In [1]:
```
import pandas as pd
import numpy as np
import requests
import bs4 as bs
import urllib.request
```

Extracting features of 2020 movies from Wikipedia

In [2]:
```
link = "https://en.wikipedia.org/wiki/List_of_American_films_of_2020"
```

In [3]:
```
source = urllib.request.urlopen(link).read()
soup = bs.BeautifulSoup(source,'lxml')
```

In [4]:
```
tables = soup.find_all('table',class_='wikitable sortable')
```

In [5]:
```
len(tables)
```

Out[5]:
```
4
```

In [6]:
```
type(tables[0])
```

Out[6]:
```
bs4.element.Tag
```

In [7]:
```
df1 = pd.read_html(str(tables[0]))[0]
df2 = pd.read_html(str(tables[1]))[0]
df3 = pd.read_html(str(tables[2]))[0]
df4 = pd.read_html(str(tables[3]).replace("'1\"\'",'"1"'))[0] # avoided "Value
```

In [8]:
```
df = df1.append(df2.append(df3.append(df4,ignore_index=True),ignore_index=True
```

In [9]:
```
df
```

Out[9]:

	Opening	Opening.1	Title	Production company	Cast and crew	Ref.
0	JANUARY	3	The Grudge	Screen Gems / Stage 6 Films / Ghost House Pict...	Nicolas Pesce (director/screenplay); Andrea Ri...	[2]
1	JANUARY	10	Underwater	20th Century Fox / TSG Entertainment / Chernin...	William Eubank (director); Brian Duffield, Ada...	[3]
2	JANUARY	10	Three Christs	IFC Films	Jon Avnet (director/screenplay); Eric Nazarian...	NaN
3	JANUARY	10	Like a Boss	Paramount Pictures	Miguel Arteta (director); Sam Pitman, Adam Col...	[4]
4	JANUARY	10	Inherit the Viper	Barry Films / Tycor International Film Company	Anthony Jerjen (director); Andrew Crabtree (sc...	[5]
...	...	...	...	...	...	...
250	DECEMBER	25	News of the World	Universal Pictures / Netflix / Playtone	Paul Greengrass (director/screenplay); Luke Da...	[226]
251	DECEMBER	25	One Night in Miami	Amazon Studios	Regina King (director); Kemp Powers (screenpla...	[227]
252	DECEMBER	25	Promising Young Woman	Focus Features / FilmNation Entertainment	Emerald Fennell (director/screenplay); Carey M...	[228]
253	DECEMBER	25	Sylvie's Love	Amazon Studios	Eugene Ashe (director/screenplay); Tessa Thomp...	[229]
254	DECEMBER	30	Pieces of a Woman	Netflix / Bron Studios	Kornél Mundruczó (director); Kata Wéber (scree...	[230]

255 rows × 6 columns

In [10]:
```
df_2020 = df[['Title','Cast and crew']]
```

In [11]:
```
df_2020
```

Out[11]:

	Title	Cast and crew
0	The Grudge	Nicolas Pesce (director/screenplay); Andrea Ri...
1	Underwater	William Eubank (director); Brian Duffield, Ada...
2	Three Christs	Jon Avnet (director/screenplay); Eric Nazarian...
3	Like a Boss	Miguel Arteta (director); Sam Pitman, Adam Col...
4	Inherit the Viper	Anthony Jerjen (director); Andrew Crabtree (sc...
...	...	...
250	News of the World	Paul Greengrass (director/screenplay); Luke Da...
251	One Night in Miami	Regina King (director); Kemp Powers (screenpla...
252	Promising Young Woman	Emerald Fennell (director/screenplay); Carey M...
253	Sylvie's Love	Eugene Ashe (director/screenplay); Tessa Thomp...
254	Pieces of a Woman	Kornél Mundruczó (director); Kata Wéber (scree...

255 rows × 2 columns

In [12]:
```
!pip install tmdbv3api
```

```
Collecting tmdbv3api
  Downloading https://files.pythonhosted.org/packages/fa/cb/72ca70a05b7364c2b
41e6cf1f615729b0c99109a3be0e61e22d42859d48f/tmdbv3api-1.7.1-py2.py3-none-any.
whl (https://files.pythonhosted.org/packages/fa/cb/72ca70a05b7364c2b41e6cf1f6
15729b0c99109a3be0e61e22d42859d48f/tmdbv3api-1.7.1-py2.py3-none-any.whl)
Requirement already satisfied: requests in /usr/local/lib/python3.6/dist-pack
ages (from tmdbv3api) (2.23.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.
6/dist-packages (from requests->tmdbv3api) (2020.11.8)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.6/dist-
packages (from requests->tmdbv3api) (2.10)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /us
r/local/lib/python3.6/dist-packages (from requests->tmdbv3api) (1.24.3)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.6/
dist-packages (from requests->tmdbv3api) (3.0.4)
Installing collected packages: tmdbv3api
Successfully installed tmdbv3api-1.7.1
```

In [13]:
```
from tmdbv3api import TMDb
import json
import requests
tmdb = TMDb()
tmdb.api_key = ''
```

In [14]:

```
from tmdbv3api import Movie
tmdb_movie = Movie()
def get_genre(x):
    genres = []
    result = tmdb_movie.search(x)
    if not result:
      return np.NaN
    else:
      movie_id = result[0].id
      response = requests.get('https://api.themoviedb.org/3/movie/{}?api_key={
      data_json = response.json()
      if data_json['genres']:
          genre_str = " "
          for i in range(0,len(data_json['genres'])):
              genres.append(data_json['genres'][i]['name'])
          return genre_str.join(genres)
      else:
          return np.NaN
```

In [15]:

```
df_2020['genres'] = df_2020['Title'].map(lambda x: get_genre(str(x)))
```

```
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:1: SettingWithCo
pyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/s
table/user_guide/indexing.html#returning-a-view-versus-a-copy (https://panda
s.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-ver
sus-a-copy)
  """Entry point for launching an IPython kernel.
```

In [16]:
```
df_2020
```

Out[16]:

	Title	Cast and crew	genres
0	The Grudge	Nicolas Pesce (director/screenplay); Andrea Ri...	Horror Mystery Thriller
1	Underwater	William Eubank (director); Brian Duffield, Ada...	Action Horror Science Fiction Thriller
2	Three Christs	Jon Avnet (director/screenplay); Eric Nazarian...	Drama
3	Like a Boss	Miguel Arteta (director); Sam Pitman, Adam Col...	Comedy
4	Inherit the Viper	Anthony Jerjen (director); Andrew Crabtree (sc...	Drama Thriller Crime
...	...	...	...
250	News of the World	Paul Greengrass (director/screenplay); Luke Da...	Drama Western
251	One Night in Miami	Regina King (director); Kemp Powers (screenpla...	Drama
252	Promising Young Woman	Emerald Fennell (director/screenplay); Carey M...	Thriller Crime Drama
253	Sylvie's Love	Eugene Ashe (director/screenplay); Tessa Thomp...	Drama
254	Pieces of a Woman	Kornél Mundruczó (director); Kata Wéber (scree...	Drama

255 rows × 3 columns

In [17]:
```
def get_director(x):
    if " (director)" in x:
        return x.split(" (director)")[0]
    elif " (directors)" in x:
        return x.split(" (directors)")[0]
    else:
        return x.split(" (director/screenplay)")[0]
```

In [18]:
```
df_2020['director_name'] = df_2020['Cast and crew'].map(lambda x: get_director
```

```
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:1: SettingWithCo
pyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/s
table/user_guide/indexing.html#returning-a-view-versus-a-copy (https://panda
s.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-ver
sus-a-copy)
  """Entry point for launching an IPython kernel.
```

In [19]:
```
def get_actor1(x):
    return ((x.split("screenplay); ")[-1]).split(", ")[0])
```

In [20]:
```
df_2020['actor_1_name'] = df_2020['Cast and crew'].map(lambda x: get_actor1(st
```

```
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:1: SettingWithCo
pyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/s
table/user_guide/indexing.html#returning-a-view-versus-a-copy (https://panda
s.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-ver
sus-a-copy)
  """Entry point for launching an IPython kernel.
```

In [21]:
```
def get_actor2(x):
    if len((x.split("screenplay); ")[-1]).split(", ")) < 2:
        return np.NaN
    else:
        return ((x.split("screenplay); ")[-1]).split(", ")[1])
```

In [22]:
```
df_2020['actor_2_name'] = df_2020['Cast and crew'].map(lambda x: get_actor2(st
```

In [23]:
```
def get_actor3(x):
    if len((x.split("screenplay); ")[-1]).split(", ")) < 3:
        return np.NaN
    else:
        return ((x.split("screenplay); ")[-1]).split(", ")[2])
```

In [24]:
```
df_2020['actor_3_name'] = df_2020['Cast and crew'].map(lambda x: get_actor3(st
```

In [25]:
```
df_2020
```

Out[25]:

	Title	Cast and crew	genres	director_name	actor_1_name	actor_2_name	acto
0	The Grudge	Nicolas Pesce (director/screenplay); Andrea Ri...	Horror Mystery Thriller	Nicolas Pesce	Andrea Riseborough	Demián Bichir	
1	Underwater	William Eubank (director); Brian Duffield, Ada...	Action Horror Science Fiction Thriller	William Eubank	Kristen Stewart	Vincent Cassel	
2	Three Christs	Jon Avnet (director/screenplay); Eric Nazarian...	Drama	Jon Avnet	Richard Gere	Peter Dinklage	
3	Like a Boss	Miguel Arteta (director); Sam Pitman, Adam Col...	Comedy	Miguel Arteta	Tiffany Haddish	Rose Byrne	Sal
4	Inherit the Viper	Anthony Jerjen (director); Andrew Crabtree (sc...	Drama Thriller Crime	Anthony Jerjen	Josh Hartnett	Margarita Levieva	
...	...	...	...	...	...	...	
250	News of the World	Paul Greengrass (director/screenplay); Luke Da...	Drama Western	Paul Greengrass	Tom Hanks	Helena Zengel	
251	One Night in Miami	Regina King (director); Kemp Powers (screenpla...	Drama	Regina King	Kingsley Ben-Adir	Eli Goree	Al
252	Promising Young Woman	Emerald Fennell (director/screenplay); Carey M...	Thriller Crime Drama	Emerald Fennell	Carey Mulligan	Bo Burnham	/
253	Sylvie's Love	Eugene Ashe (director/screenplay); Tessa Thomp...	Drama	Eugene Ashe	Tessa Thompson	Nnamdi Asomugha	Rya
254	Pieces of a Woman	Kornél Mundruczó (director); Kata Wéber (scree...	Drama	Kornél Mundruczó	Vanessa Kirby	Shia LaBeouf	M

255 rows × 7 columns

In [26]:
```
df_2020 = df_2020.rename(columns={'Title':'movie_title'})
```

In [27]:
```
new_df20 = df_2020.loc[:,['director_name','actor_1_name','actor_2_name','actor
```

In [28]:

```
new_df20
```

Out[28]:

	director_name	actor_1_name	actor_2_name	actor_3_name	genres	movie_title
0	Nicolas Pesce	Andrea Riseborough	Demián Bichir	John Cho	Horror Mystery Thriller	The Grudge
1	William Eubank	Kristen Stewart	Vincent Cassel	Jessica Henwick	Action Horror Science Fiction Thriller	Underwater
2	Jon Avnet	Richard Gere	Peter Dinklage	Walton Goggins	Drama	Three Christs
3	Miguel Arteta	Tiffany Haddish	Rose Byrne	Salma Hayek	Comedy	Like a Boss
4	Anthony Jerjen	Josh Hartnett	Margarita Levieva	Chandler Riggs	Drama Thriller Crime	Inherit the Viper
...	...	...	...	...	...	...
250	Paul Greengrass	Tom Hanks	Helena Zengel	NaN	Drama Western	News of the World
251	Regina King	Kingsley Ben-Adir	Eli Goree	Aldis Hodge	Drama	One Night in Miami
252	Emerald Fennell	Carey Mulligan	Bo Burnham	Alison Brie	Thriller Crime Drama	Promising Young Woman
253	Eugene Ashe	Tessa Thompson	Nnamdi Asomugha	Ryan Michelle Bathe	Drama	Sylvie's Love
254	Kornél Mundruczó	Vanessa Kirby	Shia LaBeouf	Molly Parker	Drama	Pieces of a Woman

255 rows × 6 columns

In [29]:

```
new_df20['comb'] = new_df20['actor_1_name'] + ' ' + new_df20['actor_2_name'] +
```

In [30]:

```
new_df20.isna().sum()
```

Out[30]:

```
director_name     0
actor_1_name      0
actor_2_name      2
actor_3_name     27
genres            2
movie_title       0
comb             28
dtype: int64
```

In [31]:

```
new_df20 = new_df20.dropna(how='any')
```

In [32]:
```
new_df20.isna().sum()
```

Out[32]:
```
director_name    0
actor_1_name     0
actor_2_name     0
actor_3_name     0
genres           0
movie_title      0
comb             0
dtype: int64
```

In [33]:
```
new_df20['movie_title'] = new_df20['movie_title'].str.lower()
```

```
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:1: SettingWithCo
pyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/s
table/user_guide/indexing.html#returning-a-view-versus-a-copy (https://panda
s.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-ver
sus-a-copy)
  """Entry point for launching an IPython kernel.
```

In [34]:

```
new_df20
```

Out[34]:

	director_name	actor_1_name	actor_2_name	actor_3_name	genres	movie_title	cc
0	Nicolas Pesce	Andrea Riseborough	Demián Bichir	John Cho	Horror Mystery Thriller	the grudge	Anc Riseboro Den Bichir J Cho Nic
1	William Eubank	Kristen Stewart	Vincent Cassel	Jessica Henwick	Action Horror Science Fiction Thriller	underwater	Kris Stev Vinc Ca Jes Henwic
2	Jon Avnet	Richard Gere	Peter Dinklage	Walton Goggins	Drama	three christs	Rich Gere P Dinkl Wa Gogg Jc
3	Miguel Arteta	Tiffany Haddish	Rose Byrne	Salma Hayek	Comedy	like a boss	Tiff Hadc Rose By Sa Ha Migue
4	Anthony Jerjen	Josh Hartnett	Margarita Levieva	Chandler Riggs	Drama Thriller Crime	inherit the viper	J Hart Marga Levi Chan Rigg
...	...	...	...	...	...	...	
249	Paul W. S. Anderson	Milla Jovovich	Tony Jaa	Tip "T.I." Harris	Fantasy Action Adventure	monster hunter	M Jovo Tony Jaa "T.I." Ha Pa
251	Regina King	Kingsley Ben-Adir	Eli Goree	Aldis Hodge	Drama	one night in miami	Kings Ben-Adi Goree A Ho Regir
252	Emerald Fennell	Carey Mulligan	Bo Burnham	Alison Brie	Thriller Crime Drama	promising young woman	Ca Mulligan Burnh Alison Emeral
253	Eugene Ashe	Tessa Thompson	Nnamdi Asomugha	Ryan Michelle Bathe	Drama	sylvie's love	Te Thomp Nna Asomu R Michelle
254	Kornél Mundruczó	Vanessa Kirby	Shia LaBeouf	Molly Parker	Drama	pieces of a woman	Vane Kirby S LaBe Molly Pa Korn

227 rows × 7 columns

```
In [ ]: old_df = pd.read_csv('final_data.csv')
```

In [ ]:

```
old_df
```

Out[39]:

	director_name	actor_1_name	actor_2_name	actor_3_name	genres	movie_title	
0	James Cameron	CCH Pounder	Joel David Moore	Wes Studi	Action Adventure Fantasy Sci-Fi	avatar	Po I Wes
1	Gore Verbinski	Johnny Depp	Orlando Bloom	Jack Davenport	Action Adventure Fantasy	pirates of the caribbean: at world's end	J Or E Dave G
2	Sam Mendes	Christoph Waltz	Rory Kinnear	Stephanie Sigman	Action Adventure Thriller	spectre	Chr Ki Step Sign
3	Christopher Nolan	Tom Hardy	Christian Bale	Joseph Gordon-Levitt	Action Thriller	the dark knight rises	Ch J Go Le
4	Doug Walker	Doug Walker	Rob Walker	unknown	Documentary	star wars: episode vii - the force awakens ...	V V unk V
...	...	...	...	...	...	...	
5864	Greta Gerwig	Saoirse Ronan	Emma Watson	Florence Pugh	Drama Romance	little women	Sa F E W Flo
5865	Sam Mendes	George MacKay	Dean-Charles Chapman	Mark Strong	War Drama Action History	1917	Gr G Ma I Cl Cha Str
5866	Destin Daniel Cretton	Michael B. Jordan	Jamie Foxx	Brie Larson	Drama Crime	just mercy	M B. J Fox L D

	director_name	actor_1_name	actor_2_name	actor_3_name	genres	movie_title	
5867	Chinonye Chukwu	Alfre Woodard	Wendell Pierce	Aldis Hodge	Drama	clemency	Wo W F H Cl
5868	Waymon Boone	Mena Suvari	Kevin Pollak	unknown	Horror Thriller	apparition	S I unk Wa Bo

5869 rows × 7 columns

```
In [ ]: final_df = old_df.append(new_df20,ignore_index=True)
In [ ]: final_df
```

Out[41]:

	director_name	actor_1_name	actor_2_name	actor_3_name	genres	movie_title	
0	James Cameron	CCH Pounder	Joel David Moore	Wes Studi	Action Adventure Fantasy Sci-Fi	avatar	Pc Joel We: Jam
1	Gore Verbinski	Johnny Depp	Orlando Bloom	Jack Davenport	Action Adventure Fantasy	pirates of the caribbean: at world's end	J C Dav C
2	Sam Mendes	Christoph Waltz	Rory Kinnear	Stephanie Sigman	Action Adventure Thriller	spectre	Ch Walt K Ste Sigı
3	Christopher Nolan	Tom Hardy	Christian Bale	Joseph Gordon-Levitt	Action Thriller	the dark knight rises	Tom Ch J G L
4	Doug Walker	Doug Walker	Rob Walker	unknown	Documentary	star wars: episode vii - the force awakens ...	un
...	...	...	...	...	...	...	
6005	Joseph Kosinski	Tom Cruise	Miles Teller	Jennifer Connelly	Action Drama	top gun: maverick	J(C(
6006	Joel Crawford	Nicolas Cage	Emma Stone	Ryan Reynolds	Animation Adventure Family	the croods 2	N Re Joe
6007	Liesl Tommy	Jennifer Hudson	Forest Whitaker	Marlon Wayans	Music Drama	respect	J H W Way

	director_name	actor_1_name	actor_2_name	actor_3_name	genres	movie_title	
6008	Ridley Scott	Matt Damon	Adam Driver	Jodie Comer	Drama	the last duel	[
6009	Paul Greengrass	Tom Hanks	Helena Zengel	Neil Sandilands	Drama Western	news of the world	I Sanc Pa

6010 rows × 7 columns

In [ ]:
```
final_df.to_csv('main_data.csv',index=False)
```