Movie Recommender System Project | Content Based Recommender System

In [1]:
```
import pandas as pd
```

Steps:

- Gathering and Handling data
- EDA - Data preprocessing
- Text preprocessing
- Top 5 Movie selection logic
- Saving model using Joblib
- Website using streamlit

Gathering and Handling data

We have 2 CSV to work tmdb_5000_movies.csv and tmdb_5000_credits.csv

In [2]:
```
movies=pd.read_csv('tmdb_5000_movies.csv')
```

In [3]:
```
movies.head(1)
```

Out[3]:

	budget	genres	homepage	id	keywords	original_language	original_title
0	237000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.avatarmovie.com/	19995	[{"id": 1463, "name": "culture clash"}, {"id":...	en	Avatar

In [4]:
```
credits=pd.read_csv('tmdb_5000_credits.csv')
```

In [5]:
```
credits.head(1)
```

Out[5]:

	movie_id	title	cast	crew
0	19995	Avatar	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...

In order to understand better, we will merge both csv using title

In [7]:
```
movies_merged=movies.merge(credits,on='title',how='inner')
```

In [8]:
```
movies_merged.shape
```

Out[8]:
```
(4809, 23)
```

In [9]:
```
movies_merged.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4809 entries, 0 to 4808
Data columns (total 23 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   budget                4809 non-null   int64  
 1   genres                4809 non-null   object 
 2   homepage              1713 non-null   object 
 3   id                    4809 non-null   int64  
 4   keywords              4809 non-null   object 
 5   original_language     4809 non-null   object 
 6   original_title        4809 non-null   object 
 7   overview              4806 non-null   object 
 8   popularity            4809 non-null   float64
 9   production_companies  4809 non-null   object 
 10  production_countries  4809 non-null   object 
 11  release_date          4808 non-null   object 
 12  revenue               4809 non-null   int64  
 13  runtime               4807 non-null   float64
 14  spoken_languages      4809 non-null   object 
 15  status                4809 non-null   object 
 16  tagline               3965 non-null   object 
 17  title                 4809 non-null   object 
 18  vote_average          4809 non-null   float64
 19  vote_count            4809 non-null   int64  
 20  movie_id              4809 non-null   int64  
 21  cast                  4809 non-null   object 
 22  crew                  4809 non-null   object 
dtypes: float64(3), int64(5), object(15)
memory usage: 901.7+ KB
```

Finalized the required column which help us in working with content/text. We are selecting below columns.

In [10]:
```
df_movies=movies_merged[['genres','keywords','title','overview','movie_id','cast','cre
```

In [11]:
```
df_movies.head()
```

Out[11]:

	genres	keywords	title	overview	movie_id	cast	crew
0	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	[{"id": 1463, "name": "culture clash"}, {"id":...	Avatar	In the 22nd century, a paraplegic Marine is di...	19995	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
1	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	285	[{"cast_id": 4, "character": "Captain Jack Spa...	[{"credit_id": "52fe4232c3a36847f800b579", "de...
2	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	[{"id": 470, "name": "spy"}, {"id": 818, "name...	Spectre	A cryptic message from Bond's past sends him o...	206647	[{"cast_id": 1, "character": "James Bond", "cr...	[{"credit_id": "54805967c3a36829b5002c41", "de...
3	[{"id": 28, "name": "Action"}, {"id": 80, "nam...	[{"id": 849, "name": "dc comics"}, {"id": 853,...	The Dark Knight Rises	Following the death of District Attorney Harve...	49026	[{"cast_id": 2, "character": "Bruce Wayne / Ba...	[{"credit_id": "52fe4781c3a36847f81398c3", "de...
4	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	[{"id": 818, "name": "based on novel"}, {"id":...	John Carter	John Carter is a war-weary, former military ca...	49529	[{"cast_id": 5, "character": "John Carter", "c...	[{"credit_id": "52fe479ac3a36847f813eaa3", "de...

EDA - Data preprocessing

In [12]:
```
df_movies.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4809 entries, 0 to 4808
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   genres    4809 non-null   object
 1   keywords  4809 non-null   object
 2   title     4809 non-null   object
 3   overview  4806 non-null   object
 4   movie_id  4809 non-null   int64 
 5   cast      4809 non-null   object
 6   crew      4809 non-null   object
dtypes: int64(1), object(6)
memory usage: 300.6+ KB
```

In features genres,keywords,cast and crew, data is present in list of dictionary, so need to write function to get the name value

In [13]:
```
df_movies['genres'][0]
```

Out[13]:
```
'[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 878, "name": "Science Fiction"}]'
```

If we see above first output of genres column, it shows a string of list that can be handled by ast.

In [14]:
```
import ast
```

In [15]:
```
def converter(input_object):
    temp_list = []
    for i in ast.literal_eval(input_object):
        temp_list.append(i['name'])

    return temp_list
```

In [16]:
```
converter('[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14,
```

Out[16]:
```
['Action', 'Adventure', 'Fantasy', 'Science Fiction']
```

Converter function is giving us the required values from the features.

Now Fetch out genres words from genres columns and append in the same dataframe

In [18]:
```
df_movies['genres']=df_movies['genres'].apply(converter)
```

```
C:\Users\rupeshv\AppData\Local\Temp\ipykernel_11104\2749150962.py:1: SettingWithCopyW
arning:
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/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  df_movies['genres']=df_movies['genres'].apply(converter)
```

In [19]:
```
df_movies.head(5)
```

Out[19]:

	genres	keywords	title	overview	movie_id	cast	crew
0	[Action, Adventure, Fantasy, Science Fiction]	[{"id": 1463, "name": "culture clash"}, {"id":...	Avatar	In the 22nd century, a paraplegic Marine is di...	19995	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
1	[Adventure, Fantasy, Action]	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	285	[{"cast_id": 4, "character": "Captain Jack Spa...	[{"credit_id": "52fe4232c3a36847f800b579", "de...
2	[Action, Adventure, Crime]	[{"id": 470, "name": "spy"}, {"id": 818, "name...	Spectre	A cryptic message from Bond's past sends him o...	206647	[{"cast_id": 1, "character": "James Bond", "cr...	[{"credit_id": "54805967c3a36829b5002c41", "de...
3	[Action, Crime, Drama, Thriller]	[{"id": 849, "name": "dc comics"}, {"id": 853,...	The Dark Knight Rises	Following the death of District Attorney Harve...	49026	[{"cast_id": 2, "character": "Bruce Wayne / Ba...	[{"credit_id": "52fe4781c3a36847f81398c3", "de...
4	[Action, Adventure, Science Fiction]	[{"id": 818, "name": "based on novel"}, {"id":...	John Carter	John Carter is a war-weary, former military ca...	49529	[{"cast_id": 5, "character": "John Carter", "c...	[{"credit_id": "52fe479ac3a36847f813eaa3", "de...

Same apply this converter logic to keywords

In [20]:
```
df_movies['keywords']=df_movies['keywords'].apply(converter)
```

```
C:\Users\rupeshv\AppData\Local\Temp\ipykernel_11104\1298947391.py:2: SettingWithCopyW
arning:
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/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  df_movies['keywords']=df_movies['keywords'].apply(converter)
```

In [21]:
```
df_movies.head(5)
```

Out[21]:

	genres	keywords	title	overview	movie_id	cast	crew
0	[Action, Adventure, Fantasy, Science Fiction]	[culture clash, future, space war, space colon...	Avatar	In the 22nd century, a paraplegic Marine is di...	19995	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
1	[Adventure, Fantasy, Action]	[ocean, drug abuse, exotic island, east india ...	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	285	[{"cast_id": 4, "character": "Captain Jack Spa...	[{"credit_id": "52fe4232c3a36847f800b579", "de...
2	[Action, Adventure, Crime]	[spy, based on novel, secret agent, sequel, mi...	Spectre	A cryptic message from Bond's past sends him o...	206647	[{"cast_id": 1, "character": "James Bond", "cr...	[{"credit_id": "54805967c3a36829b5002c41", "de...
3	[Action, Crime, Drama, Thriller]	[dc comics, crime fighter, terrorist, secret i...	The Dark Knight Rises	Following the death of District Attorney Harve...	49026	[{"cast_id": 2, "character": "Bruce Wayne / Ba...	[{"credit_id": "52fe4781c3a36847f81398c3", "de...
4	[Action, Adventure, Science Fiction]	[based on novel, mars, medallion, space travel...	John Carter	John Carter is a war-weary, former military ca...	49529	[{"cast_id": 5, "character": "John Carter", "c...	[{"credit_id": "52fe479ac3a36847f813eaa3", "de...

From cast feature, we need top 3 person, below is the function

In [22]:
```
def converter_top3(input_object):
    temp_list = []
    counter = 0
    for i in ast.literal_eval(input_object):
```

```
        if counter != 3:
            temp_list.append(i['name'])
            counter+=1
        else:
            break

    return temp_list
```

In [23]:
```
df_movies['cast']=df_movies['cast'].apply(converter_top3)
```

```
C:\Users\rupeshv\AppData\Local\Temp\ipykernel_11104\1415257880.py:1: SettingWithCopyW
arning:
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/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  df_movies['cast']=df_movies['cast'].apply(converter_top3)
```

In [24]:
```
df_movies.head(5)
```

Out[24]:

	genres	keywords	title	overview	movie_id	cast	crew
0	[Action, Adventure, Fantasy, Science Fiction]	[culture clash, future, space war, space colon...	Avatar	In the 22nd century, a paraplegic Marine is di...	19995	[Sam Worthington, Zoe Saldana, Sigourney Weaver]	[{"credit_id": "52fe48009251416c750aca23", "de...
1	[Adventure, Fantasy, Action]	[ocean, drug abuse, exotic island, east india ...	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	285	[Johnny Depp, Orlando Bloom, Keira Knightley]	[{"credit_id": "52fe4232c3a36847f800b579", "de...
2	[Action, Adventure, Crime]	[spy, based on novel, secret agent, sequel, mi...	Spectre	A cryptic message from Bond’s past sends him o...	206647	[Daniel Craig, Christoph Waltz, Léa Seydoux]	[{"credit_id": "54805967c3a36829b5002c41", "de...
3	[Action, Crime, Drama, Thriller]	[dc comics, crime fighter, terrorist, secret i...	The Dark Knight Rises	Following the death of District Attorney Harve...	49026	[Christian Bale, Michael Caine, Gary Oldman]	[{"credit_id": "52fe4781c3a36847f81398c3", "de...
4	[Action, Adventure, Science Fiction]	[based on novel, mars, medallion, space travel...	John Carter	John Carter is a war-weary, former military ca...	49529	[Taylor Kitsch, Lynn Collins, Samantha Morton]	[{"credit_id": "52fe479ac3a36847f813eaa3", "de...

From the crew feature we are going to select only director name and so function converter_director is created accordingly

In [25]:
```
df_movies['crew'][0]
```

Out[25]:

'[{"credit_id": "52fe48009251416c750aca23", "department": "Editing", "gender": 0, "id": 1721, "job": "Editor", "name": "Stephen E. Rivkin"}, {"credit_id": "539c47ecc3a36810e3001f87", "department": "Art", "gender": 2, "id": 496, "job": "Production Design", "name": "Rick Carter"}, {"credit_id": "54491c89c3a3680fb4001cf7", "department": "Sound", "gender": 0, "id": 900, "job": "Sound Designer", "name": "Christopher Boyes"}, {"credit_id": "54491cb70e0a267480001bd0", "department": "Sound", "gender": 0, "id": 900, "job": "Supervising Sound Editor", "name": "Christopher Boyes"}, {"credit_id": "539c4a4cc3a36810c9002101", "department": "Production", "gender": 1, "id": 1262, "job": "Casting", "name": "Mali Finn"}, {"credit_id": "5544ee3b925141499f0008fc", "department": "Sound", "gender": 2, "id": 1729, "job": "Original Music Composer", "name": "James Horner"}, {"credit_id": "52fe48009251416c750ac9c3", "department": "Directing", "gender": 2, "id": 2710, "job": "Director", "name": "James Cameron"}, {"credit_id": "52fe48009251416c750ac9d9", "department": "Writing", "gender": 2, "id": 2710, "job": "Writer", "name": "James Cameron"}, {"credit_id": "52fe48009251416c750aca17", "department": "Editing", "gender": 2, "id": 2710, "job": "Editor", "name": "James Cameron"}, {"credit_id": "52fe48009251416c750aca29", "department": "Production", "gender": 2, "id": 2710, "job": "Producer", "name": "James Cameron"}, {"credit_id": "52fe48009251416c750aca3f", "department": "Writing", "gender": 2, "id": 2710, "job": "Screenplay", "name": "James Cameron"}, {"credit_id": "539c4987c3a36810ba0021a4", "department": "Art", "gender": 2, "id": 7236, "job": "Art Direction", "name": "Andrew Menzies"}, {"credit_id": "549598c3c3a3686ae9004383", "department": "Visual Effects", "gender": 0, "id": 6690, "job": "Visual Effects Producer", "name": "Jill Brooks"}, {"credit_id": "52fe48009251416c750aca4b", "department": "Production", "gender": 1, "id": 6347, "job": "Casting", "name": "Margery Simkin"}, {"credit_id": "570b6f419251417da70032fe", "department": "Art", "gender": 2, "id": 6878, "job": "Supervising Art Director", "name": "Kevin Ishioka"}, {"credit_id": "5495a0fac3a3686ae9004468", "department": "Sound", "gender": 0, "id": 6883, "job": "Music Editor", "name": "Dick Bernstein"}, {"credit_id": "54959706c3a3686af3003e81", "department": "Sound", "gender": 0, "id": 8159, "job": "Sound Effects Editor", "name": "Shannon Mills"}, {"credit_id": "54491d58c3a3680fb1001ccb", "department": "Sound", "gender": 0, "id": 8160, "job": "Foley", "name": "Dennie Thorpe"}, {"credit_id": "54491d6cc3a3680fa5001b2c", "department": "Sound", "gender": 0, "id": 8163, "job": "Foley", "name": "Jana Vance"}, {"credit_id": "52fe48009251416c750aca57", "department": "Costume & Make-Up", "gender": 1, "id": 8527, "job": "Costume Design", "name": "Deborah Lynn Scott"}, {"credit_id": "52fe48009251416c750aca2f", "department": "Production", "gender": 2, "id": 8529, "job": "Producer", "name": "Jon Landau"}, {"credit_id": "539c4937c3a36810ba002194", "department": "Art", "gender": 0, "id": 9618, "job": "Art Direction", "name": "Sean Haworth"}, {"credit_id": "539c49b6c3a36810c10020e6", "department": "Art", "gender": 1, "id": 12653, "job": "Set Decoration", "name": "Kim Sinclair"}, {"credit_id": "570b6f2f9251413a0e00020d", "department": "Art", "gender": 1, "id": 12653, "job": "Supervising Art Director", "name": "Kim Sinclair"}, {"credit_id": "54491a6c0e0a26748c001b19", "department": "Art", "gender": 2, "id": 14350, "job": "Set Designer", "name": "Richard F. Mays"}, {"credit_id": "56928cf4c3a3684cff0025c4", "department": "Production", "gender": 1, "id": 20294, "job": "Executive Producer", "name": "Laeta Kalogridis"}, {"credit_id": "52fe48009251416c750aca51", "department": "Costume & Make-Up", "gender": 0, "id": 17675, "job": "Costume Design", "name": "Mayes C. Rubeo"}, {"credit_id": "52fe48009251416c750aca11", "department": "Camera", "gender": 2, "id": 18265, "job": "Director of Photography", "name": "Mauro Fiore"}, {"credit_id": "5449194d0e0a26748f001b39", "department": "Art", "gender": 0, "id": 42281, "job": "Set Designer", "name": "Scott Herbertson"}, {"credit_id": "52fe48009251416c750aca05", "department": "Crew", "gender": 0, "id": 42288, "job": "Stunts", "name": "Woody Schultz"}, {"credit_id": "5592aefb92514152de0010f5", "department": "Costume & Make-Up", "gender": 0, "id": 29067, "job": "Makeup Artist", "name": "Linda DeVetta"}, {"credit_id": "5592afa492514152de00112c", "department": "Costume & Make-Up", "gender": 0, "id": 29067, "job": "Hairstylist", "name": "Linda DeVetta"}, {"credit_id": "54959ed592514130fc002e5d", "department": "Camera", "gender": 2, "id": 33302, "job": "Camera Operator", "name": "Richard Bluck"}, {"credit_id": "539c4891c3a36810ba002147", "department": "Art", "gender": 2, "id": 33303, "job": "Art Direction", "name": "Simon Bright"}, {"credit_id": "54959c069251417a81001f3a", "department": "Visual Effects", "gender": 0, "id": 113145, "job": "Visual Effects Supervisor", "name": "Richard Martin"}, {"credit_id": "54959a0dc3a3680ff5002c8d", "department": "C

rew", "gender": 2, "id": 58188, "job": "Visual Effects Editor", "name": "Steve R. Moo
re"}, {"credit_id": "52fe48009251416c750aca1d", "department": "Editing", "gender": 2,
"id": 58871, "job": "Editor", "name": "John Refoua"}, {"credit_id": "54491a4dc3a3680f
c30018ca", "department": "Art", "gender": 0, "id": 92359, "job": "Set Designer", "nam
e": "Karl J. Martin"}, {"credit_id": "52fe48009251416c750aca35", "department": "Camer
a", "gender": 1, "id": 72201, "job": "Director of Photography", "name": "Chiling Li
n"}, {"credit_id": "52fe48009251416c750ac9ff", "department": "Crew", "gender": 0, "i
d": 89714, "job": "Stunts", "name": "Ilram Choi"}, {"credit_id": "54959c529251416e2b0
04394", "department": "Visual Effects", "gender": 2, "id": 93214, "job": "Visual Effe
cts Supervisor", "name": "Steven Quale"}, {"credit_id": "54491edf0e0a267489001c37",
"department": "Crew", "gender": 1, "id": 122607, "job": "Dialect Coach", "name": "Car
la Meyer"}, {"credit_id": "539c485bc3a368653d001a3a", "department": "Art", "gender":
2, "id": 132585, "job": "Art Direction", "name": "Nick Bassett"}, {"credit_id": "539c
4903c3a368653d001a74", "department": "Art", "gender": 0, "id": 132596, "job": "Art Di
rection", "name": "Jill Cormack"}, {"credit_id": "539c4967c3a368653d001a94", "departm
ent": "Art", "gender": 0, "id": 132604, "job": "Art Direction", "name": "Andy McLare
n"}, {"credit_id": "52fe48009251416c750aca45", "department": "Crew", "gender": 0, "i
d": 236696, "job": "Motion Capture Artist", "name": "Terry Notary"}, {"credit_id": "5
4959e02c3a3680fc60027d2", "department": "Crew", "gender": 2, "id": 956198, "job": "St
unt Coordinator", "name": "Garrett Warren"}, {"credit_id": "54959ca3c3a3686ae300438
c", "department": "Visual Effects", "gender": 2, "id": 957874, "job": "Visual Effects
Supervisor", "name": "Jonathan Rothbart"}, {"credit_id": "570b6f519251412c74001b2f",
"department": "Art", "gender": 0, "id": 957889, "job": "Supervising Art Director", "n
ame": "Stefan Dechant"}, {"credit_id": "570b6f62c3a3680b77007460", "department": "Ar
t", "gender": 2, "id": 959555, "job": "Supervising Art Director", "name": "Todd Chern
iawsky"}, {"credit_id": "539c4a3ac3a36810da0021cc", "department": "Production", "gend
er": 0, "id": 1016177, "job": "Casting", "name": "Miranda Rivers"}, {"credit_id": "53
9c482cc3a36810c1002062", "department": "Art", "gender": 0, "id": 1032536, "job": "Pro
duction Design", "name": "Robert Stromberg"}, {"credit_id": "539c4b65c3a36810c900212
5", "department": "Costume & Make-Up", "gender": 2, "id": 1071680, "job": "Costume De
sign", "name": "John Harding"}, {"credit_id": "54959e6692514130fc002e4e", "departmen
t": "Camera", "gender": 0, "id": 1177364, "job": "Steadicam Operator", "name": "Rober
to De Angelis"}, {"credit_id": "539c49f1c3a368653d001aac", "department": "Costume & M
ake-Up", "gender": 2, "id": 1202850, "job": "Makeup Department Head", "name": "Mike S
mithson"}, {"credit_id": "5495999ec3a3686ae100460c", "department": "Visual Effects",
"gender": 0, "id": 1204668, "job": "Visual Effects Producer", "name": "Alain Lalann
e"}, {"credit_id": "54959cdfc3a3681153002729", "department": "Visual Effects", "gende
r": 0, "id": 1206410, "job": "Visual Effects Supervisor", "name": "Lucas Salton"},
{"credit_id": "549596239251417a81001eae", "department": "Crew", "gender": 0, "id": 12
34266, "job": "Post Production Supervisor", "name": "Janace Tashjian"}, {"credit_id":
"54959c859251416e1e003efe", "department": "Visual Effects", "gender": 0, "id": 127193
2, "job": "Visual Effects Supervisor", "name": "Stephen Rosenbaum"}, {"credit_id": "5
592af28c3a368775a00105f", "department": "Costume & Make-Up", "gender": 0, "id": 13100
64, "job": "Makeup Artist", "name": "Frankie Karena"}, {"credit_id": "539c4adfc3a3681
0e300203b", "department": "Costume & Make-Up", "gender": 1, "id": 1319844, "job": "Co
stume Supervisor", "name": "Lisa Lovaas"}, {"credit_id": "54959b579251416e2b004371",
"department": "Visual Effects", "gender": 0, "id": 1327028, "job": "Visual Effects Su
pervisor", "name": "Jonathan Fawkner"}, {"credit_id": "539c48a7c3a36810b5001fa7", "de
partment": "Art", "gender": 0, "id": 1330561, "job": "Art Direction", "name": "Robert
Bavin"}, {"credit_id": "539c4a71c3a36810da0021e0", "department": "Costume & Make-Up",
"gender": 0, "id": 1330567, "job": "Costume Supervisor", "name": "Anthony Almaraz"},
{"credit_id": "539c4a8ac3a36810ba0021e4", "department": "Costume & Make-Up", "gende
r": 0, "id": 1330570, "job": "Costume Supervisor", "name": "Carolyn M. Fenton"}, {"cr
edit_id": "539c4ab6c3a36810da0021f0", "department": "Costume & Make-Up", "gender": 0,
"id": 1330574, "job": "Costume Supervisor", "name": "Beth Koenigsberg"}, {"credit_i
d": "54491ab70e0a267480001ba2", "department": "Art", "gender": 0, "id": 1336191, "jo
b": "Set Designer", "name": "Sam Page"}, {"credit_id": "544919d9c3a3680fc30018bd", "d
epartment": "Art", "gender": 0, "id": 1339441, "job": "Set Designer", "name": "Tex Ka
donaga"}, {"credit_id": "54491cf50e0a267483001b0c", "department": "Editing", "gende
r": 0, "id": 1352422, "job": "Dialogue Editor", "name": "Kim Foscato"}, {"credit_id":

"544919f40e0a26748c001b09", "department": "Art", "gender": 0, "id": 1352962, "job": "Set Designer", "name": "Tammy S. Lee"}, {"credit_id": "5495a115c3a3680ff5002d71", "department": "Crew", "gender": 0, "id": 1357070, "job": "Transportation Coordinator", "name": "Denny Caira"}, {"credit_id": "5495a12f92514130fc002e94", "department": "Crew", "gender": 0, "id": 1357071, "job": "Transportation Coordinator", "name": "James Waitkus"}, {"credit_id": "5495976fc3a36811530026b0", "department": "Sound", "gender": 0, "id": 1360103, "job": "Supervising Sound Editor", "name": "Addison Teague"}, {"credit_id": "54491837c3a3680fb1001c5a", "department": "Art", "gender": 2, "id": 1376887, "job": "Set Designer", "name": "C. Scott Baker"}, {"credit_id": "54491878c3a3680fb4001c9d", "department": "Art", "gender": 0, "id": 1376888, "job": "Set Designer", "name": "Luke Caska"}, {"credit_id": "544918dac3a3680fa5001ae0", "department": "Art", "gender": 0, "id": 1376889, "job": "Set Designer", "name": "David Chow"}, {"credit_id": "544919110e0a267486001b68", "department": "Art", "gender": 0, "id": 1376890, "job": "Set Designer", "name": "Jonathan Dyer"}, {"credit_id": "54491967c3a3680faa001b5e", "department": "Art", "gender": 0, "id": 1376891, "job": "Set Designer", "name": "Joseph Hiura"}, {"credit_id": "54491997c3a3680fb1001c8a", "department": "Art", "gender": 0, "id": 1376892, "job": "Art Department Coordinator", "name": "Rebecca Jellie"}, {"credit_id": "544919ba0e0a26748f001b42", "department": "Art", "gender": 0, "id": 1376893, "job": "Set Designer", "name": "Robert Andrew Johnson"}, {"credit_id": "54491b1dc3a3680faa001b8c", "department": "Art", "gender": 0, "id": 1376895, "job": "Assistant Art Director", "name": "Mike Stassi"}, {"credit_id": "54491b79c3a3680fbb001826", "department": "Art", "gender": 0, "id": 1376897, "job": "Construction Coordinator", "name": "John Villarino"}, {"credit_id": "54491baec3a3680fb4001ce6", "department": "Art", "gender": 2, "id": 1376898, "job": "Assistant Art Director", "name": "Jeffrey Wisniewski"}, {"credit_id": "54491d2fc3a3680fb4001d07", "department": "Editing", "gender": 0, "id": 1376899, "job": "Dialogue Editor", "name": "Cheryl Nardi"}, {"credit_id": "54491d86c3a3680fa5001b2f", "department": "Editing", "gender": 0, "id": 1376901, "job": "Dialogue Editor", "name": "Marshall Winn"}, {"credit_id": "54491d9dc3a3680faa001bb0", "department": "Sound", "gender": 0, "id": 1376902, "job": "Supervising Sound Editor", "name": "Gwendolyn Yates Whittle"}, {"credit_id": "54491dc10e0a267486001bce", "department": "Sound", "gender": 0, "id": 1376903, "job": "Sound Re-Recording Mixer", "name": "William Stein"}, {"credit_id": "54491f500e0a26747c001c07", "department": "Crew", "gender": 0, "id": 1376909, "job": "Choreographer", "name": "Lula Washington"}, {"credit_id": "549599239251412c4e002a2e", "department": "Visual Effects", "gender": 0, "id": 1391692, "job": "Visual Effects Producer", "name": "Chris Del Conte"}, {"credit_id": "54959d54c3a36831b8001d9a", "department": "Visual Effects", "gender": 2, "id": 1391695, "job": "Visual Effects Supervisor", "name": "R. Christopher White"}, {"credit_id": "54959bdf9251412c4e002a66", "department": "Visual Effects", "gender": 0, "id": 1394070, "job": "Visual Effects Supervisor", "name": "Dan Lemmon"}, {"credit_id": "5495971d92514132ed002922", "department": "Sound", "gender": 0, "id": 1394129, "job": "Sound Effects Editor", "name": "Tim Nielsen"}, {"credit_id": "5592b25792514152cc0011aa", "department": "Crew", "gender": 0, "id": 1394286, "job": "CG Supervisor", "name": "Michael Mulholland"}, {"credit_id": "54959a329251416e2b004355", "department": "Crew", "gender": 0, "id": 1394750, "job": "Visual Effects Editor", "name": "Thomas Nittmann"}, {"credit_id": "54959d6dc3a3686ae9004401", "department": "Visual Effects", "gender": 0, "id": 1394755, "job": "Visual Effects Supervisor", "name": "Edson Williams"}, {"credit_id": "5495a08fc3a3686ae300441c", "department": "Editing", "gender": 0, "id": 1394953, "job": "Digital Intermediate", "name": "Christine Carr"}, {"credit_id": "55402d659251413d6d000249", "department": "Visual Effects", "gender": 0, "id": 1395269, "job": "Visual Effects Supervisor", "name": "John Bruno"}, {"credit_id": "54959e7b9251416e1e003f3e", "department": "Camera", "gender": 0, "id": 1398970, "job": "Steadicam Operator", "name": "David Emmerichs"}, {"credit_id": "54959734c3a3686ae10045e0", "department": "Sound", "gender": 0, "id": 1400906, "job": "Sound Effects Editor", "name": "Christopher Scarabosio"}, {"credit_id": "549595dd92514130fc002d79", "department": "Production", "gender": 0, "id": 1401784, "job": "Production Supervisor", "name": "Jennifer Teves"}, {"credit_id": "549596009251413af70028cc", "department": "Production", "gender": 0, "id": 1401785, "job": "Production Manager", "name": "Brigitte Yorke"}, {"credit_id": "549596e892514130fc002d99", "department": "Sound", "gender": 0, "id": 1401786, "job": "Sound Effects Editor", "name": "Ken Fischer"}, {"credit_id": "549598229251412c4e002a1c", "department": "Crew", "gender": 0, "id": 140178

7, "job": "Special Effects Coordinator", "name": "Iain Hutton"}, {"credit_id": "54959
8349251416e2b00432b", "department": "Crew", "gender": 0, "id": 1401788, "job": "Speci
al Effects Coordinator", "name": "Steve Ingram"}, {"credit_id": "54959905c3a3686ae300
4324", "department": "Visual Effects", "gender": 0, "id": 1401789, "job": "Visual Eff
ects Producer", "name": "Joyce Cox"}, {"credit_id": "5495994b92514132ed002951", "depa
rtment": "Visual Effects", "gender": 0, "id": 1401790, "job": "Visual Effects Produce
r", "name": "Jenny Foster"}, {"credit_id": "549599cbc3a3686ae1004613", "department":
"Crew", "gender": 0, "id": 1401791, "job": "Visual Effects Editor", "name": "Christop
her Marino"}, {"credit_id": "549599f2c3a3686ae100461e", "department": "Crew", "gende
r": 0, "id": 1401792, "job": "Visual Effects Editor", "name": "Jim Milton"}, {"credit
_id": "54959a51c3a3686af3003eb5", "department": "Visual Effects", "gender": 0, "id":
1401793, "job": "Visual Effects Producer", "name": "Cyndi Ochs"}, {"credit_id": "5495
9a7cc3a36811530026f4", "department": "Crew", "gender": 0, "id": 1401794, "job": "Visu
al Effects Editor", "name": "Lucas Putnam"}, {"credit_id": "54959b91c3a3680ff5002cb
4", "department": "Visual Effects", "gender": 0, "id": 1401795, "job": "Visual Effect
s Supervisor", "name": "Anthony \'Max\' Ivins"}, {"credit_id": "54959bb69251412c4e002
a5f", "department": "Visual Effects", "gender": 0, "id": 1401796, "job": "Visual Effe
cts Supervisor", "name": "John Knoll"}, {"credit_id": "54959cbbc3a3686ae3004391", "de
partment": "Visual Effects", "gender": 2, "id": 1401799, "job": "Visual Effects Super
visor", "name": "Eric Saindon"}, {"credit_id": "54959d06c3a3686ae90043f6", "departmen
t": "Visual Effects", "gender": 0, "id": 1401800, "job": "Visual Effects Supervisor",
"name": "Wayne Stables"}, {"credit_id": "54959d259251416e1e003f11", "department": "Vi
sual Effects", "gender": 0, "id": 1401801, "job": "Visual Effects Supervisor", "nam
e": "David Stinnett"}, {"credit_id": "54959db49251413af7002975", "department": "Visua
l Effects", "gender": 0, "id": 1401803, "job": "Visual Effects Supervisor", "name":
"Guy Williams"}, {"credit_id": "54959de4c3a3681153002750", "department": "Crew", "gen
der": 0, "id": 1401804, "job": "Stunt Coordinator", "name": "Stuart Thorp"}, {"credit
_id": "54959ef2c3a3680fc60027f2", "department": "Lighting", "gender": 0, "id": 140180
5, "job": "Best Boy Electric", "name": "Giles Coburn"}, {"credit_id": "54959f07c3a368
0fc60027f9", "department": "Camera", "gender": 2, "id": 1401806, "job": "Still Photog
rapher", "name": "Mark Fellman"}, {"credit_id": "54959f47c3a3681153002774", "departme
nt": "Lighting", "gender": 0, "id": 1401807, "job": "Lighting Technician", "name": "S
cott Sprague"}, {"credit_id": "54959f8cc3a36831b8001df2", "department": "Visual Effec
ts", "gender": 0, "id": 1401808, "job": "Animation Director", "name": "Jeremy Hollobo
n"}, {"credit_id": "54959fa0c3a36831b8001dfb", "department": "Visual Effects", "gende
r": 0, "id": 1401809, "job": "Animation Director", "name": "Orlando Meunier"}, {"cred
it_id": "54959fb6c3a3686af3003f54", "department": "Visual Effects", "gender": 0, "i
d": 1401810, "job": "Animation Director", "name": "Taisuke Tanimura"}, {"credit_id":
"54959fd2c3a36831b8001e02", "department": "Costume & Make-Up", "gender": 0, "id": 140
1812, "job": "Set Costumer", "name": "Lilia Mishel Acevedo"}, {"credit_id": "54959ff9
c3a3686ae300440c", "department": "Costume & Make-Up", "gender": 0, "id": 1401814, "jo
b": "Set Costumer", "name": "Alejandro M. Hernandez"}, {"credit_id": "5495a0ddc3a3686
ae10046fe", "department": "Editing", "gender": 0, "id": 1401815, "job": "Digital Inte
rmediate", "name": "Marvin Hall"}, {"credit_id": "5495a1f7c3a3686ae3004443", "departm
ent": "Production", "gender": 0, "id": 1401816, "job": "Publicist", "name": "Judy All
ey"}, {"credit_id": "5592b29fc3a36869d100002f", "department": "Crew", "gender": 0, "i
d": 1418381, "job": "CG Supervisor", "name": "Mike Perry"}, {"credit_id": "5592b23a92
51415df8001081", "department": "Crew", "gender": 0, "id": 1426854, "job": "CG Supervi
sor", "name": "Andrew Morley"}, {"credit_id": "55491e1192514104c40002d8", "departmen
t": "Art", "gender": 0, "id": 1438901, "job": "Conceptual Design", "name": "Seth Engs
trom"}, {"credit_id": "5525d5809251417276002b06", "department": "Crew", "gender": 0,
"id": 1447362, "job": "Visual Effects Art Director", "name": "Eric Oliver"}, {"credit
_id": "554427ca925141586500312a", "department": "Visual Effects", "gender": 0, "id":
1447503, "job": "Modeling", "name": "Matsune Suzuki"}, {"credit_id": "551906889251415
aab001c88", "department": "Art", "gender": 0, "id": 1447524, "job": "Art Department M
anager", "name": "Paul Tobin"}, {"credit_id": "5592af8492514152cc0010de", "departmen
t": "Costume & Make-Up", "gender": 0, "id": 1452643, "job": "Hairstylist", "name": "R
oxane Griffin"}, {"credit_id": "553d3c109251415852001318", "department": "Lighting",
"gender": 0, "id": 1453938, "job": "Lighting Artist", "name": "Arun Ram-Mohan"}, {"cr
edit_id": "5592af4692514152d5001355", "department": "Costume & Make-Up", "gender": 0,

"id": 1457305, "job": "Makeup Artist", "name": "Georgia Lockhart-Adams"}, {"credit_id": "5592b2eac3a36877470012a5", "department": "Crew", "gender": 0, "id": 1466035, "job": "CG Supervisor", "name": "Thrain Shadbolt"}, {"credit_id": "5592b032c3a36877450015f1", "department": "Crew", "gender": 0, "id": 1483220, "job": "CG Supervisor", "name": "Brad Alexander"}, {"credit_id": "5592b05592514152d80012f6", "department": "Crew", "gender": 0, "id": 1483221, "job": "CG Supervisor", "name": "Shadi Almassizadeh"}, {"credit_id": "5592b090c3a36877570010b5", "department": "Crew", "gender": 0, "id": 1483222, "job": "CG Supervisor", "name": "Simon Clutterbuck"}, {"credit_id": "5592b0dbc3a368774b00112c", "department": "Crew", "gender": 0, "id": 1483223, "job": "CG Supervisor", "name": "Graeme Demmocks"}, {"credit_id": "5592b0fe92514152db0010c1", "department": "Crew", "gender": 0, "id": 1483224, "job": "CG Supervisor", "name": "Adrian Fernandes"}, {"credit_id": "5592b11f9251415df8001059", "department": "Crew", "gender": 0, "id": 1483225, "job": "CG Supervisor", "name": "Mitch Gates"}, {"credit_id": "5592b15dc3a3687745001645", "department": "Crew", "gender": 0, "id": 1483226, "job": "CG Supervisor", "name": "Jerry Kung"}, {"credit_id": "5592b18e925141645a0004ae", "department": "Crew", "gender": 0, "id": 1483227, "job": "CG Supervisor", "name": "Andy Lomas"}, {"credit_id": "5592b1bfc3a368775d0010e7", "department": "Crew", "gender": 0, "id": 1483228, "job": "CG Supervisor", "name": "Sebastian Marino"}, {"credit_id": "5592b2049251415df8001078", "department": "Crew", "gender": 0, "id": 1483229, "job": "CG Supervisor", "name": "Matthias Menz"}, {"credit_id": "5592b27b92514152d800136a", "department": "Crew", "gender": 0, "id": 1483230, "job": "CG Supervisor", "name": "Sergei Nevshupov"}, {"credit_id": "5592b2c3c3a36869e800003c", "department": "Crew", "gender": 0, "id": 1483231, "job": "CG Supervisor", "name": "Philippe Rebours"}, {"credit_id": "5592b317c3a36877470012af", "department": "Crew", "gender": 0, "id": 1483232, "job": "CG Supervisor", "name": "Michael Takarangi"}, {"credit_id": "5592b345c3a36877470012bb", "department": "Crew", "gender": 0, "id": 1483233, "job": "CG Supervisor", "name": "David Weitzberg"}, {"credit_id": "5592b37cc3a368775100113b", "department": "Crew", "gender": 0, "id": 1483234, "job": "CG Supervisor", "name": "Ben White"}, {"credit_id": "573c8e2f9251413f5d000094", "department": "Crew", "gender": 1, "id": 1621932, "job": "Stunts", "name": "Min Windle"}]'

In [26]:
```
def converter_director(input_object):
    temp_list = []
    for i in ast.literal_eval(input_object):
        if i['job'] == 'Director':
            temp_list.append(i['name'])
            break

    return temp_list
```

In [27]:
```
df_movies['crew'].apply(converter_director)
```

Out[27]:
```
0           [James Cameron]
1          [Gore Verbinski]
2              [Sam Mendes]
3       [Christopher Nolan]
4          [Andrew Stanton]
               ...
4804     [Robert Rodriguez]
4805         [Edward Burns]
4806          [Scott Smith]
4807          [Daniel Hsia]
4808     [Brian Herzlinger]
Name: crew, Length: 4809, dtype: object
```

In [28]:
```
df_movies['crew']=df_movies['crew'].apply(converter_director)
```

```
C:\Users\rupeshv\AppData\Local\Temp\ipykernel_11104\2755874778.py:1: SettingWithCopyW
arning:
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/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  df_movies['crew']=df_movies['crew'].apply(converter_director)
```

In [29]:
```
df_movies.head(5)
```

Out[29]:

	genres	keywords	title	overview	movie_id	cast	crew
0	[Action, Adventure, Fantasy, Science Fiction]	[culture clash, future, space war, space colon...	Avatar	In the 22nd century, a paraplegic Marine is di...	19995	[Sam Worthington, Zoe Saldana, Sigourney Weaver]	[James Cameron]
1	[Adventure, Fantasy, Action]	[ocean, drug abuse, exotic island, east india ...	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	285	[Johnny Depp, Orlando Bloom, Keira Knightley]	[Gore Verbinski]
2	[Action, Adventure, Crime]	[spy, based on novel, secret agent, sequel, mi...	Spectre	A cryptic message from Bond's past sends him o...	206647	[Daniel Craig, Christoph Waltz, Léa Seydoux]	[Sam Mendes]
3	[Action, Crime, Drama, Thriller]	[dc comics, crime fighter, terrorist, secret i...	The Dark Knight Rises	Following the death of District Attorney Harve...	49026	[Christian Bale, Michael Caine, Gary Oldman]	[Christopher Nolan]
4	[Action, Adventure, Science Fiction]	[based on novel, mars, medallion, space travel...	John Carter	John Carter is a war-weary, former military ca...	49529	[Taylor Kitsch, Lynn Collins, Samantha Morton]	[Andrew Stanton]

Remove space from words - like culture clash to cultureclash

In [30]:
```
df_movies['genres'] = df_movies['genres'].apply(lambda x:[i.replace(" ","") for i in x
df_movies['keywords']=df_movies['keywords'].apply(lambda x:[i.replace(" ","") for i in
df_movies['cast']=df_movies['cast'].apply(lambda x:[i.replace(" ","") for i in x])
df_movies['crew']=df_movies['crew'].apply(lambda x:[i.replace(" ","") for i in x])
```

```
C:\Users\rupeshv\AppData\Local\Temp\ipykernel_11104\1500789193.py:3: SettingWithCopyW
arning:
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/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  df_movies['genres'] = df_movies['genres'].apply(lambda x:[i.replace(" ","") for i i
n x])
C:\Users\rupeshv\AppData\Local\Temp\ipykernel_11104\1500789193.py:4: SettingWithCopyW
arning:
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/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  df_movies['keywords']=df_movies['keywords'].apply(lambda x:[i.replace(" ","") for i
in x])
C:\Users\rupeshv\AppData\Local\Temp\ipykernel_11104\1500789193.py:5: SettingWithCopyW
arning:
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/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  df_movies['cast']=df_movies['cast'].apply(lambda x:[i.replace(" ","") for i in x])
C:\Users\rupeshv\AppData\Local\Temp\ipykernel_11104\1500789193.py:6: SettingWithCopyW
arning:
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/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  df_movies['crew']=df_movies['crew'].apply(lambda x:[i.replace(" ","") for i in x])
```

In [31]:
```
df_movies.head(5)
```

Out[31]:

	genres	keywords	title	overview	movie_id	cast	crew
0	[Action, Adventure, Fantasy, ScienceFiction]	[cultureclash, future, spacewar, spacecolony, ...	Avatar	In the 22nd century, a paraplegic Marine is di...	19995	[SamWorthington, ZoeSaldana, SigourneyWeaver]	[JamesCameron]
1	[Adventure, Fantasy, Action]	[ocean, drugabuse, exoticisland, eastindiatrad...	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	285	[JohnnyDepp, OrlandoBloom, KeiraKnightley]	[GoreVerbinski]
2	[Action, Adventure, Crime]	[spy, basedonnovel, secretagent, sequel, mi6, ...	Spectre	A cryptic message from Bond's past sends him o...	206647	[DanielCraig, ChristophWaltz, LéaSeydoux]	[SamMendes]
3	[Action, Crime, Drama, Thriller]	[dccomics, crimefighter, terrorist, secretiden...	The Dark Knight Rises	Following the death of District Attorney Harve...	49026	[ChristianBale, MichaelCaine, GaryOldman]	[ChristopherNolan]
4	[Action, Adventure, ScienceFiction]	[basedonnovel, mars, medallion, spacetravel, p...	John Carter	John Carter is a war-weary, former military ca...	49529	[TaylorKitsch, LynnCollins, SamanthaMorton]	[AndrewStanton]

In [32]:
```
df_movies.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4809 entries, 0 to 4808
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   genres    4809 non-null   object
 1   keywords  4809 non-null   object
 2   title     4809 non-null   object
 3   overview  4806 non-null   object
 4   movie_id  4809 non-null   int64 
 5   cast      4809 non-null   object
 6   crew      4809 non-null   object
dtypes: int64(1), object(6)
memory usage: 429.6+ KB
```

Overview column is showing 3 movies where content is null. As we have 4809 movie data so we can remove 3 null overview movies. We missed earlier to drop. Lets drop these 3 entries.

In [33]:
```
#drop null values
df_movies.isnull().sum()
df_movies.dropna(inplace=True)
```

```
C:\Users\rupeshv\AppData\Local\Temp\ipykernel_11104\2594570376.py:3: SettingWithCopyW
arning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  df_movies.dropna(inplace=True)
```

In [34]:
```
df_movies.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4806 entries, 0 to 4808
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   genres    4806 non-null   object
 1   keywords  4806 non-null   object
 2   title     4806 non-null   object
 3   overview  4806 non-null   object
 4   movie_id  4806 non-null   int64 
 5   cast      4806 non-null   object
 6   crew      4806 non-null   object
dtypes: int64(1), object(6)
memory usage: 300.4+ KB
```

Only overview column is in string format so lets convert this into list so that we can merge all 5 text columns

In [35]:
```
#convert overview to list

df_movies['overview']=df_movies['overview'].apply(lambda x : x.split())
```

```
C:\Users\rupeshv\AppData\Local\Temp\ipykernel_11104\4072967743.py:3: SettingWithCopyW
arning:
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/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  df_movies['overview']=df_movies['overview'].apply(lambda x : x.split())
```

Merge all required columns so that we will have a complete text field in one column called TAGS. That will be easy to handle.

In [36]:
```
df_movies['tags'] = df_movies['overview'] + df_movies['genres'] + df_movies['keywords'
```

```
C:\Users\rupeshv\AppData\Local\Temp\ipykernel_11104\3991703209.py:3: SettingWithCopyW
arning:
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/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  df_movies['tags'] = df_movies['overview'] + df_movies['genres'] + df_movies['keywor
ds'] + df_movies['cast'] + df_movies['crew']
```

In [37]:
```
df_movies.head()
```

Out[37]:

	genres	keywords	title	overview	movie_id	cast	crew
0	[Action, Adventure, Fantasy, ScienceFiction]	[cultureclash, future, spacewar, spacecolony, ...	Avatar	[In, the, 22nd, century,, a, paraplegic, Marin...	19995	[SamWorthington, ZoeSaldana, SigourneyWeaver]	[JamesCameron]
1	[Adventure, Fantasy, Action]	[ocean, drugabuse, exoticisland, eastindiatrad...	Pirates of the Caribbean: At World's End	[Captain, Barbossa,, long, believed, to, be, d...	285	[JohnnyDepp, OrlandoBloom, KeiraKnightley]	[GoreVerbinski]
2	[Action, Adventure, Crime]	[spy, basedonnovel, secretagent, sequel, mi6, ...	Spectre	[A, cryptic, message, from, Bond's, past, send...	206647	[DanielCraig, ChristophWaltz, LéaSeydoux]	[SamMendes]
3	[Action, Crime, Drama, Thriller]	[dccomics, crimefighter, terrorist, secretiden...	The Dark Knight Rises	[Following, the, death, of, District, Attorney...	49026	[ChristianBale, MichaelCaine, GaryOldman]	[ChristopherNolan]
4	[Action, Adventure, ScienceFiction]	[basedonnovel, mars, medallion, spacetravel, p...	John Carter	[John, Carter, is, a, war-weary,, former, mili...	49529	[TaylorKitsch, LynnCollins, SamanthaMorton]	[AndrewStanton]

Drop all previous columns which are merged in tags field

In [38]:
```
final_movies = df_movies[['movie_id','title','tags']]
```

In [39]:
```
final_movies.head()
```

Out[39]:

	movie_id	title	tags
0	19995	Avatar	[In, the, 22nd, century,, a, paraplegic, Marin...
1	285	Pirates of the Caribbean: At World's End	[Captain, Barbossa,, long, believed, to, be, d...
2	206647	Spectre	[A, cryptic, message, from, Bond's, past, send...
3	49026	The Dark Knight Rises	[Following, the, death, of, District, Attorney...
4	49529	John Carter	[John, Carter, is, a, war-weary,, former, mili...

Do below operations

- now convert list to string
- lower the tags columns

In [40]:
```
final_movies['tags']=final_movies['tags'].apply(lambda x:" ".join(x))
```

```
C:\Users\rupeshv\AppData\Local\Temp\ipykernel_11104\3481905278.py:2: SettingWithCopyW
arning:
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/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  final_movies['tags']=final_movies['tags'].apply(lambda x:" ".join(x))
```

In [41]:
```
final_movies['tags']=final_movies['tags'].apply(lambda x : x.lower())
```

```
C:\Users\rupeshv\AppData\Local\Temp\ipykernel_11104\1495876804.py:2: SettingWithCopyW
arning:
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/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  final_movies['tags']=final_movies['tags'].apply(lambda x : x.lower())
```

Final dataframe after all preprocessing steps

In [42]:
```
final_movies
```

Out[42]:

	movie_id	title	tags
0	19995	Avatar	in the 22nd century, a paraplegic marine is di...
1	285	Pirates of the Caribbean: At World's End	captain barbossa, long believed to be dead, ha...
2	206647	Spectre	a cryptic message from bond’s past sends him o...
3	49026	The Dark Knight Rises	following the death of district attorney harve...
4	49529	John Carter	john carter is a war-weary, former military ca...
...	...	...	...
4804	9367	El Mariachi	el mariachi just wants to play his guitar and ...
4805	72766	Newlyweds	a newlywed couple's honeymoon is upended by th...
4806	231617	Signed, Sealed, Delivered	"signed, sealed, delivered" introduces a dedic...
4807	126186	Shanghai Calling	when ambitious new york attorney sam is sent t...
4808	25975	My Date with Drew	ever since the second grade when he first saw ...

4806 rows × 3 columns

Text preprocessing

In [44]:
```
#count vectorization

from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(max_features=5000, stop_words='english')
```

Create a vector of 4806 movie with top 5000 words

In [45]:
```
vectors=cv.fit_transform(final_movies['tags']).toarray()
```

In [46]:
```
vectors.shape
```

Out[46]:
```
(4806, 5000)
```

In [47]:
```
cv.get_feature_names()
```

```
C:\Users\rupeshv\Anaconda3\lib\site-packages\sklearn\utils\deprecation.py:87: FutureWarning: Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and will be removed in 1.2. Please use get_feature_names_out instead.
  warnings.warn(msg, category=FutureWarning)
```

Out[47]:
```
['000',
 '007',
 '10',
 '100',
 '11',
 '12',
 '13',
 '14',
 '15',
 '16',
 '17',
 '18',
 '18th',
 '19',
 '1930s',
 '1940s',
 '1950',
 '1950s',
 '1960s',
 '1970s',
 '1980',
 '1980s',
 '1985',
 '1990s',
 '1999',
 '19th',
 '19thcentury',
 '20',
 '200',
 '2009',
 '20th',
 '24',
 '25',
 '30',
 '300',
 '3d',
 '40',
 '50',
 '500',
 '60',
 '60s',
 '70',
 '70s',
 'aaron',
 'aaroneckhart',
 'abandoned',
 'abducted',
 'abigailbreslin',
 'abilities',
 'ability',
 'able',
 'aboard',
 'abuse',
 'abusive',
 'academy',
 'accept',
 'accepted',
 'accepts',
 'access',
 'accident',
```

```
 'accidental',
 'accidentally',
 'accompanied',
 'accomplish',
 'account',
 'accountant',
 'accused',
 'ace',
 'achieve',
 'act',
 'acting',
 'action',
 'actionhero',
 'actions',
 'activist',
 'activities',
 'activity',
 'actor',
 'actors',
 'actress',
 'acts',
 'actual',
 'actually',
 'adam',
 'adams',
 'adamsandler',
 'adamshankman',
 'adaptation',
 'adapted',
 'addict',
 'addicted',
 'addiction',
 'adolescence',
 'adolescent',
 'adopt',
 'adopted',
 'adoption',
 'adopts',
 'adrienbrody',
 'adult',
 'adultanimation',
 'adultery',
 'adulthood',
 'adults',
 'advantage',
 'adventure',
 'adventures',
 'advertising',
 'advice',
 'affair',
 'affairs',
 'affection',
 'affections',
 'afghanistan',
 'africa',
 'african',
 'africanamerican',
 'aftercreditsstinger',
 'afterlife',
 'aftermath',
```

```
 'age',
 'aged',
 'agedifference',
 'agency',
 'agenda',
 'agent',
 'agents',
 'aggressive',
 'aging',
 'ago',
 'agree',
 'agrees',
 'ahead',
 'aid',
 'aided',
 'aids',
 'ailing',
 'air',
 'airplane',
 'airplanecrash',
 'airport',
 'aka',
 'al',
 'alabama',
 'alan',
 'alaska',
 'albert',
 'alcohol',
 'alcoholic',
 'alcoholism',
 'alecbaldwin',
 'alex',
 'alfredhitchcock',
 'ali',
 'alice',
 'alien',
 'alieninvasion',
 'alienlife',
 'aliens',
 'alike',
 'alive',
 'allen',
 'alliance',
 'allied',
 'allies',
 'allow',
 'allowing',
 'allows',
 'ally',
 'alongside',
 'alpacino',
 'alpha',
 'alter',
 'alternate',
 'alternative',
 'alzheimer',
 'amanda',
 'amandapeet',
 'amandaseyfried',
 'amateur',
```

```
 'amazing',
 'ambassador',
 'ambition',
 'ambitious',
 'ambulance',
 'ambush',
 'america',
 'american',
 'americanabroad',
 'americanfootball',
 'americans',
 'amid',
 'amidst',
 'amnesia',
 'amp',
 'amsterdam',
 'amusementpark',
 'amy',
 'amyadams',
 'amysmart',
 'ana',
 'anakin',
 'analyst',
 'anarchiccomedy',
 'ancient',
 'ancientrome',
 'ancientworld',
 'anderson',
 'andiemacdowell',
 'andrew',
 'android',
 'andy',
 'andygarcía',
 'angel',
 'angelabassett',
 'angeles',
 'angelinajolie',
 'angels',
 'anger',
 'anglee',
 'angry',
 'animal',
 'animalattack',
 'animalhorror',
 'animals',
 'animated',
 'animation',
 'anna',
 'annafaris',
 'anne',
 'annehathaway',
 'annemoss',
 'annettebening',
 'annie',
 'anniversary',
 'annual',
 'answer',
 'answers',
 'ant',
 'anthology',
```

```
 'anthony',
 'anthonyanderson',
 'anthonyhopkins',
 'anthropomorphism',
 'anti',
 'antics',
 'antihero',
 'antoinefuqua',
 'antoniobanderas',
 'antonyelchin',
 'apart',
 'apartheid',
 'apartment',
 'ape',
 'apes',
 'apocalypse',
 'apocalyptic',
 'apparent',
 'apparently',
 'appear',
 'appears',
 'apple',
 'appointed',
 'apprentice',
 'approach',
 'approaches',
 'approaching',
 'april',
 'aquarium',
 'arab',
 'arch',
 'archaeologist',
 'architect',
 'arctic',
 'area',
 'aren',
 'arena',
 'argument',
 'arise',
 'aristocrat',
 'armed',
 'arms',
 'army',
 'arnold',
 'arnoldschwarzenegger',
 'arrangedmarriage',
 'arrest',
 'arrested',
 'arrival',
 'arrive',
 'arrives',
 'arriving',
 'arrogant',
 'art',
 'arthur',
 'artificialintelligence',
 'artist',
 'artistic',
 'artists',
 'arts',
```

```
'ashley',
'ashleyjudd',
'ashtonkutcher',
'asia',
'asian',
'aside',
'ask',
'asked',
'asking',
'asks',
'aspirations',
'aspiring',
'assassin',
'assassinate',
'assassination',
'assassins',
'assault',
'assigned',
'assignment',
'assistant',
'assumes',
'asteroid',
'astronaut',
'astronauts',
'asylum',
'athlete',
'atomicbomb',
'attack',
'attacked',
'attacks',
'attempt',
'attempting',
'attempts',
'attending',
'attends',
'attention',
'attitude',
'attorney',
'attracted',
'attraction',
'attractive',
'audience',
'audiences',
'audition',
'august',
'aunt',
'austin',
'australia',
'australian',
'author',
'authorities',
'authority',
'autism',
'auto',
'automobileracing',
'avenge',
'average',
'avoid',
'awaits',
'awakens',
```

```
 'award',
 'away',
 'awry',
 'ax',
 'babe',
 'baby',
 'bachelor',
 'backdrop',
 'backed',
 'background',
 'backgrounds',
 'bad',
 'badly',
 'bag',
 'bahamas',
 'bail',
 'balance',
 'ball',
 'ballet',
 'baltimore',
 'band',
 'bandits',
 'bangkok',
 'banished',
 'bank',
 'banker',
 'bankrobber',
 'bankrobbery',
 'bar',
 'barely',
 'bargained',
 'barn',
 'barney',
 'barry',
 'barrylevinson',
 'bars',
 'base',
 'baseball',
 'based',
 'basedoncomicbook',
 'basedongraphicnovel',
 'basedonnovel',
 'basedonplay',
 'basedonstagemusical',
 'basedontrueevents',
 'basedontruestory',
 'basedontvseries',
 'basedonvideogame',
 'basedonyoungadultnovel',
 'basement',
 'basketball',
 'batman',
 'battle',
 'battlefield',
 'battles',
 'battling',
 'bay',
 'beach',
 'bear',
 'bears',
```

```
 'beast',
 'beasts',
 'beat',
 'beating',
 'beautiful',
 'beautifulwoman',
 'beauty',
 'becky',
 'becominganadult',
 'bed',
 'bedroom',
 'beer',
 'befriends',
 'began',
 'begin',
 'beginning',
 'begins',
 'behavior',
 'beings',
 'beliefs',
 'believe',
 'believed',
 'believes',
 'believing',
 'beloved',
 'ben',
 'benaffleck',
 'beneath',
 'benfoster',
 'beniciodeltoro',
 'benjamin',
 'benjaminbratt',
 'benkingsley',
 'bennett',
 'benstiller',
 'bent',
 'berlin',
 'best',
 'bestfriend',
 'bestfriendsinlove',
 'bet',
 'beth',
 'betrayal',
 'betrayed',
 'bettemidler',
 'better',
 'betty',
 'beverly',
 'bible',
 'big',
 'bigger',
 'biggest',
 'biker',
 'bikini',
 'billhader',
 'billionaire',
 'billmurray',
 'billnighy',
 'billpaxton',
 'billpullman',
```

```
'billy',
'billybobthornton',
'billycrudup',
'billycrystal',
'biography',
'bird',
'birth',
'birthday',
'bisexual',
'bishop',
'bit',
'bite',
'bitter',
'bizarre',
'black',
'blackmagic',
'blackmail',
'blackpeople',
'blacksmith',
'blade',
'blame',
'blind',
'bliss',
'block',
'blonde',
'blood',
'bloodsplatter',
'bloodthirsty',
'bloody',
'blow',
'blue',
'board',
'boarding',
'boardingschool',
'boat',
'bob',
'bobby',
'bobbyfarrelly',
'bobhoskins',
'bodies',
'body',
'bodyguard',
'bold',
'bollywood',
'bomb',
'bombing',
'bond',
'bonds',
'bone',
'book',
'books',
'border',
'bored',
'boredom',
'boring',
'born',
'boss',
'boston',
'botched',
'bound',
```

```
'boundaries',
'bounty',
'bountyhunter',
'bourne',
'box',
'boxer',
'boxing',
'boy',
'boyfriend',
'boys',
'bradleycooper',
'bradpitt',
'brain',
'brand',
'brave',
'bravery',
'brazil',
'brazilian',
'break',
'breakdown',
'breaking',
'breaks',
'brendanfraser',
'brendangleeson',
'brent',
'brettratner',
'brian',
'briandepalma',
'bride',
'bridesmaid',
'bridge',
'brief',
'brielarson',
'brien',
'bright',
'brilliant',
'bring',
'bringing',
'brings',
'brink',
'britain',
'british',
'britishsecretservice',
'brittanymurphy',
'broadway',
'broke',
'broken',
'broker',
'bronx',
'brooklyn',
'brooks',
'broom',
'brothel',
'brother',
'brotherbrotherrelationship',
'brothers',
'brothersisterrelationship',
'brought',
'brown',
'bruce',
```

```
 'brucegreenwood',
 'brucewillis',
 'brutal',
 'brutality',
 'brutally',
 'bryansinger',
 'buck',
 'buddies',
 'buddy',
 'buddycomedy',
 'budget',
 'build',
 'building',
 'built',
 'bully',
 'bullying',
 'bumbling',
 'bunny',
 'burglar',
 'buried',
 'bus',
 'bush',
 'business',
 'businessman',
 'bust',
 'busy',
 'butcher',
 'butler',
 'buy',
 'cabin',
 'caesar',
 'cage',
 'cairo',
 'cal',
 'california',
 'called',
 'calls',
 'calvin',
 'camcorder',
 'came',
 'camera',
 'cameraman',
 'cameras',
 'camerondiaz',
 'camp',
 'campaign',
 'campbell',
 'camping',
 'campus',
 'canada',
 'canadian',
 'cancer',
 'candidate',
 'candy',
 'cannibal',
 'capable',
 'capital',
 'capitalism',
 'capt',
 'captain',
```

```
'captive',
'capture',
'captured',
'captures',
'car',
'caraccident',
'carchase',
'carcrash',
'card',
'care',
'career',
'carefree',
'caretaker',
'caribbean',
'carjourney',
'carl',
'carlagugino',
'carmen',
'carol',
'carolina',
'carrace',
'carrie',
'carry',
'carrying',
'cars',
'cartel',
'carter',
'cartoon',
'caryelwes',
'case',
'caseyaffleck',
'cash',
'casino',
'cast',
'castle',
'cat',
'cataclysm',
'catastrophe',
'catch',
'catches',
'cateblanchett',
'catherinedeneuve',
'catherinekeener',
'catherinezeta',
'catholic',
'catholicism',
'cattle',
'caught',
'cause',
'caused',
'causes',
'causing',
'cavalry',
'cave',
'cavemen',
'celebrate',
'celebrated',
'celebration',
'celebrity',
'cellphone',
```

```
 'cemetery',
 'center',
 'centered',
 'centers',
 'central',
 'centuries',
 'century',
 'ceremony',
 'certain',
 'chain',
 'chainsaw',
 'challenge',
 'challenged',
 'challenges',
 'champion',
 'championship',
 'chance',
 'change',
 'changed',
 'changes',
 'changing',
 'channingtatum',
 'chaos',
 'chaotic',
 'chapter',
 'character',
 'characters',
 'charge',
 'charged',
 'charismatic',
 'charles',
 'charlie',
 'charliesheen',
 'charlizetheron',
 'charlotte',
 'charm',
 'charming',
 'chase',
 'chased',
 'chases',
 'chauffeur',
 'chazzpalminteri',
 'cheating',
 'cheerleader',
 'chef',
 'chemical',
 'cher',
 'chicago',
 'chicken',
 'chief',
 'child',
 'childabuse',
 'childhero',
 'childhood',
 'children',
 'chilling',
 'china',
 'chinese',
 'chip',
 'chiwetelejiofor',
```

```
'chloëgracemoretz',
'chloësevigny',
'chocolate',
'choice',
'choices',
'choose',
'chosen',
'chowyun',
'chris',
'chriscolumbus',
'chriscooper',
'chrisevans',
'chrishemsworth',
'chrisklein',
'chrispine',
'chrisrock',
'christ',
'christian',
'christianbale',
'christianity',
'christianslater',
'christinaapplegate',
'christinaricci',
'christine',
'christmas',
'christmasparty',
'christopher',
'christopherlloyd',
'christophernolan',
'christopherplummer',
'christopherwalken',
'christophwaltz',
'chrisweitz',
'chronicle',
'chronicles',
'chuck',
'church',
'cia',
'cigarettesmoking',
'cillianmurphy',
'cindy',
'cinema',
'circle',
'circuit',
'circumstances',
'circus',
'cities',
'citizens',
'city',
'civil',
'civilian',
'civilization',
'civilwar',
'claim',
'claims',
'claire',
'clairedanes',
'clan',
'clark',
'clash',
```

```
'class',
'classes',
'classic',
'classmate',
'classmates',
'classroom',
'claudevandamme',
'clay',
'clean',
'clear',
'clerk',
'client',
'clients',
'climate',
'climbing',
'clinteastwood',
'clique',
'cliveowen',
'clock',
'clone',
'cloning',
'close',
'closed',
'closer',
'club',
'clubs',
'clues',
'clutches',
'coach',
'coast',
'cocaine',
'code',
'cody',
'coffin',
'cohen',
'col',
'cold',
'coldwar',
'cole',
'colin',
'colinfarrell',
'colinfirth',
'collapse',
'colleague',
'colleagues',
'collect',
'collection',
'collector',
'college',
'collide',
'collision',
'colonel',
'colony',
'color',
'colorado',
'colorful',
'coma',
'combat',
'combined',
'come',
```

```
 'comeback',
 'comedian',
 'comedic',
 'comedy',
 'comes',
 'comet',
 'comfort',
 'comic',
 'comics',
 'coming',
 'comingofage',
 'comingout',
 'command',
 'commander',
 'commercial',
 'commit',
 'commitment',
 'committed',
 'common',
 'communication',
 'community',
 'companion',
 'company',
 'compete',
 'competing',
 'competition',
 'complete',
 'completely',
 'complex',
 'complicated',
 'complications',
 'composer',
 'computer',
 'computervirus',
 'conan',
 'concert',
 'conclusion',
 'condition',
 'confederate',
 'confession',
 'confidence',
 'confident',
 'conflict',
 'confront',
 'confronted',
 'confronts',
 'confused',
 'congress',
 'conman',
 'connected',
 'connection',
 'connell',
 'connor',
 'conquer',
 'conscience',
 'consequences',
 'conservative',
 'considered',
 'conspiracy',
 'conspire',
```

```
 'constant',
 'constantly',
 'construction',
 'consumed',
 'contact',
 'contain',
 'contemporary',
 'contend',
 'contest',
 'continue',
 'continues',
 'continuing',
 'contract',
 'control',
 'controlled',
 'controlling',
 'controversial',
 'convention',
 'converge',
 'convict',
 'convicted',
 'convince',
 'convinced',
 'convinces',
 'cook',
 'cooking',
 'cool',
 'cooper',
 'cop',
 'cope',
 'cops',
 'core',
 'corner',
 'corners',
 'corporate',
 'corporation',
 'corpse',
 'corrupt',
 'corruption',
 'cost',
 ...]
```

We need to do stemming here in final_movies and then apply count vector again. This step was done because we have words like accident, accidental and accidentally - to make it one we use stemming

For stem we will include nltk lib

In [48]:
```
from nltk.stem.porter import PorterStemmer
```

In [49]:
```
ps=PorterStemmer()
```

In [50]:
```
def stem_text(text):
    word_list = []
    for i in text.split():
        word_list.append(ps.stem(i))
```

```
    return " ".join(word_list)
```

In [51]:
```
final_movies['tags']=final_movies['tags'].apply(stem_text)
```

```
C:\Users\rupeshv\AppData\Local\Temp\ipykernel_11104\1979869361.py:1: SettingWithCopyW
arning:
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/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  final_movies['tags']=final_movies['tags'].apply(stem_text)
```

In [52]:
```
final_movies.head()
```

Out[52]:

	movie_id	title	tags
0	19995	Avatar	in the 22nd century, a parapleg marin is dispa...
1	285	Pirates of the Caribbean: At World's End	captain barbossa, long believ to be dead, ha c...
2	206647	Spectre	a cryptic messag from bond' past send him on a...
3	49026	The Dark Knight Rises	follow the death of district attorney harvey d...
4	49529	John Carter	john carter is a war-weary, former militari ca...

In [53]:
```
#do count vector again
```

In [54]:
```
vectors_new = cv.fit_transform(final_movies['tags']).toarray()
```

In [55]:
```
vectors_new.shape
```

Out[55]:
```
(4806, 5000)
```

In [56]:
```
cv.get_feature_names()
```

```
C:\Users\rupeshv\Anaconda3\lib\site-packages\sklearn\utils\deprecation.py:87: FutureW
arning: Function get_feature_names is deprecated; get_feature_names is deprecated in
1.0 and will be removed in 1.2. Please use get_feature_names_out instead.
  warnings.warn(msg, category=FutureWarning)
```

Out[56]:
```
['000',
 '007',
 '10',
 '100',
 '11',
 '12',
 '13',
 '14',
 '15',
 '16',
 '17',
 '17th',
 '18',
 '18th',
 '18thcenturi',
 '19',
 '1910',
 '1920',
 '1930',
 '1940',
 '1944',
 '1950',
 '1950s',
 '1960',
 '1960s',
 '1970',
 '1970s',
 '1971',
 '1974',
 '1976',
 '1980',
 '1985',
 '1990',
 '1999',
 '19th',
 '19thcenturi',
 '20',
 '200',
 '2003',
 '2009',
 '20th',
 '21st',
 '23',
 '24',
 '25',
 '30',
 '300',
 '3d',
 '40',
 '50',
 '500',
 '60',
 '70',
 '80',
 'aaron',
 'aaroneckhart',
 'abandon',
 'abduct',
 'abigailbreslin',
 'abil',
```

```
'abl',
'aboard',
'abov',
'abus',
'academ',
'academi',
'accept',
'access',
'accid',
'accident',
'acclaim',
'accompani',
'accomplish',
'account',
'accus',
'ace',
'achiev',
'acquaint',
'act',
'action',
'actionhero',
'activ',
'activist',
'activities',
'actor',
'actress',
'actual',
'ad',
'adam',
'adamsandl',
'adamshankman',
'adapt',
'add',
'addict',
'adjust',
'admir',
'admit',
'adolesc',
'adopt',
'ador',
'adrienbrodi',
'adult',
'adultanim',
'adulteri',
'adulthood',
'advanc',
'adventur',
'adventure',
'adventures',
'advertis',
'advic',
'advis',
'affair',
'affect',
'afghanistan',
'africa',
'african',
'africanamerican',
'aftercreditssting',
'afterlif',
```

```
 'aftermath',
 'ag',
 'age',
 'agediffer',
 'agenc',
 'agency',
 'agenda',
 'agent',
 'agents',
 'aggress',
 'ago',
 'agre',
 'ahead',
 'aid',
 'aidanquinn',
 'ail',
 'aim',
 'air',
 'airplan',
 'airplanecrash',
 'airport',
 'aka',
 'al',
 'alabama',
 'alan',
 'alaska',
 'albert',
 'alcatraz',
 'alcohol',
 'alecbaldwin',
 'alex',
 'alexkendrick',
 'alfredhitchcock',
 'alfredmolina',
 'ali',
 'alic',
 'alice',
 'alien',
 'alieninvas',
 'alienlife',
 'alienplanet',
 'aliens',
 'alik',
 'aliv',
 'alive',
 'allen',
 'alli',
 'allianc',
 'allow',
 'alon',
 'alongsid',
 'alpacino',
 'alpha',
 'alreadi',
 'alter',
 'altern',
 'alway',
 'alyssa',
 'alzheimer',
 'amanda',
```

```
 'amandapeet',
 'amandaseyfri',
 'amateur',
 'amaz',
 'amazon',
 'ambassador',
 'ambit',
 'ambiti',
 'ambul',
 'ambush',
 'america',
 'american',
 'americanabroad',
 'americancivilwar',
 'americanfootbal',
 'americanfootballplay',
 'amid',
 'amidst',
 'amnesia',
 'amp',
 'amsterdam',
 'amus',
 'amusementpark',
 'amy',
 'amyadam',
 'amysmart',
 'ana',
 'anakin',
 'analyst',
 'anarchiccomedi',
 'ancient',
 'ancientrom',
 'ancientworld',
 'anderson',
 'andi',
 'andiemacdowel',
 'andrew',
 'android',
 'andy',
 'andygarcía',
 'angel',
 'angela',
 'angelabassett',
 'angeles',
 'angelinajoli',
 'anger',
 'angle',
 'angri',
 'ani',
 'anim',
 'animalattack',
 'animalhorror',
 'animals',
 'anjelicahuston',
 'ann',
 'anna',
 'annafari',
 'annakendrick',
 'anne',
 'annehathaway',
```

```
'annemoss',
'annetteben',
'anni',
'annie',
'anniversari',
'announc',
'annual',
'anonym',
'anoth',
'answer',
'ant',
'antholog',
'anthoni',
'anthonyanderson',
'anthonyhopkin',
'anthropomorph',
'anti',
'antic',
'antihero',
'antiqu',
'antoinefuqua',
'antoniobandera',
'antonyelchin',
'anyon',
'anyth',
'apart',
'apartheid',
'apartment',
'ape',
'apocalyps',
'apocalypse',
'apocalypt',
'appar',
'appear',
'appl',
'apple',
'appoint',
'appreci',
'apprentic',
'approach',
'april',
'aquarium',
'arab',
'arch',
'archaeologist',
'archeolog',
'archer',
'architect',
'arctic',
'area',
'aren',
'arena',
'argument',
'aris',
'aristocrat',
'arm',
'armi',
'armor',
'armsdeal',
'army',
```

```
'arnold',
'arnoldschwarzenegg',
'arrang',
'arrangedmarriag',
'arrest',
'arriv',
'arrog',
'art',
'arthur',
'artifact',
'artifici',
'artificialintellig',
'artist',
'ash',
'ashley',
'ashleyjudd',
'ashtonkutch',
'asia',
'asian',
'asid',
'ask',
'aspect',
'aspir',
'assassin',
'assault',
'assembl',
'assign',
'assist',
'assistant',
'associ',
'assum',
'asteroid',
'astronaut',
'asylum',
'atheist',
'athlet',
'atom',
'atomicbomb',
'attack',
'attacks',
'attempt',
'attend',
'attent',
'attic',
'attitud',
'attorney',
'attract',
'auction',
'audienc',
'audit',
'august',
'aunt',
'austin',
'australia',
'australian',
'author',
'autism',
'auto',
'automobilerac',
'aveng',
```

```
'averag',
'avoid',
'await',
'awak',
'awaken',
'awar',
'award',
'away',
'awkward',
'awri',
'awry',
'ax',
'babe',
'babi',
'baby',
'bachelor',
'backdrop',
'background',
'backpack',
'bad',
'bag',
'bahama',
'bail',
'balanc',
'ball',
'ballet',
'balloon',
'baltimor',
'ban',
'band',
'bandit',
'bangkok',
'banish',
'bank',
'banker',
'bankrobb',
'bankrobberi',
'bar',
'barbrastreisand',
'bare',
'bargain',
'barn',
'barney',
'baron',
'barri',
'barrylevinson',
'barrysonnenfeld',
'bas',
'base',
'basebal',
'basedoncomicbook',
'basedongraphicnovel',
'basedonnovel',
'basedonplay',
'basedonstagemus',
'basedontrueev',
'basedontruestori',
'basedontvseri',
'basedonvideogam',
'basedonyoungadultnovel',
```

```
 'basement',
 'basketbal',
 'basketball',
 'bat',
 'batman',
 'battl',
 'battle',
 'battlefield',
 'bay',
 'beach',
 'beam',
 'bear',
 'beard',
 'beast',
 'beat',
 'beauti',
 'beautiful',
 'beautifulwoman',
 'beauty',
 'becam',
 'becaus',
 'becki',
 'becom',
 'becominganadult',
 'bed',
 'bedroom',
 'bee',
 'beer',
 'befor',
 'befriend',
 'began',
 'begin',
 'begins',
 'behavior',
 'belief',
 'believ',
 'bell',
 'bella',
 'belong',
 'belov',
 'ben',
 'benaffleck',
 'bend',
 'beneath',
 'benefit',
 'benfost',
 'beniciodeltoro',
 'benjamin',
 'benjaminbratt',
 'benkingsley',
 'bennett',
 'benstil',
 'bent',
 'berlin',
 'best',
 'bestfriend',
 'bestfriendsinlov',
 'bet',
 'beth',
 'betray',
```

```
'bettemidl',
'better',
'betti',
'beverli',
'bibl',
'bid',
'big',
'bigger',
'biggest',
'bike',
'biker',
'bikini',
'billhad',
'billi',
'billionair',
'billmurray',
'billnighi',
'billpaxton',
'billpullman',
'billybobthornton',
'billycrudup',
'billycryst',
'biographi',
'biolog',
'bird',
'birth',
'birthday',
'bisexu',
'bishop',
'bit',
'bite',
'bitter',
'bizarr',
'black',
'blackmag',
'blackmail',
'blackpeopl',
'blacksmith',
'blade',
'blame',
'blend',
'blind',
'bliss',
'blizzard',
'block',
'blond',
'blood',
'bloodi',
'bloodsplatt',
'bloodthirsti',
'blow',
'blue',
'board',
'boardingschool',
'boat',
'bob',
'bobbi',
'bobbyfarrelli',
'bobhoskin',
'bodi',
```

```
 'body',
 'bodyguard',
 'bold',
 'bollywood',
 'bomb',
 'bond',
 'bone',
 'book',
 'border',
 'bore',
 'boredom',
 'born',
 'boss',
 'boston',
 'botch',
 'bound',
 'boundari',
 'bounti',
 'bountyhunt',
 'bout',
 'box',
 'boxer',
 'boy',
 'boyfriend',
 'boys',
 'bradleycoop',
 'bradpitt',
 'brain',
 'brainwash',
 'brand',
 'brandon',
 'brave',
 'braveri',
 'brazil',
 'brazilian',
 'break',
 'breakdown',
 'breast',
 'breath',
 'breed',
 'brendanfras',
 'brendangleeson',
 'brent',
 'brettratn',
 'brian',
 'briandepalma',
 'bride',
 'bridesmaid',
 'bridg',
 'brief',
 'brielarson',
 'brien',
 'bright',
 'brilliant',
 'bring',
 'brink',
 'britain',
 'british',
 'britishsecretservic',
 'brittanymurphi',
```

```
 'broadcast',
 'broadway',
 'broke',
 'broken',
 'broker',
 'bronx',
 'brook',
 'brooklyn',
 'broom',
 'brothel',
 'brother',
 'brotherbrotherrelationship',
 'brothers',
 'brothersisterrelationship',
 'brought',
 'brown',
 'bruce',
 'brucegreenwood',
 'brucewilli',
 'brutal',
 'bryansing',
 'bu',
 'buck',
 'bud',
 'buddi',
 'buddy',
 'buddycomedi',
 'buddycop',
 'budget',
 'build',
 'building',
 'built',
 'bullet',
 'bulli',
 'bumbl',
 'bunch',
 'bunker',
 'bunni',
 'burglar',
 'buri',
 'burn',
 'bush',
 'busi',
 'business',
 'businessman',
 'bust',
 'butcher',
 'butler',
 'butt',
 'button',
 'buy',
 'buzz',
 'cabin',
 'caesar',
 'cage',
 'cairo',
 'cal',
 'california',
 'calvin',
 'camcord',
```

```
 'came',
 'camera',
 'cameraman',
 'camerondiaz',
 'camp',
 'campaign',
 'campbell',
 'campu',
 'canada',
 'canadian',
 'cancer',
 'candi',
 'candid',
 'canin',
 'cannib',
 'canuxploit',
 'capabl',
 'caper',
 'capit',
 'capt',
 'captain',
 'captiv',
 'captur',
 'capture',
 'car',
 'caraccid',
 'carchas',
 'carcrash',
 'card',
 'care',
 'career',
 'carefre',
 'caretak',
 'careymulligan',
 'caribbean',
 'carjourney',
 'carl',
 'carlagugino',
 'carmen',
 'carol',
 'carolina',
 'carrac',
 'carri',
 'carrie',
 'cartel',
 'carter',
 'cartoon',
 'caryelw',
 'case',
 'caseyaffleck',
 'cash',
 'casino',
 'cast',
 'castl',
 'cat',
 'cataclysm',
 'catastroph',
 'catch',
 'cateblanchett',
 'catherinedeneuv',
```

```
 'catherinekeen',
 'catherinezeta',
 'cathol',
 'catholic',
 'cattl',
 'caught',
 'caus',
 'cavalri',
 'cave',
 'cavemen',
 'celebr',
 'celebration',
 'cell',
 'cellphon',
 'cemeteri',
 'center',
 'centr',
 'central',
 'centuri',
 'centuries',
 'century',
 'ceo',
 'certain',
 'chad',
 'chain',
 'chainsaw',
 'challeng',
 'chamber',
 'champion',
 'championship',
 'chanc',
 'chance',
 'chang',
 'change',
 'changed',
 'changes',
 'channingtatum',
 'chao',
 'chaos',
 'chaotic',
 'chapter',
 'charact',
 'character',
 'characters',
 'charg',
 'charismat',
 'charl',
 'charli',
 'charlie',
 'charliesheen',
 'charlizetheron',
 'charm',
 'chart',
 'chase',
 'chauffeur',
 'chazzpalminteri',
 'cheat',
 'check',
 'cheerlead',
 'chef',
```

```
'chemic',
'cher',
'chevychas',
'chicago',
'chicken',
'chief',
'child',
'childabus',
'childhero',
'childhood',
'childprodigi',
'children',
'chill',
'chimp',
'china',
'chines',
'chip',
'chipmunk',
'chiwetelejiofor',
'chloe',
'chloëgracemoretz',
'chloësevigni',
'chocol',
'choic',
'choice',
'choos',
'chosen',
'chowyun',
'chri',
'chriscolumbu',
'chriscoop',
'chrisevan',
'chrishemsworth',
'chrisklein',
'chrispin',
'chrisrock',
'christ',
'christian',
'christianbal',
'christianslat',
'christin',
'christinaappleg',
'christinaricci',
'christma',
'christmas',
'christmasparti',
'christmastre',
'christoph',
'christopherlambert',
'christopherlloyd',
'christophernolan',
'christopherplumm',
'christopherwalken',
'christophwaltz',
'chrisweitz',
'chronicl',
'chuck',
'church',
'cia',
'ciaránhind',
```

```
'cigarettesmok',
'cillianmurphi',
'cinema',
'circl',
'circu',
'circuit',
'circumst',
'citi',
'citizen',
'city',
'civil',
'civilian',
'civilwar',
'claim',
'clair',
'clairedan',
'claireforlani',
'clan',
'clark',
'clash',
'class',
'classdiffer',
'classic',
'classmat',
'classroom',
'claudevandamm',
'clay',
'clean',
'clear',
'clerk',
'clever',
'client',
'clients',
'cliff',
'climat',
'climb',
'clinteastwood',
'cliveowen',
'clock',
'clone',
'close',
'closer',
'cloud',
'clown',
'club',
'clue',
'clueless',
'clutch',
'coach',
'coast',
'cocain',
'code',
'coffin',
'cohen',
'col',
'cold',
'coldwar',
'cole',
'colin',
'colinfarrel',
```

```
 'colinfirth',
 'collaps',
 'colleagu',
 'collect',
 'collector',
 'colleg',
 'college',
 'collid',
 'collis',
 'colombia',
 'colonel',
 'coloni',
 'color',
 'colorado',
 'coma',
 'combat',
 'combin',
 'come',
 'comeback',
 'comed',
 'comedi',
 'comedian',
 'comedy',
 'comet',
 'comfort',
 'comic',
 'coming',
 'comingofag',
 'comingout',
 'command',
 'commando',
 'commerci',
 'commiss',
 'commit',
 'common',
 'commun',
 'communist',
 'community',
 'compani',
 'companion',
 'company',
 'compet',
 'competit',
 'competition',
 'complet',
 'complex',
 'complic',
 'compos',
 'compuls',
 'comput',
 'computerviru',
 'conan',
 'concern',
 'concert',
 'concoct',
 'condit',
 'condition',
 'conduct',
 'confeder',
 'confess',
```

```
 'confid',
 'confin',
 'conflict',
 'confront',
 'confus',
 'congress',
 'conman',
 'connect',
 'connecticut',
 'connel',
 'connor',
 'conquer',
 'consequ',
 'consequences',
 'conserv',
 'consid',
 'conspir',
 'conspiraci',
 'conspiracy',
 'constant',
 'constantli',
 'construct',
 'consum',
 'contact',
 'contain',
 'contemporari',
 'contend',
 'content',
 'contest',
 'continu',
 'contract',
 'contractor',
 'control',
 'controversi',
 'convent',
 'converg',
 'convers',
 'convict',
 'convinc',
 'cook',
 ...]
```

Now we need to find the simlarity b/w different movies - we will find the distance b/w movies with respect to angle

- Here euclidean distance is not good measure to find distance b/w differnt movie vector
- We will use cosaine distance in term of similarity

In [57]:
```
from sklearn.metrics.pairwise import cosine_similarity
cosine_similarity(vectors_new)
```

Out[57]:
```
array([[1.        , 0.08346223, 0.0860309 , ..., 0.04499213, 0.        ,
        0.        ],
       [0.08346223, 1.        , 0.06063391, ..., 0.02378257, 0.        ,
        0.02615329],
       [0.0860309 , 0.06063391, 1.        , ..., 0.02451452, 0.        ,
        0.        ],
       ...,
       [0.04499213, 0.02378257, 0.02451452, ..., 1.        , 0.03962144,
        0.04229549],
       [0.        , 0.        , 0.        , ..., 0.03962144, 1.        ,
        0.08714204],
       [0.        , 0.02615329, 0.        , ..., 0.04229549, 0.08714204,
        1.        ]])
```

In [58]:
```
similarity = cosine_similarity(vectors_new)
```

Here we have distance of our movie with all movie and which has less disstance that movie will be considered

In [70]:
```
similarity[0]
```

Out[70]:
```
array([1.        , 0.08346223, 0.0860309 , ..., 0.04499213, 0.        ,
       0.        ])
```

In [60]:
```
sorted(list(enumerate(similarity[0])),reverse=True,key=lambda x :x[1])[1:6]
```

Out[60]:
```
[(1216, 0.28676966733820225),
 (2409, 0.26901379342448517),
 (3730, 0.2605130246476754),
 (507, 0.255608593705383),
 (539, 0.25038669783359574)]
```

In [61]:
```
final_movies.head(2)
```

Out[61]:

	movie_id	title	tags
0	19995	Avatar	in the 22nd century, a parapleg marin is dispa...
1	285	Pirates of the Caribbean: At World's End	captain barbossa, long believ to be dead, ha c...

Top 5 Movie selection logic

In [62]:
```
def recommend(movie):
    movie_index = final_movies[final_movies['title'] == movie].index[0]
    distances=similarity[movie_index]
    movie_list=sorted(list(enumerate(distances)),reverse=True,key=lambda x :x[1])[1:6]

    for i in movie_list:
        print(final_movies.iloc[i[0]]['title'])
```

In [63]:
```
recommend('Batman Begins')
```

```
The Dark Knight
Batman
Batman
The Dark Knight Rises
10th & Wolf
```

In [64]:
```
final_movies.iloc[2409]['title']
```

Out[64]:
```
'Aliens'
```

In [65]:
```
final_movies.head()
```

Out[65]:

	movie_id	title	tags
0	19995	Avatar	in the 22nd century, a parapleg marin is dispa...
1	285	Pirates of the Caribbean: At World's End	captain barbossa, long believ to be dead, ha c...
2	206647	Spectre	a cryptic messag from bond' past send him on a...
3	49026	The Dark Knight Rises	follow the death of district attorney harvey d...
4	49529	John Carter	john carter is a war-weary, former militari ca...

Saving model using Joblib

In [66]:
```
import joblib
```

In [68]:
```
joblib.dump(final_movies,'df_final_movies.pkl')
```

In [71]:
```
joblib.dump(similarity,'similarity.pkl')
```

Out[71]:
```
['similarity.pkl']
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: