Who are the true trending actors?

1. Preparation and execution of scraping (collect_data.py)

Import Library

```
from bs4 import BeautifulSoup
import requests
import pandas as pd
```

Config

```
config = {
  "START": XXX, # ex) 1
  "COUNT": XXX, # ex) 200
}

url = f'https://www.imdb.com/search/title/?title_type=feature,tv_series&count={config["COUNT"]}&start={config["START"]}&ref_=adv_nxt'

html = requests.get(url)
soup = BeautifulSoup(html.content, 'html.parser')

data = {'names': [],
        'rates': [],
        'director': [],
        'actors': []}
```

(scrape from IMDb Feature Film/TV Series (Sorted by Popularity Ascending))

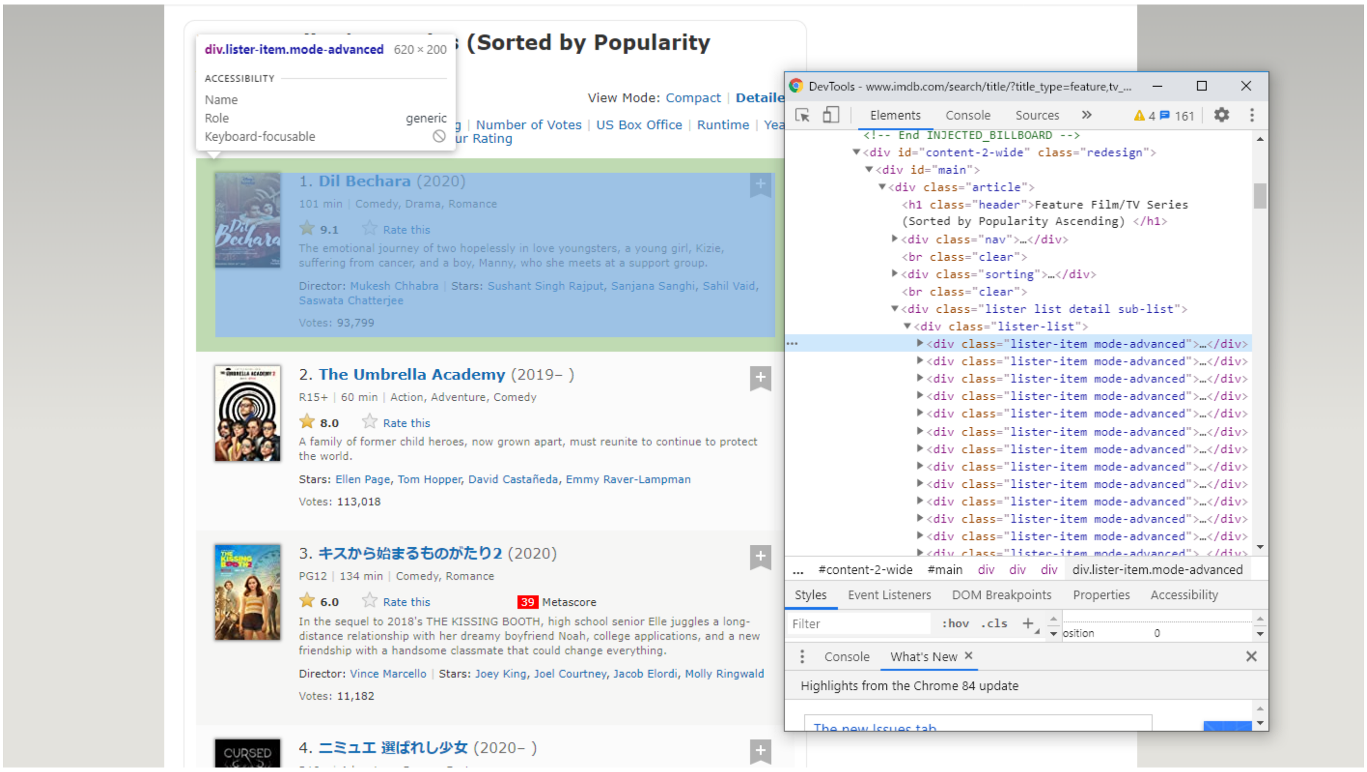
You can get the specified number of data by setting the COUNT and START queries for the IMDB URLs.

The information you will get is name, rate, director and actor.

Scraping

```
movie_list = soup.select('div[class="lister-list"] div[class="lister-item mode-advanced"]')

for movie in movie_list:
  name = movie.h3.a.text

  try:
    rate = float(movie.select('div[class="inline-block ratings-imdb-rating"] strong')[0].text)
  except:
    rate = ''

  staff_li = movie.select('p')[2]
  staff_li = staff_li.text.replace('\n', '').split('|')


  director = ""
  actor = ""

  for staff in staff_li:
    if "Director:" in staff:
      director = staff.replace("Director:", "").strip()
    if "Stars:" in staff:
      actor = staff.replace("Stars:", "").strip()

  data["names"].append(name)
  data["rates"].append(rate)
  data["director"].append(director)
  data["actors"].append(actor)


df = pd.DataFrame(data)

df.to_csv(f'{config["START"]}to{config["START"] + config["COUNT"]}_movies.csv',
index=False)
```

You will see that there is a list of movie information you want to get to the child class of `div[class="inline-block ratings-imdb-rating"] strong`



div.lister-item.mode-advanced 620 × 200

ACCESSIBILITY
Name
Role generic
Keyboard-focusable ⃠

(Sorted by Popularity

View Mode: Compact | Detaile

g | Number of Votes | US Box Office | Runtime | Yea
ur Rating

1. Dil Bechara (2020)
101 min | Comedy, Drama, Romance
9.1 Rate this
The emotional journey of two hopelessly in love youngsters, a young girl, Kizie, suffering from cancer, and a boy, Manny, who she meets at a support group.
Director: Mukesh Chhabra | Stars: Sushant Singh Rajput, Sanjana Sanghi, Sahil Vaid, Saswata Chatterjee
Votes: 93,799

2. The Umbrella Academy (2019–)
R15+ | 60 min | Action, Adventure, Comedy
8.0 Rate this
A family of former child heroes, now grown apart, must reunite to continue to protect the world.
Stars: Ellen Page, Tom Hopper, David Castañeda, Emmy Raver-Lampman
Votes: 113,018

3. キスから始まるものがたり2 (2020)
PG12 | 134 min | Comedy, Romance
6.0 Rate this 39 Metascore
In the sequel to 2018's THE KISSING BOOTH, high school senior Elle juggles a long-distance relationship with her dreamy boyfriend Noah, college applications, and a new friendship with a handsome classmate that could change everything.
Director: Vince Marcello | Stars: Joey King, Joel Courtney, Jacob Elordi, Molly Ringwald
Votes: 11,182

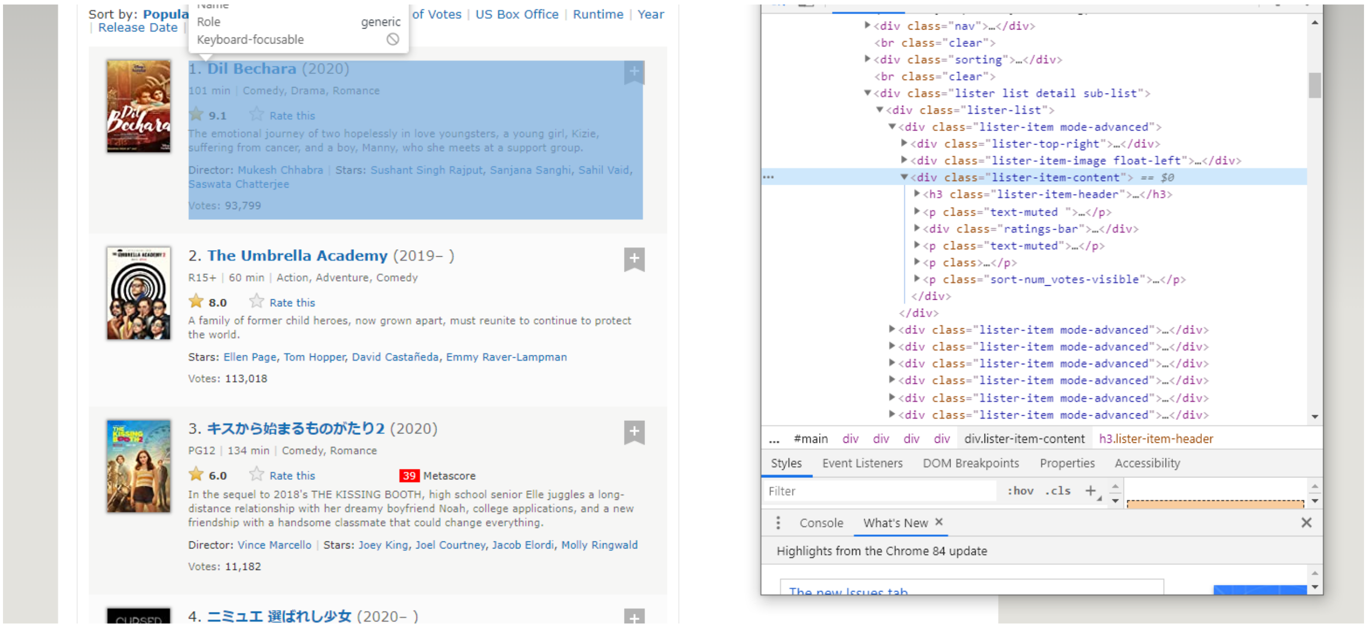
4. ニミュエ 選ばれし少女 (2020–)

DevTools - www.imdb.com/search/title/?title_type=feature,tv_...

```
<!-- End INJECTED_BILLBOARD -->
▼<div id="content-2-wide" class="redesign">
  ▼<div id="main">
    ▼<div class="article">
      <h1 class="header">Feature Film/TV Series
      (Sorted by Popularity Ascending) </h1>
     ▶<div class="nav">…</div>
      <br class="clear">
     ▶<div class="sorting">…</div>
      <br class="clear">
     ▼<div class="lister list detail sub-list">
       ▼<div class="lister-list">
         ▶<div class="lister-item mode-advanced">…</div>
         ▶<div class="lister-item mode-advanced">…</div>
         ▶<div class="lister-item mode-advanced">…</div>
         ▶<div class="lister-item mode-advanced">…</div>
         ▶<div class="lister-item mode-advanced">…</div>
         ▶<div class="lister-item mode-advanced">…</div>
         ▶<div class="lister-item mode-advanced">…</div>
         ▶<div class="lister-item mode-advanced">…</div>
         ▶<div class="lister-item mode-advanced">…</div>
         ▶<div class="lister-item mode-advanced">…</div>
         ▶<div class="lister-item mode-advanced">…</div>
         ▶<div class="lister-item mode-advanced">…</div>
         ▶<div class="lister-item mode-advanced">…</div>
```

... #content-2-wide #main div div div div.lister-item.mode-advanced

And you'll find all sorts of information in `div[class="lister-item-content"]` We're going to write the code in the same way below, checking it with devtool.



1. Dil Bechara (2020)
101 min | Comedy, Drama, Romance
9.1 Rate this
The emotional journey of two hopelessly in love youngsters, a young girl, Kizie, suffering from cancer, and a boy, Manny, who she meets at a support group.
Director: Mukesh Chhabra | Stars: Sushant Singh Rajput, Sanjana Sanghi, Sahil Vaid, Saswata Chatterjee
Votes: 93,799

2. The Umbrella Academy (2019–)
R15+ | 60 min | Action, Adventure, Comedy
8.0 Rate this
A family of former child heroes, now grown apart, must reunite to continue to protect the world.
Stars: Ellen Page, Tom Hopper, David Castañeda, Emmy Raver-Lampman
Votes: 113,018

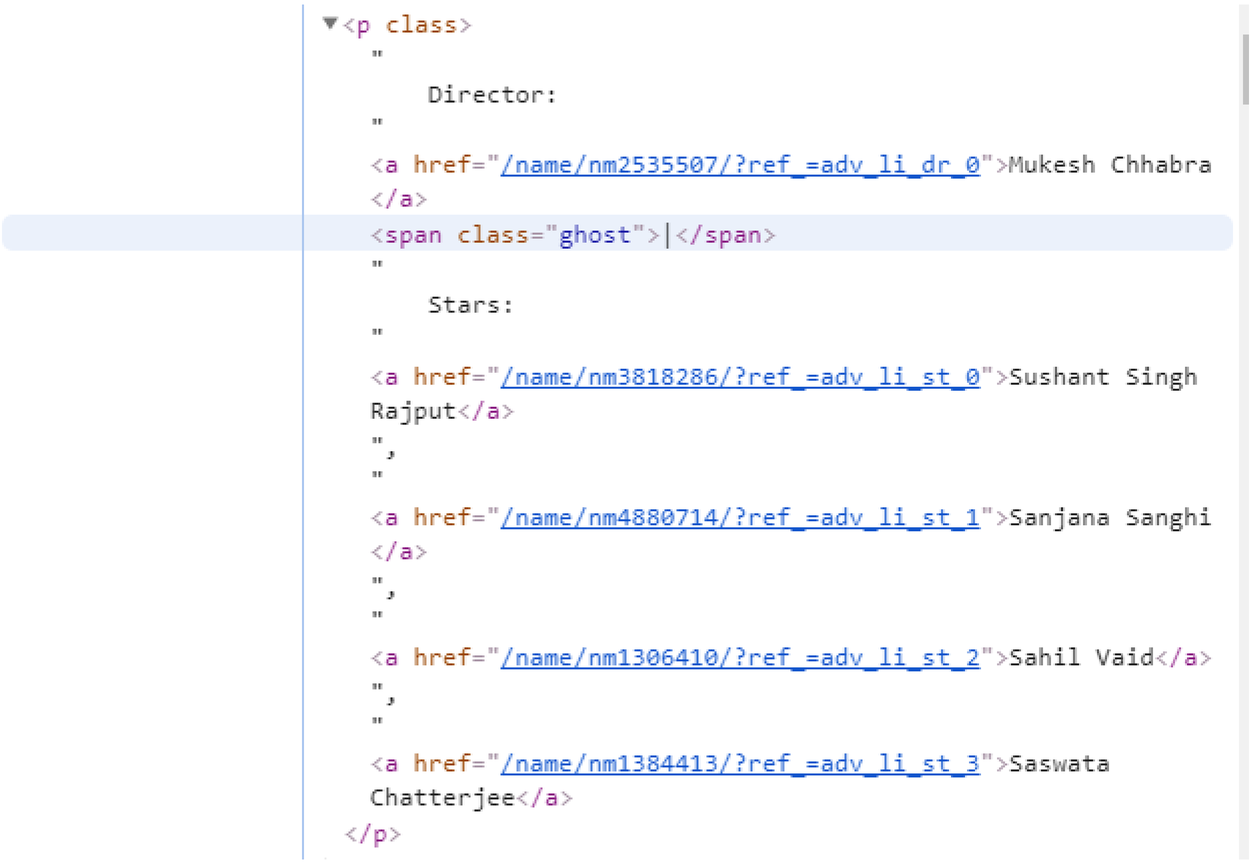
3. キスから始まるものがたり2 (2020)
PG12 | 134 min | Comedy, Romance
6.0 Rate this 39 Metascore
In the sequel to 2018's THE KISSING BOOTH, high school senior Elle juggles a long-distance relationship with her dreamy boyfriend Noah, college applications, and a new friendship with a handsome classmate that could change everything.
Director: Vince Marcello | Stars: Joey King, Joel Courtney, Jacob Elordi, Molly Ringwald
Votes: 11,182

4. ニミュエ 選ばれし少女 (2020–)

```
▶<div class="nav">…</div>
 <br class="clear">
▶<div class="sorting">…</div>
 <br class="clear">
▼<div class="lister list detail sub-list">
  ▼<div class="lister-list">
    ▼<div class="lister-item mode-advanced">
      ▶<div class="lister-top-right">…</div>
      ▶<div class="lister-item-image float-left">…</div>
      ▼<div class="lister-item-content"> == $0
        ▶<h3 class="lister-item-header">…</h3>
        ▶<p class="text-muted ">…</p>
        ▶<div class="ratings-bar">…</div>
        ▶<p class="text-muted">…</p>
        ▶<p class>…</p>
        ▶<p class="sort-num_votes-visible">…</p>
       </div>
     </div>
    ▶<div class="lister-item mode-advanced">…</div>
    ▶<div class="lister-item mode-advanced">…</div>
    ▶<div class="lister-item mode-advanced">…</div>
    ▶<div class="lister-item mode-advanced">…</div>
    ▶<div class="lister-item mode-advanced">…</div>
    ▶<div class="lister-item mode-advanced">…</div>
```

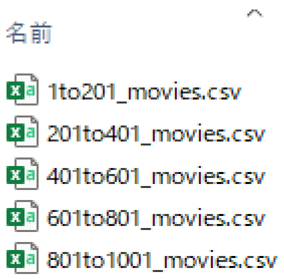
However, getting a director and an actor is a bit tricky because sometimes there is only one director and one actor, and sometimes there is only one director and one actor.

If there are both, the only thing that helps is that they are separated by "|", so I was able to write the conditional branch carefully.



```
▼<p class>
   "
       Director:
   "
   <a href="/name/nm2535507/?ref_=adv_li_dr_0">Mukesh Chhabra
   </a>
   <span class="ghost">|</span>
   "
       Stars:
   "
   <a href="/name/nm3818286/?ref_=adv_li_st_0">Sushant Singh
   Rajput</a>
   "
    ,
   "
   <a href="/name/nm4880714/?ref_=adv_li_st_1">Sanjana Sanghi
   </a>
   "
    ,
   "
   <a href="/name/nm1306410/?ref_=adv_li_st_2">Sahil Vaid</a>
   "
    ,
   "
   <a href="/name/nm1384413/?ref_=adv_li_st_3">Saswata
   Chatterjee</a>
 </p>
```

That's it, we're done getting the data !



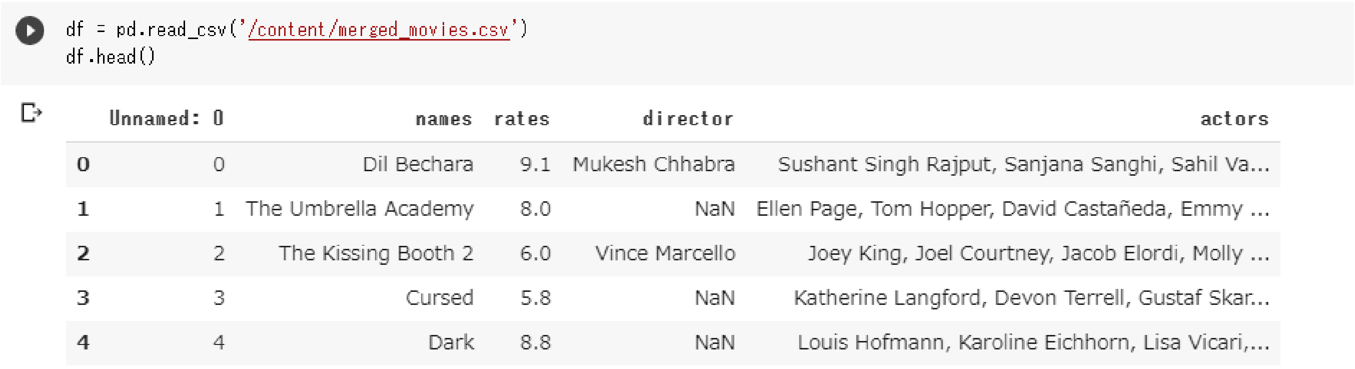
名前

- 1to201_movies.csv
- 201to401_movies.csv
- 401to601_movies.csv
- 601to801_movies.csv
- 801to1001_movies.csv

2. Merge CSV Data (merge_data.py)

```
import pandas as pd
import os

datas = os.listdir('data')

print(datas)

df_merged = pd.read_csv(f'data/{datas[0]}')

# print(df_merged)

for i in range(len(datas)):
  if i == 0:
    continue
  data = pd.read_csv(f'data/{datas[i]}')
  df_merged = pd.concat([df_merged, data])

print(df_merged.shape)

df_merged.to_csv('merged_movies.csv', index=False)
```

The pandas concat merges 5 csv files and outputs 1000 movie information in a single file.



```
df = pd.read_csv('/content/merged_movies.csv')
df.head()
```

	Unnamed: 0	names	rates	director	actors
0	0	Dil Bechara	9.1	Mukesh Chhabra	Sushant Singh Rajput, Sanjana Sanghi, Sahil Va...
1	1	The Umbrella Academy	8.0	NaN	Ellen Page, Tom Hopper, David Castañeda, Emmy ...
2	2	The Kissing Booth 2	6.0	Vince Marcello	Joey King, Joel Courtney, Jacob Elordi, Molly ...
3	3	Cursed	5.8	NaN	Katherine Langford, Devon Terrell, Gustaf Skar...
4	4	Dark	8.8	NaN	Louis Hofmann, Karoline Eichhorn, Lisa Vicari,...

When viewed as a DataFrame type, this is what it looks like 😃

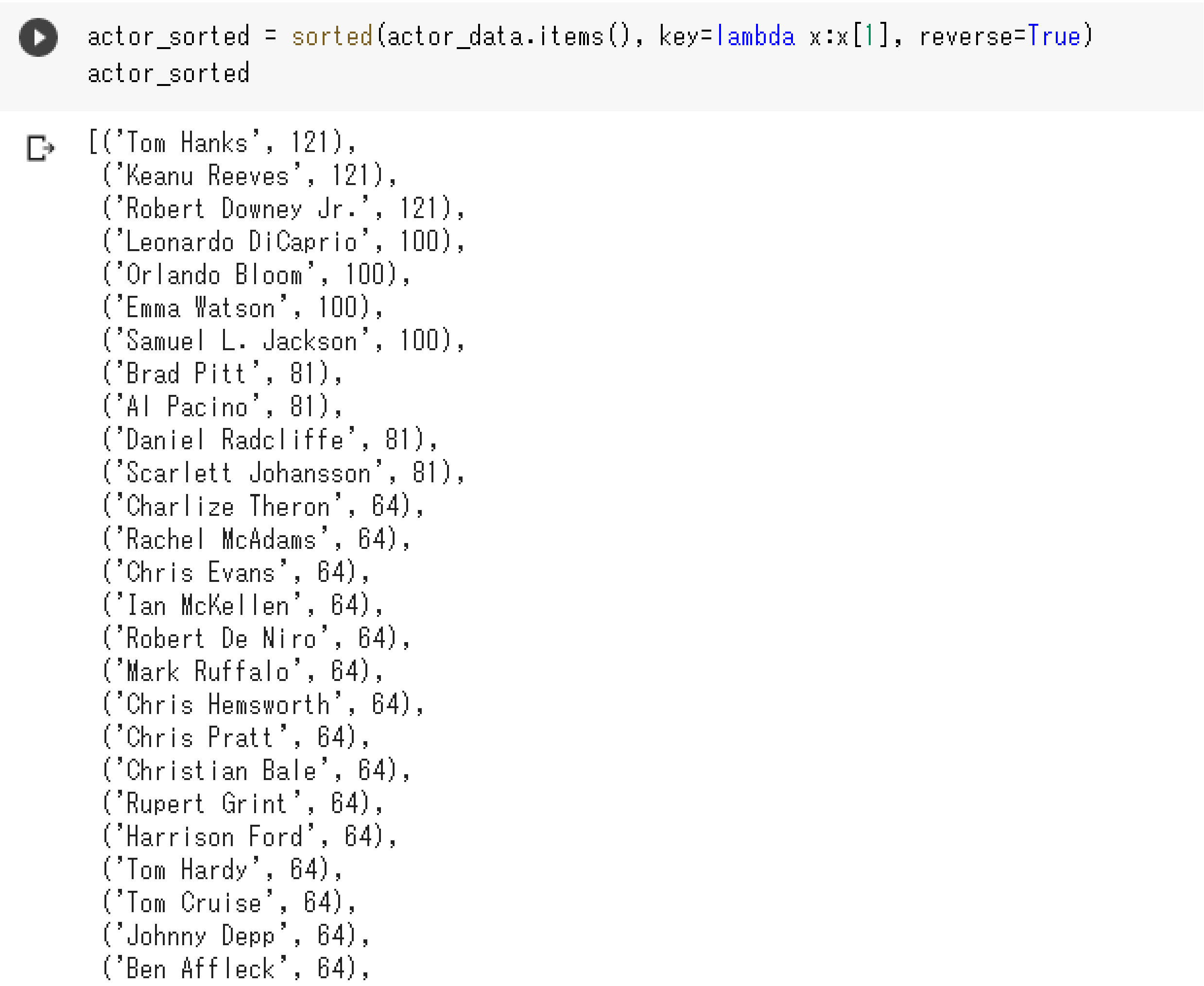
3. Analysis: Who are the trending actors ? (analysis_movie.py)

```python
# -*- coding: utf-8 -*-
"""analysis_movie

Automatically generated by Colaboratory.

Original file is located at
    https://colab.research.google.com/drive/1Vig8lNhzV8C_498oAWbW0nA5V75XnVUQ
"""

import pandas as pd
import numpy as np

df = pd.read_csv('/content/merged_movies.csv')
df.head()

df['actors'].isnull().sum()

# >> 2

df = df.dropna(subset = ['actors'])

df.head()

actor_li = []
for actors in df['actors']:
  for actor in str(actors).split(','):
    actor_li.append(actor.strip())

from collections import defaultdict
d = defaultdict(int)

for actor in df['actors']:
  for ref_actor in actor_li:
    if ref_actor in actor:
      d[ref_actor]+=1

actor_data = dict(d)

actor_sorted = sorted(actor_data.items(), key=lambda x:x[1], reverse=True)
actor_sorted
```

Here's a tally of which actors have appeared in these 1,000 movies

The actors column contains multiple actors' names separated by commas, so we paid attention to this.

Here are the results in total !!



```
actor_sorted = sorted(actor_data.items(), key=lambda x:x[1], reverse=True)
actor_sorted
```

```
[('Tom Hanks', 121),
 ('Keanu Reeves', 121),
 ('Robert Downey Jr.', 121),
 ('Leonardo DiCaprio', 100),
 ('Orlando Bloom', 100),
 ('Emma Watson', 100),
 ('Samuel L. Jackson', 100),
 ('Brad Pitt', 81),
 ('Al Pacino', 81),
 ('Daniel Radcliffe', 81),
 ('Scarlett Johansson', 81),
 ('Charlize Theron', 64),
 ('Rachel McAdams', 64),
 ('Chris Evans', 64),
 ('Ian McKellen', 64),
 ('Robert De Niro', 64),
 ('Mark Ruffalo', 64),
 ('Chris Hemsworth', 64),
 ('Chris Pratt', 64),
 ('Christian Bale', 64),
 ('Rupert Grint', 64),
 ('Harrison Ford', 64),
 ('Tom Hardy', 64),
 ('Tom Cruise', 64),
 ('Johnny Depp', 64),
 ('Ben Affleck', 64),
```

(A little abbreviated because there are a lot of them)

I'll list the top 15 on this list !

rank, name, count

1. 'Tom Hanks', 121
2. 'Keanu Reeves', 121
3. 'Robert Downey Jr.', 121
4. 'Leonardo DiCaprio', 100
5. 'Orlando Bloom', 100
6. 'Emma Watson', 100
7. 'Samuel L. Jackson', 100
8. 'Brad Pitt', 81
9. 'Al Pacino', 81
10. 'Daniel Radcliffe', 81
11. 'Scarlett Johansson', 81
12. 'Charlize Theron', 64
13. 'Rachel McAdams', 64
14. 'Chris Evans', 64
15. 'Ian McKellen', 64

They're all famous !

the end my report, thank you for reading 😃