In [29]:
```
import pandas as pd
import numpy as np
```

In [30]:
```
credit=pd.read_csv("tmdb_5000_credits.csv")
movies=pd.read_csv("tmdb_5000_movies.csv")
```

In [31]:
```
movies.head(1)
```

Out[31]:

	budget	genres	homepage	id	keywords	original_language	original_title	overview	popularity	production_comp
0	237000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.avatarmovie.com/	19995	[{"id": 1463, "name": "culture clash"}, {"id":...	en	Avatar	In the 22nd century, a paraplegic Marine is di...	150.437577	[{"name": "Ing Film Partners

In [32]:
```
credit.head(1)
```

Out[32]:

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

In [36]:
```
movies=movies.merge(credit,on="title")
```

In [38]:
```
movies.head(1)
```

Out[38]:

	budget	genres	homepage	id	keywords	original_language	original_title	overview	popularity	production_comp
0	237000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.avatarmovie.com/	19995	[{"id": 1463, "name": "culture clash"}, {"id":...	en	Avatar	In the 22nd century, a paraplegic Marine is di...	150.437577	[{"name": "Ing Film Partners

1 rows × 26 columns

In [40]:
```
movies.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4821 entries, 0 to 4820
Data columns (total 26 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   budget                4821 non-null   int64  
 1   genres                4821 non-null   object 
 2   homepage              1715 non-null   object 
 3   id                    4821 non-null   int64  
 4   keywords              4821 non-null   object 
 5   original_language     4821 non-null   object 
 6   original_title        4821 non-null   object 
 7   overview              4818 non-null   object 
 8   popularity            4821 non-null   float64
 9   production_companies  4821 non-null   object 
 10  production_countries  4821 non-null   object 
 11  release_date          4820 non-null   object 
 12  revenue               4821 non-null   int64  
 13  runtime               4819 non-null   float64
 14  spoken_languages      4821 non-null   object 
 15  status                4821 non-null   object 
 16  tagline               3977 non-null   object 
 17  title                 4821 non-null   object 
 18  vote_average          4821 non-null   float64
 19  vote_count            4821 non-null   int64  
 20  movie_id_x            4821 non-null   int64  
 21  cast_x                4821 non-null   object 
 22  crew_x                4821 non-null   object 
 23  movie_id_y            4821 non-null   int64  
 24  cast_y                4821 non-null   object 
 25  crew_y                4821 non-null   object 
dtypes: float64(3), int64(6), object(17)
memory usage: 1016.9+ KB
```

In [42]:
```
# we used only this columns
# genres
# id
# keywords
# title
# overview
# cast_x
# crew_x
```

```
movies=movies[["id","genres","keywords","title","overview","cast_x","crew_x"]]
```

In [43]:
```
movies.head()
```

Out[43]:

	id	genres	keywords	title	overview	cast_x	crew_x
0	19995	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	[{"id": 1463, "name": "culture clash"}, {"id":...	Avatar	In the 22nd century, a paraplegic Marine is di...	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
1	285	[{"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...	[{"cast_id": 4, "character": "Captain Jack Spa...	[{"credit_id": "52fe4232c3a36847f800b579", "de...
2	206647	[{"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...	[{"cast_id": 1, "character": "James Bond", "cr...	[{"credit_id": "54805967c3a36829b5002c41", "de...
3	49026	[{"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...	[{"cast_id": 2, "character": "Bruce Wayne / Ba...	[{"credit_id": "52fe4781c3a36847f81398c3", "de...
4	49529	[{"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...	[{"cast_id": 5, "character": "John Carter", "c...	[{"credit_id": "52fe479ac3a36847f813eaa3", "de...

In [49]:
```
movies.isnull().sum()
```

Out[49]:
```
id          0
genres      0
keywords    0
title       0
overview    3
cast_x      0
crew_x      0
dtype: int64
```

In [50]:
```
movies.dropna(inplace=True)
```

```
C:\Users\kiran\AppData\Local\Temp\ipykernel_8676\3786870272.py:1: SettingWithCopyWarning:
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/user_guide/indexing.html#returning-a-view-versus-a-copy
  movies.dropna(inplace=True)
```

In [51]:
```
movies.isnull().sum()
```

Out[51]:
```
id          0
genres      0
keywords    0
title       0
overview    0
cast_x      0
crew_x      0
dtype: int64
```

In [52]:
```
movies.duplicated().sum()
```

Out[52]:
```
12
```

In [56]:
```
movies.drop_duplicates(inplace=True)
```

```
C:\Users\kiran\AppData\Local\Temp\ipykernel_8676\1441796974.py:1: SettingWithCopyWarning:
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/user_guide/indexing.html#returning-a-view-versus-a-copy
  movies.drop_duplicates(inplace=True)
```

In [57]:
```
movies.duplicated().sum()
```

Out[57]:
```
0
```

In [58]:
```
movies.iloc[0].genres
```

Out[58]:
```
'[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 878, "name": "Science Fiction"}]'
```

In [60]:
```
import ast
```

In [59]:
```
def convert(obj):
    L= []
    for i in ast.literal_eval(obj):
        L.append(i["name"])
    return L
```

In [62]:
```
movies["genres"]=movies["genres"].apply(convert)
```

```
C:\Users\kiran\AppData\Local\Temp\ipykernel_8676\1455536712.py:1: SettingWithCopyWarning:
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/user_guide/indexing.html#returning-a-view-versus-a-copy
  movies["genres"]=movies["genres"].apply(convert)
```

In [65]:
```
movies["keywords"]=movies["keywords"].apply(convert)
```

```
C:\Users\kiran\AppData\Local\Temp\ipykernel_8676\3818525984.py:1: SettingWithCopyWarning:
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/user_guide/indexing.html#returning-a-view-versus-a-copy
  movies["keywords"]=movies["keywords"].apply(convert)
```

In [66]:
```
movies.head()
```

Out[66]:

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

In [78]:
```
def convert3(obj):
    L=[]
    counter=0
    for i in ast.literal_eval(obj):
        if counter !=3:
            L.append(i["name"])
            counter+=1
        else:
            break
    return L
```

In [79]:
```
movies["cast_x"]=movies["cast_x"].apply(convert3)
```

```
C:\Users\kiran\AppData\Local\Temp\ipykernel_8676\1570072864.py:1: SettingWithCopyWarning:
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/user_guide/indexing.html#returning-a-view-versus-a-copy
  movies["cast_x"]=movies["cast_x"].apply(convert3)
```

In [80]:
```
movies.head(2)
```

Out[80]:

	id	genres	keywords	title	overview	cast_x	crew_x
0	19995	[Action, Adventure, Fantasy, Science Fiction]	[culture clash, future, space war, space colon...	Avatar	In the 22nd century, a paraplegic Marine is di...	[Sam Worthington, Zoe Saldana, Sigourney Weaver]	[{"credit_id": "52fe48009251416c750aca23", "de...
1	285	[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...	[Johnny Depp, Orlando Bloom, Keira Knightley]	[{"credit_id": "52fe4232c3a36847f800b579", "de...

In [81]:
```
def convert4(obj):
    L= []
    for i in ast.literal_eval(obj):
        if i["job"]=="Director":
            L.append(i["name"])
            break
    return L
```

In [85]:
```
movies["crew_x"]=movies["crew_x"].apply(convert4)
```

```
C:\Users\kiran\AppData\Local\Temp\ipykernel_8676\2344520527.py:1: SettingWithCopyWarning:
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/user_guide/indexing.html#returning-a-view-versus-a-copy
  movies["crew_x"]=movies["crew_x"].apply(convert4)
```

In [86]:
```
movies.head()
```

Out[86]:

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

In [87]:
```
movies["overview"]=movies["overview"].apply(lambda x:x.split())
```

```
C:\Users\kiran\AppData\Local\Temp\ipykernel_8676\229043305.py:1: SettingWithCopyWarning:
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/user_guide/indexing.html#returning-a-view-versus-a-copy
  movies["overview"]=movies["overview"].apply(lambda x:x.split())
```

In [88]:
```
movies.head(2)
```

Out[88]:

	id	genres	keywords	title	overview	cast_x	crew_x
0	19995	[Action, Adventure, Fantasy, Science Fiction]	[culture clash, future, space war, space colon...	Avatar	[In, the, 22nd, century,, a, paraplegic, Marin...	[Sam Worthington, Zoe Saldana, Sigourney Weaver]	[James Cameron]
1	285	[Adventure, Fantasy, Action]	[ocean, drug abuse, exotic island, east india ...	Pirates of the Caribbean: At World's End	[Captain, Barbossa,, long, believed, to, be, d...	[Johnny Depp, Orlando Bloom, Keira Knightley]	[Gore Verbinski]

In [91]:
```
movies["genres"]=movies["genres"].apply(lambda x:[i.replace(" ","")for i in x])
movies["keywords"]=movies["keywords"].apply(lambda x:[i.replace(" ","")for i in x])
movies["cast_x"]=movies["cast_x"].apply(lambda x:[i.replace(" ","")for i in x])
movies["crew_x"]=movies["crew_x"].apply(lambda x:[i.replace(" ","")for i in x])
```

```
C:\Users\kiran\AppData\Local\Temp\ipykernel_8676\3309237797.py:1: SettingWithCopyWarning:
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/user_guide/indexing.html#returning-a-view-versus-a-copy
  movies["genres"]=movies["genres"].apply(lambda x:[i.replace(" ","")for i in x])
C:\Users\kiran\AppData\Local\Temp\ipykernel_8676\3309237797.py:2: SettingWithCopyWarning:
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/user_guide/indexing.html#returning-a-view-versus-a-copy
  movies["keywords"]=movies["keywords"].apply(lambda x:[i.replace(" ","")for i in x])
C:\Users\kiran\AppData\Local\Temp\ipykernel_8676\3309237797.py:3: SettingWithCopyWarning:
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/user_guide/indexing.html#returning-a-view-versus-a-copy
  movies["cast_x"]=movies["cast_x"].apply(lambda x:[i.replace(" ","")for i in x])
C:\Users\kiran\AppData\Local\Temp\ipykernel_8676\3309237797.py:4: SettingWithCopyWarning:
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/user_guide/indexing.html#returning-a-view-versus-a-copy
  movies["crew_x"]=movies["crew_x"].apply(lambda x:[i.replace(" ","")for i in x])
```

In [92]:
```
movies.head(2)
```

Out[92]:

	id	genres	keywords	title	overview	cast_x	crew_x
0	19995	[Action, Adventure, Fantasy, ScienceFiction]	[cultureclash, future, spacewar, spacecolony, ...	Avatar	[In, the, 22nd, century,, a, paraplegic, Marin...	[SamWorthington, ZoeSaldana, SigourneyWeaver]	[JamesCameron]
1	285	[Adventure, Fantasy, Action]	[ocean, drugabuse, exoticisland, eastindiatrad...	Pirates of the Caribbean: At World's End	[Captain, Barbossa,, long, believed, to, be, d...	[JohnnyDepp, OrlandoBloom, KeiraKnightley]	[GoreVerbinski]

In [93]:
```
movies.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4806 entries, 0 to 4820
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   id        4806 non-null   int64 
 1   genres    4806 non-null   object
 2   keywords  4806 non-null   object
 3   title     4806 non-null   object
 4   overview  4806 non-null   object
 5   cast_x    4806 non-null   object
 6   crew_x    4806 non-null   object
dtypes: int64(1), object(6)
memory usage: 300.4+ KB
```

In [94]:
```
movies["tags"]= movies["genres"]+movies["keywords"]+movies["overview"]+movies["cast_x"]+movies["crew_x"]
```

```
C:\Users\kiran\AppData\Local\Temp\ipykernel_8676\2647787049.py:1: SettingWithCopyWarning:
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/user_guide/indexing.html#returning-a-view-versus-a-copy
  movies["tags"]= movies["genres"]+movies["keywords"]+movies["overview"]+movies["cast_x"]+movies["crew_x"]
```

In [95]:
```
movies.head(2)
```

Out[95]:

	id	genres	keywords	title	overview	cast_x	crew_x	tags
0	19995	[Action, Adventure, Fantasy, ScienceFiction]	[cultureclash, future, spacewar, spacecolony, ...	Avatar	[In, the, 22nd, century,, a, paraplegic, Marin...	[SamWorthington, ZoeSaldana, SigourneyWeaver]	[JamesCameron]	[Action, Adventure, Fantasy, ScienceFiction, c...
1	285	[Adventure, Fantasy, Action]	[ocean, drugabuse, exoticisland, eastindiatrad...	Pirates of the Caribbean: At World's End	[Captain, Barbossa,, long, believed, to, be, d...	[JohnnyDepp, OrlandoBloom, KeiraKnightley]	[GoreVerbinski]	[Adventure, Fantasy, Action, ocean, drugabuse,...

In [97]:
```
new_df=movies.drop(columns=["genres","keywords","overview","cast_x","crew_x"])
```

In [102…
```
new_df["tags"]=new_df["tags"].apply(lambda x:" ".join(x))
```

In [104…
```
new_df["tags"][0]
```

Out[104]:
```
'Action Adventure Fantasy ScienceFiction cultureclash future spacewar spacecolony society spacetravel futuristic romance space alien tribe alienplanet cgi marine soldier battle loveaffair antiwar powerrelations mindandsoul 3d In the 22nd century, a paraplegic Marine is dispatched to the moon Pandora on a unique mission, but becomes torn between following orders and protecting an alien civilization. SamWorthington ZoeSaldana SigourneyWeaver JamesCameron'
```

In [105…
```
new_df["tags"]=new_df["tags"].apply(lambda x:x.lower())
```

In [106…
```
new_df
```

Out[106]:

	id	title	tags
0	19995	Avatar	action adventure fantasy sciencefiction cultur...
1	285	Pirates of the Caribbean: At World's End	adventure fantasy action ocean drugabuse exoti...
2	206647	Spectre	action adventure crime spy basedonnovel secret...
3	49026	The Dark Knight Rises	action crime drama thriller dccomics crimefigh...
4	49529	John Carter	action adventure sciencefiction basedonnovel m...
...	...	...	...
4816	9367	El Mariachi	action crime thriller unitedstates–mexicobarri...
4817	72766	Newlyweds	comedy romance a newlywed couple's honeymoon i...
4818	231617	Signed, Sealed, Delivered	comedy drama romance tvmovie date loveatfirsts...
4819	126186	Shanghai Calling	when ambitious new york attorney sam is sent t...
4820	25975	My Date with Drew	documentary obsession camcorder crush dreamgir...

4806 rows × 3 columns

In [107…
```
# text vectorization
```

In [135…
```
from sklearn.feature_extraction.text import CountVectorizer
cv=CountVectorizer(max_features=5000,stop_words="english")
```

In [136…
```
vectors = cv.fit_transform(new_df["tags"]).toarray()
```

In [137…
```
vectors[0]
```

Out[137]:
```
array([0, 0, 0, ..., 0, 0, 0], dtype=int64)
```

In [138…
```
cv.get_feature_names()
```

Out[138]:
```
['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',
 ...]
```

In [139…
```
import nltk
```

In [140…
```
from nltk.stem.porter import PorterStemmer
ps=PorterStemmer()
```

In [141…
```
def stem(text):
    y=[]

    for i in text.split():
        y.append(ps.stem(i))

    return " ".join(y)
```

In [142…
```
new_df["tags"][0]
```

Out[142]:
```
'action adventur fantasi sciencefict cultureclash futur spacewar spacecoloni societi spacetravel futurist roma
nc space alien tribe alienplanet cgi marin soldier battl loveaffair antiwar powerrel mindandsoul 3d in the 22n
d century, a parapleg marin is dispatch to the moon pandora on a uniqu mission, but becom torn between follow
order and protect an alien civilization. samworthington zoesaldana sigourneyweav jamescameron'
```

In [143…
```
new_df["tags"]=new_df["tags"].apply(stem)
```

In [144…
```
new_df
```

Out[144]:

	id	title	tags
0	19995	Avatar	action adventur fantasi sciencefict culturecla...
1	285	Pirates of the Caribbean: At World's End	adventur fantasi action ocean drugabu exoticis...
2	206647	Spectre	action adventur crime spi basedonnovel secreta...
3	49026	The Dark Knight Rises	action crime drama thriller dccomic crimefight...
4	49529	John Carter	action adventur sciencefict basedonnovel mar m...
...	...	...	...
4816	9367	El Mariachi	action crime thriller unitedstates–mexicobarri...
4817	72766	Newlyweds	comedi romanc a newlyw couple' honeymoon is up...
4818	231617	Signed, Sealed, Delivered	comedi drama romanc tvmovi date loveatfirstsig...
4819	126186	Shanghai Calling	when ambiti new york attorney sam is sent to s...
4820	25975	My Date with Drew	documentari obsess camcord crush dreamgirl eve...

4806 rows × 3 columns

In [145…
```
from sklearn.metrics.pairwise import cosine_similarity
```

In [149…
```
similarity=cosine_similarity(vectors)
```

In [153…
```
similarity[1]
```

Out[153]:
```
array([0.08346223, 1.        , 0.06063391, ..., 0.02378257, 0.        ,
       0.02615329])
```

In [160…
```
def recommend(movie):
    movie_index= new_df[new_df["title"]==movie].index[0]
    distances =similarity[movie_index]
    movies_list = sorted(list(enumerate(distances)),reverse=True,key=lambda x:x[1])[1:6]

    for i in movies_list:
        print(new_df.iloc[i[0]].title)
```

In [161…
```
recommend("Batman Begins")
```

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

In [ ]: