Live Coding - Module 5

Rachel Holman

In [1]:
```
import numpy as np
import pandas as pd
import requests
from bs4 import BeautifulSoup
import json
```

headers

In [2]:
```
r = requests.get("https://httpbin.org/user-agent")
useragent= json.loads(r.text)['user-agent']
useragent
```

Out[2]:
```
'python-requests/2.29.0'
```

In [3]:
```
headers = {'User-Agent': useragent,
           'From': 'dnw9qk@virginia.edu'}
```

Get raw HTML from the website we are scrping

In [4]:
```
url = "https://www.rottentomatoes.com/browse/movies_in_theaters/sort:a_z?page=5

r= requests.get(url, headers=headers)
r
```

Out[4]:
```
<Response [200]>
```

In [5]:
```
mysoup = BeautifulSoup(r.text)
#mysoup
```

Movies are either in an a tag or a div tag

a tag movies

In [6]:
```
movielist = mysoup.find_all('a', 'js-tile-link')
#movielist[0].find('span', 'p--small').text.strip()
```

In [7]:
```
titles1 = [m.find('span', 'p--small').text.strip() for m in movielist]
titles1
```

Out[7]:
```
['20 Days in Mariupol',
 '7:11 PM',
 'Bad Girl Boogey',
 'Belle',
 'Bhaag Saale',
 'Blue Jean',
 'Contempt',
 'Have You Got It Yet? The Story of Syd Barrett and Pink Floyd',
 "L'immensità",
 'Lost in the Stars',
 'Lynch/Oz',
 'Millie Lies Low',
 'Neeyat',
 'ODESZA: The Last Goodbye Cinematic Experience',
 'Odd Hours, No Pay, Cool Hat',
 'Once Upon a Time in Uganda',
 'One More Chance',
 'Padmini',
 'Rangabali',
 'Rudrangi',
 'Squaring the Circle (The Story of Hipgnosis)',
 'The 50th Anniversary of Lynyrd Skynyrd',
 'The Crusades',
 'The League',
 'The Miracle Club',
 'The Mother and the Whore',
 'The Wicker Man',
 'UFC 290: Volkanovski vs. Rodríguez',
 'Werckmeister Harmonies',
 'Wham!']
```

In [8]:
```
#movielist[0].find('span', 'smaller').string.strip()
# if not open date, put empty string
def getopendate(m):
    openlist = m.find_all('span', 'smaller')
    if len(openlist) >0:
        opendate = openlist[0].string.strip()
    else:
        opendate = ''
    return opendate

opendates1 = [getopendate(m) for m in movielist]
opendates1
```

Out[8]:
```
['Opens Jul 14, 2023',
 'Opened Jul 07, 2023',
 'Opened Jul 07, 2023',
 'Opens Jul 14, 2023',
 'Opened Jul 07, 2023',
 'Opened Jun 09, 2023',
 'Opened Dec 18, 1963',
 'Opens Jul 14, 2023',
 'Opened May 12, 2023',
 'Opened Jul 07, 2023',
 'Opened Jun 02, 2023',
 'Opened Jun 30, 2023',
 'Opened Jul 07, 2023',
 'Opened Jul 07, 2023',
 'Opened Jul 07, 2023',
 'Opened Jul 04, 2023',
 'Opened Jul 07, 2023',
 'Opened Jul 07, 2023',
 'Opened Jul 07, 2023',
 'Opened Jul 07, 2023',
 'Opened Jun 07, 2023',
 'Opened Jul 08, 2023',
 'Opened Jul 07, 2023',
 'Opened Jul 07, 2023',
 'Opens Jul 14, 2023',
 'Opened Mar 25, 1974',
 'Opened Aug 07, 1974',
 'Opened Jul 08, 2023',
 '',
 'Opened Jul 05, 2023']
```

In [9]:
```
audiencescore1 =[m.find('score-pairs')['audiencescore'] for m in movielist]
audiencesentiment1 =[m.find('score-pairs')['audiencesentiment'] for m in movie
criticsscore1 =[m.find('score-pairs')['criticsscore'] for m in movielist]
criticssentiment1 =[m.find('score-pairs')['criticssentiment'] for m in movieli
```

In [10]:
```
endpoints1 = ['https://rottentomatoes.com'+ m['href'] for m in movielist]
endpoints1
```

Out[10]:
```
['https://rottentomatoes.com/m/20_days_in_mariupol',
 'https://rottentomatoes.com/m/7_11_pm',
 'https://rottentomatoes.com/m/bad_girl_boogey',
 'https://rottentomatoes.com/m/belle_2023',
 'https://rottentomatoes.com/m/bhaag_saale',
 'https://rottentomatoes.com/m/blue_jean',
 'https://rottentomatoes.com/m/contempt',
 'https://rottentomatoes.com/m/have_you_got_it_yet_the_story_of_syd_barrett_and_pink_floyd',
 'https://rottentomatoes.com/m/limmensita_2022',
 'https://rottentomatoes.com/m/lost_in_the_stars_2022',
 'https://rottentomatoes.com/m/lynch_oz',
 'https://rottentomatoes.com/m/millie_lies_low',
 'https://rottentomatoes.com/m/neeyat',
 'https://rottentomatoes.com/m/odesza_the_last_goodbye_cinematic_experience',
 'https://rottentomatoes.com/m/odd_hours_no_pay_cool_hat',
 'https://rottentomatoes.com/m/once_upon_a_time_in_uganda',
 'https://rottentomatoes.com/m/one_more_chance_2023',
 'https://rottentomatoes.com/m/padmini',
 'https://rottentomatoes.com/m/rangabali',
 'https://rottentomatoes.com/m/rudrangi',
 'https://rottentomatoes.com/m/squaring_the_circle_the_story_of_hipgnosis',
 'https://rottentomatoes.com/m/the_50th_anniversary_of_lynyrd_skynyrd',
 'https://rottentomatoes.com/m/the_crusades_2023',
 'https://rottentomatoes.com/m/the_league',
 'https://rottentomatoes.com/m/the_miracle_club',
 'https://rottentomatoes.com/m/the_mother_and_the_whore',
 'https://rottentomatoes.com/m/the_wicker_man_1973',
 'https://rottentomatoes.com/m/ufc_290_volkanovski_vs_rodriguez',
 'https://rottentomatoes.com/m/werckmeister_harmonies',
 'https://rottentomatoes.com/m/wham']
```

div tags

In [11]:
```
movielist = mysoup.find_all('div', 'js-tile-link')
titles2 = [m.find('span', 'p--small').text.strip() for m in movielist]
titles2
```

Out[11]:
```
['About My Father',
 'Afire',
 'Amanda',
 'Asteroid City',
 'Biosphere',
 'Black Ice',
 "Chile '76",
 'Close to Vermeer',
 'Dalíland',
 "Dead Man's Hand",
 'Desperate Souls, Dark City and the Legend of Midnight Cowboy',
 'Earth Mama',
 'Elemental',
 'Every Body',
 'Fast X',
 'Final Cut',
 'Fourth Grade',
 'Full Time',
 'Guardians of the Galaxy Vol. 3',
 'Indiana Jones and the Dial of Destiny',
 'Insidious: The Red Door',
 "It Ain't Over",
 'Joy Ride',
 'Lakota Nation vs. United States',
 'Master Gardener',
 'Mission: Impossible - Dead Reckoning, Part One',
 'No Hard Feelings',
 "Other People's Children",
 'Our Deadly Vows',
 'PSYCHO-PASS: Providence',
 'Past Lives',
 'Revoir Paris',
 'Ruby Gillman, Teenage Kraken',
 'Scarlet',
 'Somewhere in Queens',
 'Sound of Freedom',
 'Spider-Man: Across the Spider-Verse',
 'The Angry Black Girl and Her Monster',
 'The Blackening',
 'The Boogeyman',
 'The Channel',
 'The Childe',
 'The Cow Who Sang a Song Into the Future',
 'The Deepest Breath',
 'The Flash',
 'The Flood',
 'The Last Rider',
 'The Lesson',
 'The Little Mermaid',
 'The Modelizer',
 'The Night of the 12th',
 'The Roundup: No Way Out',
 'The Super Mario Bros. Movie',
 'The YouTube Effect',
 'Theater Camp',
 'Transformers: Rise of the Beasts',
 'Two Tickets to Greece']
```

In [12]:
```
opendates2 = [m.find('span', 'smaller').string.strip() for m in movielist]
```

```
opendates2
```

Out[12]:

```
['Opened May 26, 2023',
 'Opens Jul 14, 2023',
 'Opened Jul 07, 2023',
 'Opened Jun 23, 2023',
 'Opened Jul 07, 2023',
 'Opens Jul 14, 2023',
 'Opened May 05, 2023',
 'Opened May 26, 2023',
 'Opened Jun 09, 2023',
 'Opened Jul 07, 2023',
 'Opened Jun 23, 2023',
 'Opened Jul 07, 2023',
 'Opened Jun 16, 2023',
 'Opened Jun 30, 2023',
 'Opened May 19, 2023',
 'Opens Jul 14, 2023',
 'Opens Jul 14, 2023',
 'Opened Feb 03, 2023',
 'Opened May 05, 2023',
 'Opened Jun 30, 2023',
 'Opened Jul 07, 2023',
 'Opened May 12, 2023',
 'Opened Jul 07, 2023',
 'Opens Jul 14, 2023',
 'Opened May 19, 2023',
 'Opens Jul 12, 2023',
 'Opened Jun 23, 2023',
 'Opened Apr 21, 2023',
 'Opened Jul 07, 2023',
 'Opens Jul 14, 2023',
 'Opened Jun 30, 2023',
 'Opened Jun 23, 2023',
 'Opened Jun 30, 2023',
 'Opened Jun 09, 2023',
 'Opened Apr 21, 2023',
 'Opened Jul 04, 2023',
 'Opened Jun 02, 2023',
 'Opened Jun 09, 2023',
 'Opened Jun 16, 2023',
 'Opened Jun 02, 2023',
 'Opens Jul 14, 2023',
 'Opened Jun 30, 2023',
 'Opened May 19, 2023',
 'Opens Jul 14, 2023',
 'Opened Jun 16, 2023',
 'Opens Jul 14, 2023',
 'Opened Jun 23, 2023',
 'Opened Jul 07, 2023',
 'Opened May 26, 2023',
 'Opens Jul 14, 2023',
 'Opened May 19, 2023',
 'Opened Jun 02, 2023',
 'Opened Apr 05, 2023',
 'Opened Jul 07, 2023',
 'Opens Jul 14, 2023',
 'Opened Jun 09, 2023',
 'Opens Jul 14, 2023']
```

In [13]:
```
audiencescore2 =[m.find('score-pairs')['audiencescore'] for m in movielist]
audiencesentiment2 =[m.find('score-pairs')['audiencesentiment'] for m in movie
criticsscore2 =[m.find('score-pairs')['criticsscore'] for m in movielist]
criticssentiment2 =[m.find('score-pairs')['criticssentiment'] for m in movieli
```

In [14]:
```
endpoints2 = ['https://rottentomatoes.com'+ m.find('a', href=True)['href'] for
endpoints2
```

Out[14]:
```
['https://rottentomatoes.com/m/about_my_father_2023',
 'https://rottentomatoes.com/m/afire',
 'https://rottentomatoes.com/m/amanda_2022',
 'https://rottentomatoes.com/m/asteroid_city',
 'https://rottentomatoes.com/m/biosphere',
 'https://rottentomatoes.com/m/black_ice_2022',
 'https://rottentomatoes.com/m/chile_76',
 'https://rottentomatoes.com/m/close_to_vermeer',
 'https://rottentomatoes.com/m/daliland',
 'https://rottentomatoes.com/m/dead_mans_hand_2023',
 'https://rottentomatoes.com/m/desperate_souls_dark_city_and_the_legend_of_mid
night_cowboy',
 'https://rottentomatoes.com/m/earth_mama',
 'https://rottentomatoes.com/m/elemental_2023',
 'https://rottentomatoes.com/m/every_body',
 'https://rottentomatoes.com/m/fast_x',
 'https://rottentomatoes.com/m/final_cut_2022',
 'https://rottentomatoes.com/m/fourth_grade',
 'https://rottentomatoes.com/m/full_time',
 'https://rottentomatoes.com/m/guardians_of_the_galaxy_vol_3',
 'https://rottentomatoes.com/m/indiana_jones_and_the_dial_of_destiny',
 'https://rottentomatoes.com/m/insidious_the_red_door',
 'https://rottentomatoes.com/m/it_aint_over',
 'https://rottentomatoes.com/m/joy_ride_2023',
 'https://rottentomatoes.com/m/lakota_nation_vs_united_states',
 'https://rottentomatoes.com/m/master_gardener',
 'https://rottentomatoes.com/m/mission_impossible_dead_reckoning_part_one',
 'https://rottentomatoes.com/m/no_hard_feelings_2023',
 'https://rottentomatoes.com/m/other_peoples_children_2022',
 'https://rottentomatoes.com/m/our_deadly_vows',
 'https://rottentomatoes.com/m/psycho_pass_providence',
 'https://rottentomatoes.com/m/past_lives',
 'https://rottentomatoes.com/m/revoir_paris',
 'https://rottentomatoes.com/m/ruby_gillman_teenage_kraken',
 'https://rottentomatoes.com/m/scarlet_2022',
 'https://rottentomatoes.com/m/somewhere_in_queens',
 'https://rottentomatoes.com/m/sound_of_freedom',
 'https://rottentomatoes.com/m/spider_man_across_the_spider_verse',
 'https://rottentomatoes.com/m/the_angry_black_girl_and_her_monster',
 'https://rottentomatoes.com/m/the_blackening',
 'https://rottentomatoes.com/m/the_boogeyman',
 'https://rottentomatoes.com/m/the_channel_2023',
 'https://rottentomatoes.com/m/the_childe',
 'https://rottentomatoes.com/m/the_cow_who_sang_a_song_into_the_future',
 'https://rottentomatoes.com/m/the_deepest_breath',
 'https://rottentomatoes.com/m/the_flash_2023',
 'https://rottentomatoes.com/m/the_flood_2023',
 'https://rottentomatoes.com/m/the_last_rider_2022',
 'https://rottentomatoes.com/m/the_lesson_2023',
 'https://rottentomatoes.com/m/the_little_mermaid_2023',
 'https://rottentomatoes.com/m/the_modelizer',
 'https://rottentomatoes.com/m/the_night_of_the_12th',
 'https://rottentomatoes.com/m/the_roundup_no_way_out',
 'https://rottentomatoes.com/m/the_super_mario_bros_movie',
 'https://rottentomatoes.com/m/the_youtube_effect',
 'https://rottentomatoes.com/m/theater_camp',
 'https://rottentomatoes.com/m/transformers_rise_of_the_beasts',
 'https://rottentomatoes.com/m/two_tickets_to_greece']
```

Build dataframe

In [15]:
```
titles = titles1 + titles2
opendates = opendates1 +opendates2
audiencescore = audiencescore1 + audiencescore2
audiencesentiment = audiencesentiment1 + audiencesentiment2
criticsscore = criticsscore1 + criticsscore2
criticssentiment = criticssentiment1 + criticssentiment2
endpoints = endpoints1 + endpoints2
```

In [16]:
```
mydict = {'titles': titles,
          'opendates': opendates,
          'audiencescore': audiencescore,
          'criticsscore': criticsscore,
          'audiencesentiment': audiencesentiment,
          'criticssentiment': criticssentiment,
           'endpoints': endpoints
          }
rt_df = pd.DataFrame(mydict)
rt_df
```

Out[16]:

	titles	opendates	audiencescore	criticsscore	audiencesentiment	criticssentiment
0	20 Days in Mariupol	Opens Jul 14, 2023		100		positive
1	7:11 PM	Opened Jul 07, 2023	96		positive	
2	Bad Girl Boogey	Opened Jul 07, 2023		70		positive
3	Belle	Opens Jul 14, 2023				
4	Bhaag Saale	Opened Jul 07, 2023				
...	...	...	...	...	...	...
82	The Super Mario Bros. Movie	Opened Apr 05, 2023	95	58	positive	negative
83	The YouTube Effect	Opened Jul 07, 2023		88		positive
84	Theater Camp	Opens Jul 14, 2023	73	80	positive	positive
85	Transformers: Rise of the Beasts	Opened Jun 09, 2023	91	53	positive	negative
86	Two Tickets to Greece	Opens Jul 14, 2023				

87 rows × 7 columns

Spider

In [17]:
```
url = rt_df['endpoints'][0]
```

In [18]:
```
r= requests.get(url, headers=headers)
mysoup= BeautifulSoup(r.text)
#mysoup
```

In [19]:
```
json.loads(mysoup.find('script', type='application/ld+json').text)['director'][
```

Out[19]:
```
'Mstyslav Chernov'
```

In [20]:
```
def myspider(url):
    r= requests.get(url, headers=headers)
    mysoup= BeautifulSoup(r.text)
    director = json.loads(mysoup.find('script', type='application/ld+json').tex
    if len(director) >0:
        name = director[0]['name']
```

```
    else:
        name = ''
    return name
```

In [21]:
```
directors = []
for m in rt_df['endpoints']:
    directors = directors+ [myspider(m)]
```

In [22]:
```
directors
```

Out[22]:
```
['Mstyslav Chernov',
 'Chaitu Madala',
 'Alice Maio Mackay',
 'Max Gold',
 'Pranith Bramandapally',
 'Georgia Oakley',
 'Jean-Luc Godard',
 'Roddy Bogawa',
 'Emanuele Crialese',
 'Rui Cui',
 'Alexandre O. Philippe',
 'Michelle Savill',
 'Anu Menon',
 'Kusanagi',
 'Gary Matoso',
 'Cathryne Czubek',
 'Anthony Pun',
 'Senna Hegde',
 'Pawan Basamsetti',
 'Ajay Samrat',
 'Anton Corbijn',
 '',
 'Leo Milano',
 'Samuel D. Pollard',
 "Thaddeus O'Sullivan",
 'Jean Eustache',
 'Robin Hardy',
 'Anthony Giordano',
 'Béla Tarr',
 'Chris Smith',
 'Laura Terruso',
 'Christian Petzold',
 'Carolina Cavalli',
 'Wes Anderson',
 'Mel Eslyn',
 'Hubert Davis',
 'Manuela Martelli',
 'Suzanne Raes',
 'Mary Harron',
 'Brian Skiba',
 'Nancy Buirski',
 'Savanah Leaf',
 'Peter Sohn',
 'Julie Cohen',
 'Louis Leterrier',
 'Michel Hazanavicius',
 'Marcelo Galvão',
 'Eric Gravel',
 'James Gunn',
 'James Mangold',
 'Patrick Wilson',
 'Sean Mullin',
 'Adele Lim',
 'Jesse Short Bull',
 'Paul Schrader',
 'Christopher McQuarrie',
 'Gene Stupnitsky',
 'Rebecca Zlotowski',
 'Chris Chalk',
 'Naoyoshi Shiotani',
```

```
'Celine Song',
'Alice Winocour',
'Kirk DeMicco',
'Pietro Marcello',
'Ray Romano',
'Alejandro Monteverde',
'Joaquim Dos Santos',
'Bomani J. Story',
'Tim Story',
'Rob Savage',
'William Kaufman',
'Park Hoon-jung',
'Francisca Alegría',
'Laura McGann',
'Andy Muschietti',
'Brandon Slagle',
'Alex Holmes',
'Alice Troughton',
'Rob Marshall',
'Keoni Waxman',
'Dominik Moll',
'Lee Sang-yong',
'Aaron Horvath',
'Alex Winter',
'Molly Gordon',
'Steven Caple Jr.',
'Marc Fitoussi']
```

In [ ]: