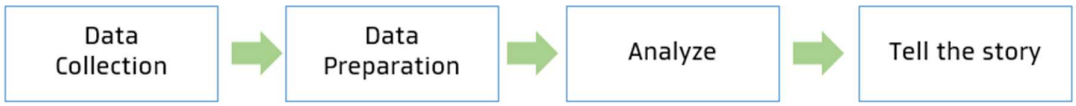
Data Journey



Data Collection

Data มาจาก 2 แหล่ง คือ

- IMDB
  - https://www.imdb.com/interfaces/
  - 9,267,898 Row, 8 Column
- Boxofficemojo
  - https://www.kaggle.com/datasets/eliasdabbas/boxofficemojo-alltime-domestic-data
  - 16543 Row, 5 Column

IMDb File (**title.basics.tsv.gz**)

	tconst	titleType	primaryTitle	originalTitle	isAdult	startYear	endYear	runtimeMinutes	genres
0	tt0000001	short	Carmencita	Carmencita	0	1894	\N	1	Documentary,Short
1	tt0000002	short	Le clown et ses chiens	Le clown et ses chiens	0	1892	\N	5	Animation,Short
2	tt0000003	short	Pauvre Pierrot	Pauvre Pierrot	0	1892	\N	4	Animation,Comedy,Romance
3	tt0000004	short	Un bon bock	Un bon bock	0	1892	\N	12	Animation,Short
4	tt0000005	short	Blacksmith Scene	Blacksmith Scene	0	1893	\N	1	Comedy,Short

Layout

	Field Name	Type	Info
1	tconst	string	unique identifier of the title
2	titleType	string	the type/format of the title
3	primaryTitle	string	popular title
4	originalTitle	string	original title
5	isAdult	bool	0: non-adult title; 1: adult title
6	startYear	YYYY	release year
7	endYear	YYYY	end year
8	runtimeMinutes	int	primary runtime of the title, in minutes
9	genres	string array	includes up to three genres associated with the title

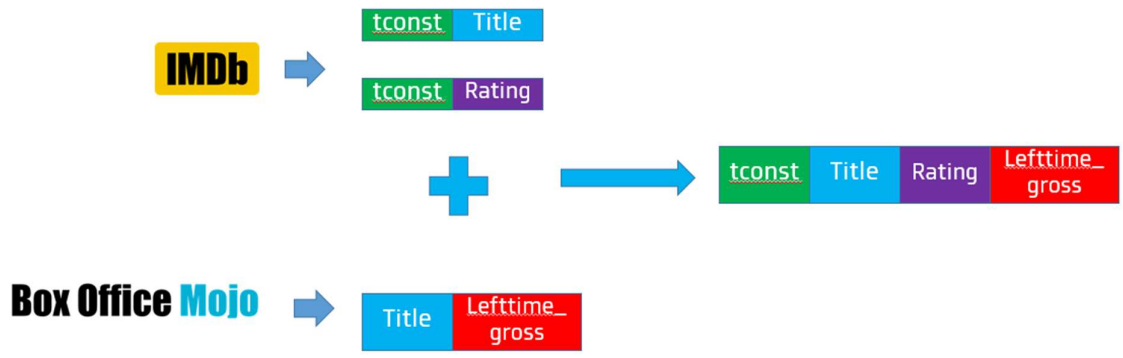
Boxofficemojo (boxoffice_2.csv)

rank	title	studio	lifetime_gross	year
1	Star Wars: The Force Awakens	BV	936662225	2015
2	Avengers: Endgame	BV	857190335	2019
3	Avatar	Fox	760507625	2009
4	Black Panther	BV	700059566	2018
5	Avengers: Infinity War	BV	678815482	2018

Layout

Field Name	Type	Info
rank	string	unique identifier of the title
title	string	the type/format of the title
studio	string	Company
lifetime_gross	int	
year	YYYY	release year

Data Preparation



เนื่องจากข้อมูลจาก IMDb มีข้อมูลประเภทของภาพยนตร์ (genres)

ส่วนข้อมูลจาก Box Office Mojo มีข้อมูลรายได้ของภาพยนตร์ (Leftime_gross)

แต่ทั้ง2 data source ไม่มี key เพื่อเชื่อมโยงกันได้ แต่เนื่องจากต้องการนำข้อมูลมาวิเคราะห์เกี่ยวข้องกับความนิยมของภาพยนตร์นั้นๆได้ จึงต้องทำการเชื่อมโยง Data source ด้วยการเปรียบเทียบชื่อของภาพยนตร์ (title)

โดยใช้ field primaryTitle (IMDb) กับ field title (Boxofficemojo)

IMDb

	Field Name	Type
1	tconst	string
2	titleType	string
3	**primaryTitle**	**string**
4	originalTitle	string
5	isAdult	bool
6	startYear	YYYY
7	endYear	YYYY
8	runtimeMinutes	int
9	genres	string array

Boxofficemojo

	Field Name	Type
1	rank	string
2	**title**	**string**
3	studio	string
4	lifetime_gross	int
5	year	YYYY

ขั้นตอน

1. สร้าง file tconst_list.csv เพื่อดึงเอาเฉพาะข้อมูลที่จำเป็นให้ไปใช้ process สำหรับการ เปรียบเทียบ

tconst	primaryTitle	startYear
tt0000009	Miss Jerry	1894
tt0000502	Bohemios	1905
tt0000574	The Story of the Kelly Gang	1906
tt0000591	The Prodigal Son	1907
tt0000615	Robbery Under Arms	1907

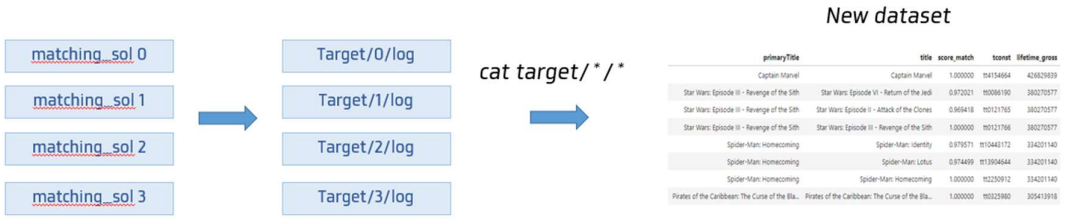
tconst_list.csv

2. ใช้ Script matching_sol.py ซึ่งจะใช้ field primaryTitle เปรียบเทียบ field title (Boxofficemojo)

Script matching_sol.py สามารถ run เป็น Multiprocessing ได้เพื่อรองรับสำหรับข้อมูลขนาดใหญ่ โดยจะแบ่งงานกันทำในแต่ละ Process โดยแบ่งแต่ละ Row ในfile เป็น key 0 – 3 ซึ่งจะ run process 4 process โดยแต่ละ process จะรับผิดชอบ ตาม key

Boxofficemojo

	rank	title	studio	lifetime_gross	year
Key 0	1	Star Wars: The Force Awakens	BV	936662225	2015
Key 1	2	Avengers: Endgame	BV	857190335	2019
Key 2	3	Avatar	Fox	760507625	2009
Key 3	4	Black Panther	BV	700059566	2018
Key 0	5	Avengers: Infinity War	BV	678815482	2018



3. ได้ Data Set ที่มีทั้ง tconst และ lifetime_gross ตามรูป

primaryTitle	title	score_match	tconst	lifetime_gross
Captain Marvel	Captain Marvel	1.000000	tt4154664	426829839
Star Wars: Episode III - Revenge of the Sith	Star Wars: Episode VI - Return of the Jedi	0.972021	tt0086190	380270577
Star Wars: Episode III - Revenge of the Sith	Star Wars: Episode II - Attack of the Clones	0.969418	tt0121765	380270577
Star Wars: Episode III - Revenge of the Sith	Star Wars: Episode III - Revenge of the Sith	1.000000	tt0121766	380270577
Spider-Man: Homecoming	Spider-Man: Identity	0.979571	tt10443172	334201140
Spider-Man: Homecoming	Spider-Man: Lotus	0.974499	tt13904644	334201140
Spider-Man: Homecoming	Spider-Man: Homecoming	1.000000	tt2250912	334201140
Pirates of the Caribbean: The Curse of the Bla...	Pirates of the Caribbean: The Curse of the Bla...	1.000000	tt0325980	305413918

การเปรียบเทียบชื่อของภาพยนตร์ (title) โดยใช้ spaCy open-source library Natural Language Procession (NLP) in Python. โดยใช้ primaryTitle เปรียบเทียบกับ title ซึ่ง Lib spaCy จะให้เป็น **Score match** ออกมา โดย Score จะอยู่ในช่วง 0 – 1.0 **ซึ่งจากการทดสอบ ยังไม่เป็นที่น่าพอใจ** เนื่องจาก หาก Score ไม่เท่ากับ **1.0** ก็มีโอกาสสูงที่จะเป็นหนังคนละเรื่อง

4. จากข้อ3 จะได้ Data Set ใหม่ ซึ่งสามารถนำ tconst ไป merge รวมกับ Data Set IMDb ได้แล้ว

tconst	titleType	primaryTitle	originalTitle	isAdult	startYear	endYear	runtimeMinutes	genres	averageRating	numVotes	lifetime_gross
tt0001184	movie	Don Juan de Serrallonga	Don Juan de Serrallonga	0	1910	\N	58	Adventure,Drama	3.9	20	22150451
tt0002405	movie	Oliver Twist	Oliver Twist	0	1912	\N	\N	Drama	5.2	38	2080321
tt0002461	movie	The Life and Death of King Richard III	Richard III	0	1912	\N	55	Drama	5.7	287	2684904
tt0002767	movie	The Count of Monte Cristo	The Count of Monte Cristo	0	1913	\N	69	Drama,History	6.0	63	54234062
tt0003337	movie	Robin Hood	Robin Hood	0	1913	\N	\N	Adventure	6.1	52	105269730

Analyze

Sort by lifetine_gross.

```
df_full_sheet.loc[df_full_sheet['startYear'] == 2015,"tconst":]
        .sort_values(by=['lifetime_gross'], ascending=False).iloc[0:20].reset_index(drop=True)
```

	tconst	titleType	primaryTitle	originalTitle	isAdult	startYear	endYear	runtimeMinutes	genres	averageRating	numVotes	lifetime_gross
0	tt2488496	movie	Star Wars: Episode VII - The Force Awakens	Star Wars: Episode VII - The Force Awakens	0	2015	\N	138	Action,Adventure,Sci-Fi	7.8	924246	936662225
1	tt0369610	movie	Jurassic World	Jurassic World	0	2015	\N	124	Action,Adventure,Sci-Fi	6.9	639828	652270625
2	tt2395427	movie	Avengers: Age of Ultron	Avengers: Age of Ultron	0	2015	\N	141	Action,Adventure,Sci-Fi	7.3	854411	459005868
3	tt2820852	movie	Furious 7	Fast & Furious 7	0	2015	\N	137	Action,Crime,Thriller	7.1	389065	353007020
4	tt2293640	movie	Minions	Minions	0	2015	\N	91	Adventure,Animation,Comedy	6.4	238379	336045770
5	tt1951266	movie	The Hunger Games: Mockingjay - Part 2	The Hunger Games: Mockingjay - Part 2	0	2015	\N	137	Action,Adventure,Sci-Fi	6.5	324793	281723902

Sort by averageRating.

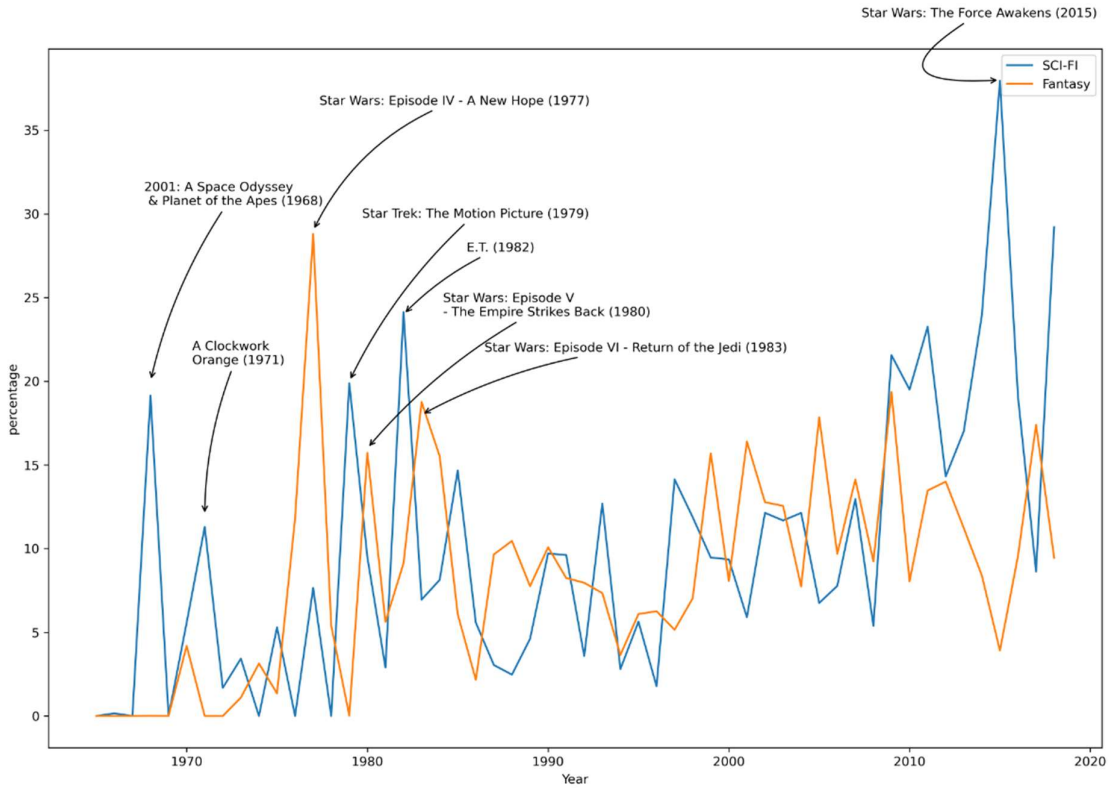
```
df_full_sheet.loc[df_full_sheet['startYear'] == 2015,"tconst":] \
        .sort_values(by=['averageRating','lifetime_gross'], ascending=False).iloc[0:20].reset_index(drop=True)
```

	tconst	titleType	primaryTitle	originalTitle	isAdult	startYear	endYear	runtimeMinutes	genres	averageRating	numVotes	lifetime_gross
0	tt3029962	movie	Brotherhood of the Popcorn	Brotherhood of the Popcorn	0	2015	\N	84	Biography,Documentary,History	9.3	26	11260096
1	tt4023328	movie	In a Perfect World	In a Perfect World	0	2015	\N	75	Documentary,Drama,News	8.8	43	2963902
2	tt4023328	movie	In a Perfect World	In a Perfect World	0	2015	\N	75	Documentary,Drama,News	8.8	43	1008098
3	tt3379352	movie	Mully	Mully	0	2015	\N	81	Adventure,Biography,Documentary	8.3	380	1489771
4	tt4422656	movie	Can You Dig This	Can You Dig This	0	2015	\N	80	Documentary,Drama,Family	8.3	60	15909
5	tt1392190	movie	Mad Max: Fury Road	Mad Max: Fury Road	0	2015	\N	120	Action,Adventure,Sci-Fi	8.1	991314	154058340

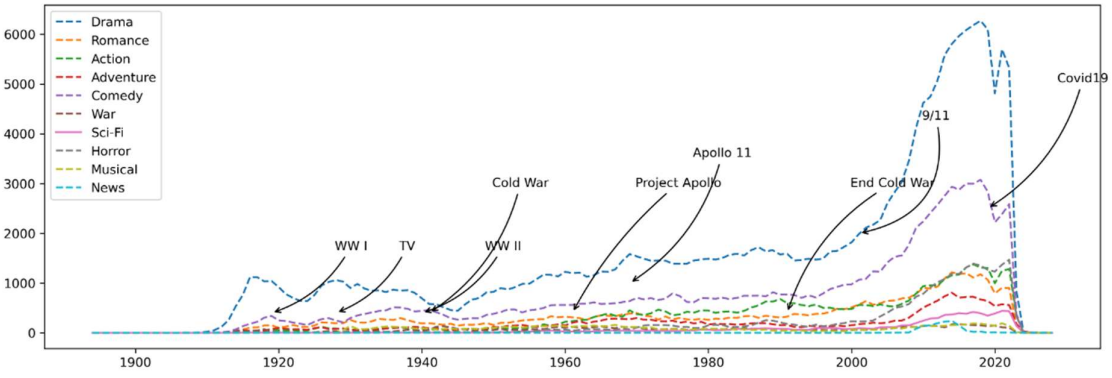
การ Sort by averageRating จากที่ตรวจสอบเป็น Score ที่ได้จาก Comment ของ user IMDb ซึ่งไม่ได้สะท้อน ความนิยมในช่วงเวลาที่ภาพยนตร์ออกฉาย

Tell the story

Star war เป็น pop culture

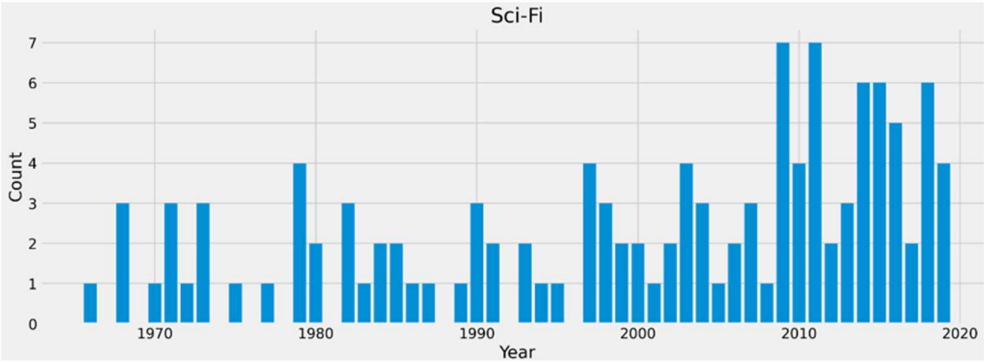


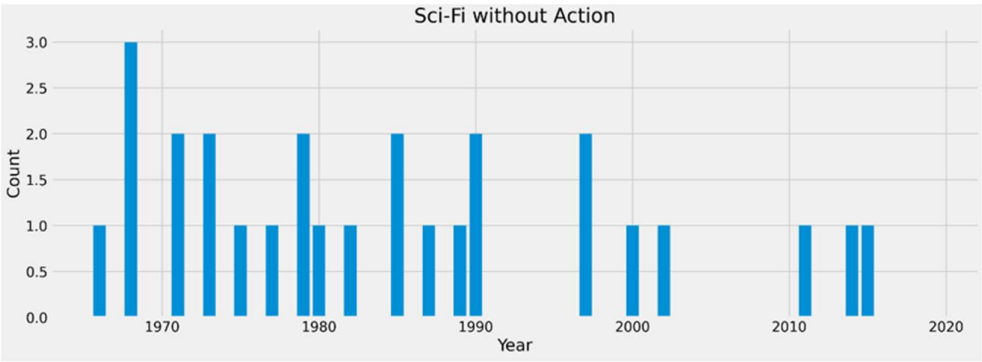
ภาพยนตร์แนว Drama มักได้รับการสร้างออกมา



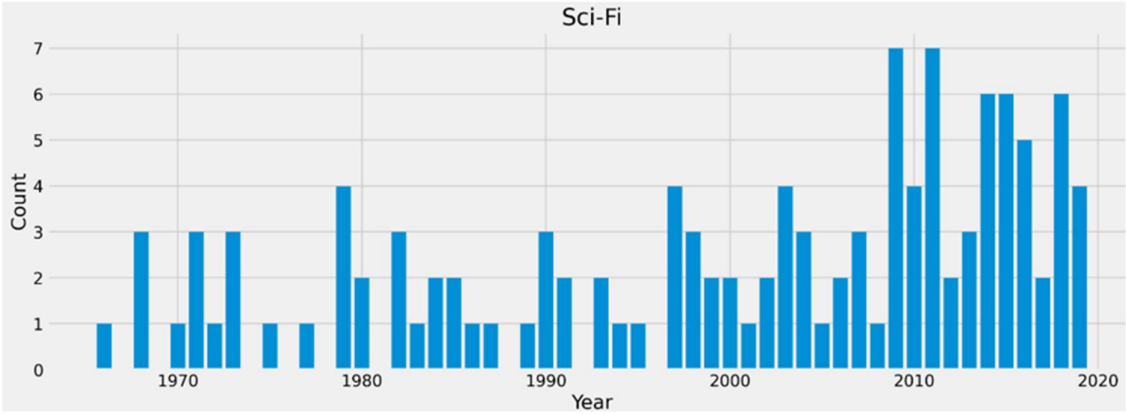
ภาพยนตร์ที่ได้รายได้เยอะ 20 อันดับ ตั้งแต่ 1965 – 2020

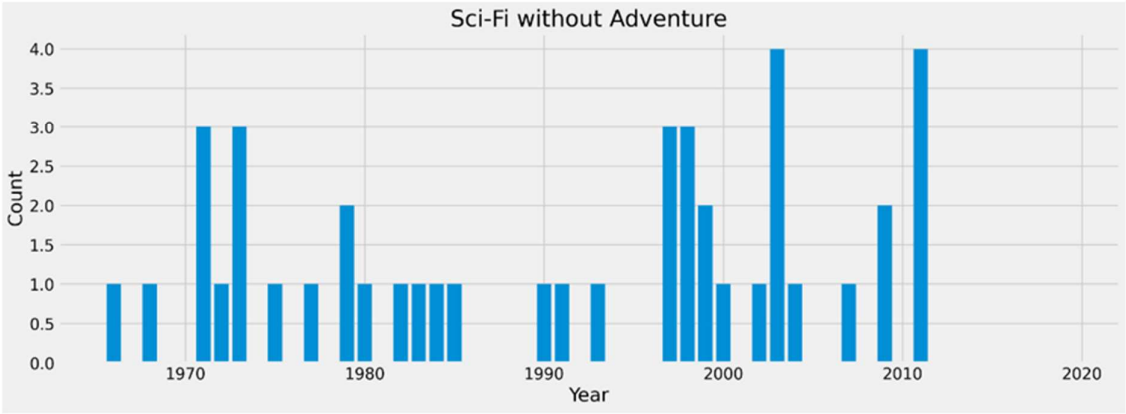
เปรียบเทียบ ภาพยนตร์แนว Sci-Fi จะพบว่าแนว Sci-Fi Action ได้รับความนิยมสูงในช่วงตั้งแต่ปี 1990 เป็นต้นไป





และ Sci-Fi Adventure ได้รับความนิยมสูงในช่วงตั้งแต่ปี 2010 เป็นต้นไป





1990 - 1999

	primaryTitle	startYear	genres	lifetime_gross
0	Star Wars: Episode I - The Phantom Menace	1999	Action,Adventure,Fantasy	474544677
1	The Lion King	1994	Adventure,Animation,Drama	422783777
2	Jurassic Park	1993	Action,Adventure,Sci-Fi	402828120
3	Forrest Gump	1994	Drama,Romance	330455270
4	The Sixth Sense	1999	Drama,Mystery,Thriller	293506292
5	Home Alone	1990	Comedy,Family	285761243
6	Men in Black	1997	Action,Adventure,Comedy	250690539
7	Toy Story 2	1999	Adventure,Animation,Comedy	245852179
8	Brave	1994	Drama,Music	237283207
9	The Lost World: Jurassic Park	1997	Action,Adventure,Sci-Fi	229086679

2000 - 2009

	primaryTitle	startYear	genres	lifetime_gross
0	Avatar	2009	Action,Adventure,Fantasy	760507625
1	The Dark Knight	2008	Action,Crime,Drama	535234033
2	Shrek 2	2004	Adventure,Animation,Comedy	441226247
3	Pirates of the Caribbean: Dead Man's Chest	2006	Action,Adventure,Fantasy	423315812
4	Spider-Man	2002	Action,Adventure,Sci-Fi	403706375
5	Transformers: Revenge of the Fallen	2009	Action,Adventure,Sci-Fi	402111870
6	Frozen	2005	Thriller	400738009
7	Finding Nemo	2003	Adventure,Animation,Comedy	380843261
8	Star Wars: Episode III - Revenge of the Sith	2005	Action,Adventure,Fantasy	380270577
9	The Lord of the Rings: The Return of the King	2003	Action,Adventure,Drama	377845905

2010 - 2019

	primaryTitle	startYear	genres	lifetime_gross
0	Star Wars: Episode VII - The Force Awakens	2015	Action,Adventure,Sci-Fi	936662225
1	Avengers: Endgame	2019	Action,Adventure,Drama	857190335
2	Black Panther	2018	Action,Adventure,Sci-Fi	700059566
3	Avengers: Infinity War	2018	Action,Adventure,Sci-Fi	678815482
4	Jurassic World	2015	Action,Adventure,Sci-Fi	652270625
5	Incredibles 2	2018	Action,Adventure,Animation	608581744
6	Finding Dory	2016	Adventure,Animation,Comedy	486295561
7	Avengers: Age of Ultron	2015	Action,Adventure,Sci-Fi	459005868
8	The Dark Knight Rises	2012	Action,Drama	448139099
9	Captain Marvel	2019	Action,Adventure,Sci-Fi	426829839