inglourious Basterds',
'Good Will Hunting',
'Star Wars: Episode VI - Return of the Jedi',
'2001: A Space Odyssey',
'Reservoir Dogs',
'Like Stars on Earth',
'M',
'Vertigo',
'Citizen Kane',
'Pather Panchali',
'Come and See',
'The Hunt',
'Requiem for a Dream',
"Singin' in the Rain",
'North by Northwest',
'Eternal Sunshine of the Spotless Mind',
'Ikiru',
'Bicycle Thieves',
'Lawrence of Arabia',
'The Kid',
'Dangal',
'Full Metal Jacket',
'The Father',
'A Clockwork Orange',
'Metropolis',
'Taxi Driver',
'The Apartment',
'Double Indemnity',
'Incendies',
'The Sting',
'A Separation',
'1917',
'Scarface',
'Amélie',
'Snatch',
'Toy Story 3',
'To Kill a Mockingbird',
'For a Few Dollars More',
'Up',
'Indiana Jones and the Last Crusade',
'L.A. Confidential',
'Heat',
'Rashomon',
'Yojimbo',
'Ran',
'Die Hard',
'Green Book',
'Downfall',
'Monty Python and the Holy Grail',
'All About Eve',
'Some Like It Hot',
'Batman Begins',
'Unforgiven',
'Children of Heaven',
"Howl's Moving Castle",
'The Wolf of Wall Street',
'The Great Escape',
'Judgment at Nuremberg',
'Casino',
'There Will Be Blood',
'The Treasure of the Sierra Madre',
"Pan's Labyrinth",
```

```
'A Beautiful Mind',
'The Secret in Their Eyes',
'Raging Bull',
'My Neighbor Totoro',
'Chinatown',
'Lock, Stock and Two Smoking Barrels',
'The Gold Rush',
'Shutter Island',
'No Country for Old Men',
'Dial M for Murder',
'Three Billboards Outside Ebbing, Missouri',
'The Seventh Seal',
'The Elephant Man',
'The Thing',
'The Sixth Sense',
'Klaus',
'The Third Man',
'V for Vendetta',
'Wild Strawberries',
'Inside Out',
'Jurassic Park',
'The Truman Show',
'Memories of Murder',
'Blade Runner',
'Trainspotting',
'The Bridge on the River Kwai',
'Warrior',
'Fargo',
'Finding Nemo',
'My Father and My Son',
'Gone with the Wind',
'Kill Bill: Vol. 1',
'Tokyo Story',
'On the Waterfront',
'Stalker',
'Wild Tales',
'The General',
'Sherlock Jr.',
'The Deer Hunter',
'Gran Torino',
'Persona',
'The Grand Budapest Hotel',
'Before Sunrise',
'Mary and Max',
'Prisoners',
'Room',
'Mr. Smith Goes to Washington',
'In the Name of the Father',
'Catch Me If You Can',
'Gone Girl',
'Barry Lyndon',
'Hacksaw Ridge',
'To Be or Not to Be',
'Andhadhun',
'The Passion of Joan of Arc',
'Ford v Ferrari',
'The Big Lebowski',
'12 Years a Slave',
'How to Train Your Dragon',
'Mad Max: Fury Road',
'Dead Poets Society',
'Ben-Hur',
'Million Dollar Baby',
'The Wages of Fear',
'Harry Potter and the Deathly Hallows: Part 2',
'Autumn Sonata',
'Network',
'Stand by Me',
'The Handmaiden',
'Cool Hand Luke',
'The 400 Blows',
'Logan',
```

```
 "Hachi: A Dog's Tale",
 'La Haine',
 'The Bandit',
 'Platoon',
 'Gangs of Wasseypur',
 'Spotlight',
 "Monty Python's Life of Brian",
 'Hotel Rwanda',
 'A Silent Voice: The Movie',
 'Monsters, Inc.',
 'Rebecca',
 'Rush',
 'Andrei Rublev',
 'Into the Wild',
 "Love's a Bitch",
 'In the Mood for Love',
 'Rocky',
 'It Happened One Night',
 'Neon Genesis Evangelion: The End of Evangelion',
 'Nausicaä of the Valley of the Wind',
 'The Battle of Algiers',
 'Before Sunset',
 'Fanny and Alexander',
 'Three Colors: Red',
 'The Princess Bride',
 'Rififi',
 'Paris, Texas',
 'Demon Slayer: Mugen Train',
 'Nights of Cabiria',
 'Sunrise',
 'Raatchasan',
 'Hera Pheri']
```

In [ ]:

```
for row in movie_titlecolumn:
    year = row.span.text # tag content extraction
    movie_year.append(year)
movie_year
```

Out[ ]:

```
['(1994)',
 '(1972)',
 '(1974)',
 '(2008)',
 '(1957)',
 '(1993)',
 '(2003)',
 '(1994)',
 '(1966)',
 '(2001)',
 '(1999)',
 '(1994)',
 '(2010)',
 '(2002)',
 '(1980)',
 '(1999)',
 '(1990)',
 '(1975)',
 '(1954)',
 '(1995)',
 '(1991)',
 '(2002)',
 '(1997)',
 '(1946)',
 '(1977)',
 '(1998)',
 '(2001)',
 '(2014)',
 '(1999)',
 '(2019)',
```

'(1994)',
'(1962)',
'(2002)',
'(1995)',
'(1991)',
'(1985)',
'(1960)',
'(1936)',
'(1994)',
'(1998)',
'(1931)',
'(1988)',
'(2000)',
'(2014)',
'(2006)',
'(2011)',
'(2006)',
'(1942)',
'(1968)',
'(1954)',
'(1988)',
'(1979)',
'(1979)',
'(2000)',
'(1981)',
'(1940)',
'(2006)',
'(2012)',
'(1957)',
'(1950)',
'(2008)',
'(1980)',
'(2018)',
'(1957)',
'(1964)',
'(2018)',
'(2019)',
'(1997)',
'(2003)',
'(2016)',
'(1984)',
'(2012)',
'(1986)',
'(2017)',
'(2020)',
'(1981)',
'(2018)',
'(2019)',
'(1963)',
'(1999)',
'(1995)',
'(2009)',
'(1984)',
'(1995)',
'(2009)',
'(1997)',
'(1983)',
'(1968)',
'(1992)',
'(2007)',
'(1931)',
'(1955)',
'(1958)',
'(1941)',
'(1985)',
'(2012)',
'(2000)',
'(1952)',
'(1959)',
'(2004)',
'(1952)',
'(1948)',

'(1962)',
'(1921)',
'(2016)',
'(1987)',
'(2020)',
'(1971)',
'(1927)',
'(1976)',
'(1960)',
'(2010)',
'(1944)',
'(1973)',
'(2011)',
'(2019)',
'(1983)',
'(2001)',
'(2000)',
'(2010)',
'(1962)',
'(1965)',
'(2009)',
'(1989)',
'(1997)',
'(1995)',
'(1950)',
'(1961)',
'(1985)',
'(1988)',
'(2018)',
'(1975)',
'(2004)',
'(1950)',
'(1959)',
'(2005)',
'(1992)',
'(1997)',
'(2004)',
'(2013)',
'(1963)',
'(1961)',
'(1995)',
'(1948)',
'(2007)',
'(2006)',
'(2001)',
'(2009)',
'(1980)',
'(1988)',
'(1974)',
'(1998)',
'(1925)',
'(2010)',
'(2007)',
'(1954)',
'(2017)',
'(1957)',
'(1980)',
'(1982)',
'(1999)',
'(2019)',
'(1949)',
'(2005)',
'(2015)',
'(1957)',
'(1993)',
'(1998)',
'(2003)',
'(1982)',
'(1996)',
'(1957)',
'(2011)',
'(1996)',

'(2003)',
'(1939)',
'(2005)',
'(2003)',
'(1953)',
'(1954)',
'(1979)',
'(2014)',
'(1926)',
'(1924)',
'(1978)',
'(2008)',
'(2014)',
'(1966)',
'(1995)',
'(2009)',
'(2013)',
'(2015)',
'(1939)',
'(1993)',
'(2002)',
'(2014)',
'(2016)',
'(1975)',
'(1942)',
'(2018)',
'(1928)',
'(2019)',
'(1998)',
'(2013)',
'(2010)',
'(2015)',
'(1989)',
'(1959)',
'(2004)',
'(1953)',
'(2011)',
'(1978)',
'(1976)',
'(1986)',
'(2016)',
'(1967)',
'(1959)',
'(2017)',
'(2009)',
'(1995)',
'(1996)',
'(1986)',
'(2015)',
'(2012)',
'(1979)',
'(2004)',
'(2001)',
'(2016)',
'(1940)',
'(2013)',
'(2007)',
'(1966)',
'(2000)',
'(2000)',
'(1976)',
'(1934)',
'(1984)',
'(1997)',
'(1966)',
'(2004)',
'(1982)',
'(1994)',
'(1987)',
'(1955)',
'(1984)',
'(2020)',

```
 '(1927)',
 '(1957)',
 '(2019)',
 '(2000)']
```

In [ ]:

```
for row in movie_ratingscolumn:
    rating = row.strong.text # tag content extraction
    movie_rating.append(rating)
movie_rating
```

Out[ ]:

```
['9.2',
 '9.1',
 '9.0',
 '9.0',
 '8.9',
 '8.9',
 '8.9',
 '8.8',
 '8.8',
 '8.8',
 '8.8',
 '8.7',
 '8.7',
 '8.7',
 '8.7',
 '8.6',
 '8.6',
 '8.6',
 '8.6',
 '8.6',
 '8.6',
 '8.6',
 '8.6',
 '8.6',
 '8.6',
 '8.6',
 '8.5',
 '8.5',
 '8.5',
 '8.5',
 '8.5',
 '8.5',
 '8.5',
 '8.5',
 '8.5',
 '8.5',
 '8.5',
 '8.5',
 '8.5',
 '8.5',
 '8.5',
 '8.5',
 '8.5',
 '8.5',
 '8.5',
 '8.5',
 '8.5',
 '8.5',
 '8.4',
 '8.4',
 '8.4',
 '8.4',
 '8.4',
 '8.4',
 '8.4',
 '8.4',
 '8.4',
 '8.4',
 '8.4',
 '8.4'
```

```
'8.4',
'8.4',
'8.4',
'8.4',
'8.4',
'8.4',
'8.4',
'8.4',
'8.4',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.3',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
'8.2',
```

```
    '8.2',
    '8.2',
    '8.2',
    '8.2',
    '8.2',
    '8.2',
    '8.2',
    '8.2',
    '8.2',
    '8.2',
    '8.2',
    '8.2',
    '8.2',
    '8.2',
    '8.2',
    '8.2',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
    '8.1',
```

```
 '8.1',
 '8.1',
 '8.1',
 '8.1',
 '8.1',
 '8.1',
 '8.1',
 '8.1',
 '8.1',
 '8.1',
 '8.1',
 '8.1',
 '8.1',
 '8.1',
 '8.1',
 '8.1',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0',
 '8.0']
```

```
# This is formatted as code
```

5: Create Dataframe

In [ ]:

```
movie_df = pd.DataFrame({'Movie Title': movie_title, 'Year of Release': movie_year, 'IMD
b Rating': movie_rating})
movie_df.head(30) #View first 30 rows of dataframe
```

Out[ ]:

	Movie Title	Year of Release	IMDb Rating
0	The Shawshank Redemption	(1994)	9.2
1	The Godfather	(1972)	9.1
2	The Godfather: Part II	(1974)	9.0
3	The Dark Knight	(2008)	9.0
4	12 Angry Men	(1957)	8.9

	Movie Title	Year of Release	IMDb Rating
5	Schindler's List	(1993)	8.9
6	The Lord of the Rings: The Return of the King	(2003)	8.9
7	Pulp Fiction	(1994)	8.8
8	The Good, the Bad and the Ugly	(1966)	8.8
9	The Lord of the Rings: The Fellowship of the Ring	(2001)	8.8
10	Fight Club	(1999)	8.8
11	Forrest Gump	(1994)	8.7
12	Inception	(2010)	8.7
13	The Lord of the Rings: The Two Towers	(2002)	8.7
14	Star Wars: Episode V - The Empire Strikes Back	(1980)	8.7
15	The Matrix	(1999)	8.6
16	Goodfellas	(1990)	8.6
17	One Flew Over the Cuckoo's Nest	(1975)	8.6
18	Seven Samurai	(1954)	8.6
19	Se7en	(1995)	8.6
20	The Silence of the Lambs	(1991)	8.6
21	City of God	(2002)	8.6
22	Life Is Beautiful	(1997)	8.6
23	It's a Wonderful Life	(1946)	8.6
24	Star Wars: Episode IV - A New Hope	(1977)	8.6
25	Saving Private Ryan	(1998)	8.5
26	Spirited Away	(2001)	8.5
27	Interstellar	(2014)	8.5
28	The Green Mile	(1999)	8.5
29	Parasite	(2019)	8.5