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

In [3]:
```
movies=pd.read_csv('tmdb_5000_movies.csv')
credits=pd.read_csv('tmdb_5000_credits.csv')
```

In [4]:
```
movies.head()
```

Out[4]:

	budget	genres	homepage	id	keywords	origi
0	237000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.avatarmovie.com/	19995	[{"id": 1463, "name": "culture clash"}, {"id":...	
1	300000000	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	http://disney.go.com/disneypictures/pirates/	285	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	
2	245000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.sonypictures.com/movies/spectre/	206647	[{"id": 470, "name": "spy"}, {"id": 818, "name...	
3	250000000	[{"id": 28, "name": "Action"}, {"id": 80, "nam...	http://www.thedarkknightrises.com/	49026	[{"id": 849, "name": "dc comics"}, {"id": 853,...	
4	260000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://movies.disney.com/john-carter	49529	[{"id": 818, "name": "based on novel"}, {"id":...	

In [6]:
```
credits.head(1)['cast'].values
```

Out[6]:
```
array(['[{"cast_id": 242, "character": "Jake Sully", "credit_id": "5602
a8a7c3a3685532001c9a", "gender": 2, "id": 65731, "name": "Sam Worthingt
on", "order": 0}, {"cast_id": 3, "character": "Neytiri", "credit_id":
"52fe48009251416c750ac9cb", "gender": 1, "id": 8691, "name": "Zoe Salda
na", "order": 1}, {"cast_id": 25, "character": "Dr. Grace Augustine",
"credit_id": "52fe48009251416c750aca39", "gender": 1, "id": 10205, "nam
e": "Sigourney Weaver", "order": 2}, {"cast_id": 4, "character": "Col.
Quaritch", "credit_id": "52fe48009251416c750ac9cf", "gender": 2, "id":
32747, "name": "Stephen Lang", "order": 3}, {"cast_id": 5, "character":
"Trudy Chacon", "credit_id": "52fe48009251416c750ac9d3", "gender": 1,
"id": 17647, "name": "Michelle Rodriguez", "order": 4}, {"cast_id": 8,
"character": "Selfridge", "credit_id": "52fe48009251416c750ac9e1", "gen
der": 2, "id": 1771, "name": "Giovanni Ribisi", "order": 5}, {"cast_i
d": 7, "character": "Norm Spellman", "credit_id": "52fe48009251416c750a
c9dd", "gender": 2, "id": 59231, "name": "Joel David Moore", "order":
6}, {"cast_id": 9, "character": "Moat", "credit_id": "52fe48009251416c7
50ac9e5", "gender": 1, "id": 30485, "name": "CCH Pounder", "order": 7},
{"cast_id": 11, "character": "Eytukan", "credit_id": "52fe48009251416c7
50ac9ed", "gender": 2, "id": 15853, "name": "Wes Studi", "order": 8},
```

In [7]:
```
credits.head(1)['crew'].values
```

Out[7]:
```
array(['[{"credit_id": "52fe48009251416c750aca23", "department": "Editi
ng", "gender": 0, "id": 1721, "job": "Editor", "name": "Stephen E. Rivk
in"}, {"credit_id": "539c47ecc3a36810e3001f87", "department": "Art", "g
ender": 2, "id": 496, "job": "Production Design", "name": "Rick Carte
r"}, {"credit_id": "54491c89c3a3680fb4001cf7", "department": "Sound",
"gender": 0, "id": 900, "job": "Sound Designer", "name": "Christopher B
oyes"}, {"credit_id": "54491cb70e0a267480001bd0", "department": "Soun
d", "gender": 0, "id": 900, "job": "Supervising Sound Editor", "name":
"Christopher Boyes"}, {"credit_id": "539c4a4cc3a36810c9002101", "depart
ment": "Production", "gender": 1, "id": 1262, "job": "Casting", "name":
"Mali Finn"}, {"credit_id": "5544ee3b925141499f0008fc", "department":
"Sound", "gender": 2, "id": 1729, "job": "Original Music Composer", "na
me": "James Horner"}, {"credit_id": "52fe48009251416c750ac9c3", "depart
ment": "Directing", "gender": 2, "id": 2710, "job": "Director", "name":
"James Cameron"}, {"credit_id": "52fe48009251416c750ac9d9", "departmen
t": "Writing", "gender": 2, "id": 2710, "job": "Writer", "name": "James
Cameron"}, {"credit_id": "52fe48009251416c750aca17", "department": "Edi
ting", "gender": 2, "id": 2710, "job": "Editor", "name": "James Camero
n"}, {"credit_id": "52fe48009251416c750aca29", "department": "Productio
```

In [11]:
```
movies=movies.merge(credits,on='title')
```

In [9]:
```
movies.shape
```

Out[9]:
```
(4803, 20)
```

In [10]:
```
credits.shape
```

Out[10]:
```
(4803, 4)
```

In [12]:
```
movies.head(1)
```

Out[12]:

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

1 rows × 23 columns

In [19]:
```
#genres
#id
#keywords
#title
#overview
#cast
#crew
movies=movies[['movie_id','title','overview','genres','keywords','cast','cr
```

In [15]:
```
movies['original_language'].value_counts()
```

Out[15]:
```
en    4510
fr      70
es      32
zh      27
de      27
hi      19
ja      16
it      14
ko      12
cn      12
ru      11
pt       9
da       7
sv       5
nl       4
fa       4
th       3
he       3
ta       2
cs       2
ro       2
id       2
ar       2
vi       1
sl       1
ps       1
no       1
ky       1
hu       1
pl       1
af       1
nb       1
tr       1
is       1
xx       1
te       1
el       1
Name: original_language, dtype: int64
```

In [16]:
```
movies.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
```

In [20]:
```
movies.head()
```

Out[20]:

	movie_id	title	overview	genres	keywords	cast	
0	19995	Avatar	In the 22nd century, a paraplegic Marine is di...	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	[{"id": 1463, "name": "culture clash"}, {"id":...	[{"cast_id": 242, "character": "Jake Sully", "...	[{"cr "52fe48009251416c750
1	285	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	[{"cast_id": 4, "character": "Captain Jack Spa...	[{"cr "52fe4232c3a36847f80
2	206647	Spectre	A cryptic message from Bond's past sends him o...	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	[{"id": 470, "name": "spy"}, {"id": 818, "name...	[{"cast_id": 1, "character": "James Bond", "cr...	[{"cr "54805967c3a36829b5(
3	49026	The Dark Knight Rises	Following the death of District Attorney Harve...	[{"id": 28, "name": "Action"}, {"id": 80, "nam...	[{"id": 849, "name": "dc comics"}, {"id": 853,...	[{"cast_id": 2, "character": "Bruce Wayne / Ba...	[{"cr "52fe4781c3a36847f81
4	49529	John Carter	John Carter is a war-weary, former military ca...	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	[{"id": 818, "name": "based on novel"}, {"id":...	[{"cast_id": 5, "character": "John Carter", "c...	[{"cr "52fe479ac3a36847f81

In [24]:
```
movies.isnull().sum()
```

Out[24]:
```
movie_id    0
title       0
overview    0
genres      0
keywords    0
cast        0
crew        0
dtype: int64
```

In [23]:
```
movies.dropna(inplace=True)
```

```
C:\Users\v mouli\AppData\Local\Temp\ipykernel_11484\3786870272.py:1: Setti
ngWithCopyWarning:
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-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  movies.dropna(inplace=True)
```

In [26]:
```
movies.duplicated().sum()
```

Out[26]:
```
0
```

In [28]:
```
movies.iloc[0].genres
```

Out[28]:
```
'[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 1
4, "name": "Fantasy"}, {"id": 878, "name": "Science Fiction"}]'
```

In [35]:
```
#['Action','Adventure','FFantasy','SciFi']
def convert(obj):
    L=[]
    for i in ast.literal_eval(obj):
        L.append(i['name'])
    return L
```

In [33]:
```
import ast
ast.literal_eval('[{"id": 28, "name": "Action"}, {"id": 12, "name": "Advent
```

Out[33]:
```
[{'id': 28, 'name': 'Action'},
 {'id': 12, 'name': 'Adventure'},
 {'id': 14, 'name': 'Fantasy'},
 {'id': 878, 'name': 'Science Fiction'}]
```

In [34]:
```
convert([{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"i
```

Out[34]:
```
['Action', 'Adventure', 'Fantasy', 'Science Fiction']
```

In [38]:
```
movies['genres']=movies['genres'].apply(convert)
```

```
C:\Users\v mouli\AppData\Local\Temp\ipykernel_11484\1567952790.py:1: Setti
ngWithCopyWarning:
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-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  movies['genres']=movies['genres'].apply(convert)
```

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

Out[39]:

	movie_id	title	overview	genres	keywords	cast	
0	19995	Avatar	In the 22nd century, a paraplegic Marine is di...	[Action, Adventure, Fantasy, Science Fiction]	[{"id": 1463, "name": "culture clash"}, {"id":...	[{"cast_id": 242, "character": "Jake Sully", "...	[{"cred "52fe48009251416c750a
1	285	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	[Adventure, Fantasy, Action]	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	[{"cast_id": 4, "character": "Captain Jack Spa...	[{"cred "52fe4232c3a36847f800
2	206647	Spectre	A cryptic message from Bond's past sends him o...	[Action, Adventure, Crime]	[{"id": 470, "name": "spy"}, {"id": 818, "name...	[{"cast_id": 1, "character": "James Bond", "cr...	[{"cred "54805967c3a36829b500
3	49026	The Dark Knight Rises	Following the death of District Attorney Harve...	[Action, Crime, Drama, Thriller]	[{"id": 849, "name": "dc comics"}, {"id": 853,...	[{"cast_id": 2, "character": "Bruce Wayne / Ba...	[{"cred "52fe4781c3a36847f813
4	49529	John Carter	John Carter is a war-weary, former military ca...	[Action, Adventure, Science Fiction]	[{"id": 818, "name": "based on novel"}, {"id":...	[{"cast_id": 5, "character": "John Carter", "c...	[{"cred "52fe479ac3a36847f813

In [42]:
```python
import ast

def convert1(obj):
    L1 = []

    # Check if the object is a string and needs to be evaluated
    if isinstance(obj, str):
        try:
            obj = ast.literal_eval(obj)
        except (ValueError, SyntaxError):
            return []  # Return empty list if parsing fails

    # If it's already a list, extract 'name' values
    if isinstance(obj, list):
        for i in obj:
            if isinstance(i, dict) and 'name' in i:  # Ensure i is a dictio
                L1.append(i['name'])
            elif isinstance(i, str):  # Handle if it's just a list of strin
                L1.append(i)

    return L1
```

In [43]:
```python
movies['keywords'] = movies['keywords'].apply(convert1)
```

In [44]:
```
movies.head()
```

Out[44]:

	movie_id	title	overview	genres	keywords	cast	
0	19995	Avatar	In the 22nd century, a paraplegic Marine is di...	[Action, Adventure, Fantasy, Science Fiction]	[culture clash, future, space war, space colon...	[{"cast_id": 242, "character": "Jake Sully", "...	[{"cred "52fe48009251416c750a
1	285	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	[Adventure, Fantasy, Action]	[ocean, drug abuse, exotic island, east india ...	[{"cast_id": 4, "character": "Captain Jack Spa...	[{"cred "52fe4232c3a36847f800
2	206647	Spectre	A cryptic message from Bond's past sends him o...	[Action, Adventure, Crime]	[spy, based on novel, secret agent, sequel, mi...	[{"cast_id": 1, "character": "James Bond", "cr...	[{"cred "54805967c3a36829b500
3	49026	The Dark Knight Rises	Following the death of District Attorney Harve...	[Action, Crime, Drama, Thriller]	[dc comics, crime fighter, terrorist, secret i...	[{"cast_id": 2, "character": "Bruce Wayne / Ba...	[{"cred "52fe4781c3a36847f813
4	49529	John Carter	John Carter is a war-weary, former military ca...	[Action, Adventure, Science Fiction]	[based on novel, mars, medallion, space travel...	[{"cast_id": 5, "character": "John Carter", "c...	[{"cred "52fe479ac3a36847f813

In [48]:
```
import ast
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 [49]:
```
movies['cast']=movies['cast'].apply(convert3)
```

In [50]:
```
movies.head()
```

Out[50]:

	movie_id	title	overview	genres	keywords	cast	
0	19995	Avatar	In the 22nd century, a paraplegic Marine is di...	[Action, Adventure, Fantasy, Science Fiction]	[culture clash, future, space war, space colon...	[Sam Worthington, Zoe Saldana, Sigourney Weaver]	[{"cre "52fe48009251416c750
1	285	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	[Adventure, Fantasy, Action]	[ocean, drug abuse, exotic island, east india ...	[Johnny Depp, Orlando Bloom, Keira Knightley]	[{"cre "52fe4232c3a36847f80
2	206647	Spectre	A cryptic message from Bond's past sends him o...	[Action, Adventure, Crime]	[spy, based on novel, secret agent, sequel, mi...	[Daniel Craig, Christoph Waltz, Léa Seydoux]	[{"cre "54805967c3a36829b50
3	49026	The Dark Knight Rises	Following the death of District Attorney Harve...	[Action, Crime, Drama, Thriller]	[dc comics, crime fighter, terrorist, secret i...	[Christian Bale, Michael Caine, Gary Oldman]	[{"cre "52fe4781c3a36847f81
4	49529	John Carter	John Carter is a war-weary, former military ca...	[Action, Adventure, Science Fiction]	[based on novel, mars, medallion, space travel...	[Taylor Kitsch, Lynn Collins, Samantha Morton]	[{"cre "52fe479ac3a36847f81

In [51]:
```
def fetch_director(obj):
    L=[]
    for i in ast.literal_eval(obj):
        if i['job']=='Director':
                    L.append(i['name'])
                    break



    return L
```

In [53]:
```
movies['crew']=movies['crew'].apply(fetch_director)
```

In [54]:
```
movies.head()
```

Out[54]:

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

In [56]:
```
movies['overview']=movies['overview'].apply(lambda x:x.split())
```

In [57]:
```
movies.head()
```

Out[57]:

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

In [59]:
```
movies['genres']=movies['genres'].apply(lambda x:[i.replace(" ","") for i i
movies['keywords']=movies['keywords'].apply(lambda x:[i.replace(" ","") for
movies['cast']=movies['cast'].apply(lambda x:[i.replace(" ","") for i in x]
movies['crew']=movies['crew'].apply(lambda x:[i.replace(" ","") for i in x]
```

In [60]:
```
movies.head()
```

Out[60]:

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

In [61]:
```
movies['tags']=movies['overview']+movies['genres']+movies['keywords']+movie
```

In [62]:
```
movies.head()
```

Out[62]:

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

In [63]:
```
new_df=movies[['movie_id','title','tags']]
```

In [64]:
```
new_df
```

Out[64]:

	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...
...	...	...	...
4804	9367	El Mariachi	[El, Mariachi, just, wants, to, play, his, gui...
4805	72766	Newlyweds	[A, newlywed, couple's, honeymoon, is, upended...
4806	231617	Signed, Sealed, Delivered	["Signed,, Sealed,, Delivered", introduces, a,...
4807	126186	Shanghai Calling	[When, ambitious, New, York, attorney, Sam, is...
4808	25975	My Date with Drew	[Ever, since, the, second, grade, when, he, fi...

4806 rows × 3 columns

In [66]:
```
new_df['tags']=new_df['tags'].apply(lambda x:" ".join(x))
```

```
C:\Users\v mouli\AppData\Local\Temp\ipykernel_11484\487797088.py:1: Settin
gWithCopyWarning:
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-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  new_df['tags']=new_df['tags'].apply(lambda x:" ".join(x))
```

In [67]:
```
new_df.head()
```

Out[67]:

	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...

In [68]:
```
new_df['tags']=new_df['tags'].apply(lambda x:x.lower())
```

```
C:\Users\v mouli\AppData\Local\Temp\ipykernel_11484\4224080999.py:1: Setti
ngWithCopyWarning:
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-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  new_df['tags']=new_df['tags'].apply(lambda x:x.lower())
```

In [69]:
```
new_df.head()
```

Out[69]:

	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...

In [70]:
```
new_df['tags'][0]
```

Out[70]:
```
'in the 22nd century, a paraplegic marine is dispatched to the moon pandor
a on a unique mission, but becomes torn between following orders and prote
cting an alien civilization. action adventure fantasy sciencefiction cultu
reclash future spacewar spacecolony society spacetravel futuristic romance
space alien tribe alienplanet cgi marine soldier battle loveaffair antiwar
powerrelations mindandsoul 3d samworthington zoesaldana sigourneyweaver ja
mescameron'
```

In [72]:
```
from sklearn.feature_extraction.text import CountVectorizer
cv=CountVectorizer(max_features=5000,stop_words='english')
```

```
C:\Users\v mouli\anaconda3\lib\site-packages\scipy\__init__.py:146: UserWa
rning: A NumPy version >=1.16.5 and <1.23.0 is required for this version o
f SciPy (detected version 1.26.4
  warnings.warn(f"A NumPy version >={np_minversion} and <{np_maxversion}"
```

In [75]:
```
vectors=cv.fit_transform(new_df['tags']).toarray()
```

In [76]:
```
vectors
```

Out[76]:
```
array([[0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       ...,
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0]], dtype=int64)
```

In [80]:
```
cv.get_feature_names()
```

Out[80]:
```
['000',
 '007',
 '10',
 '100',
 '11',
 '12',
 '13',
 '14',
 '15',
 '16',
 '17',
 '18',
 '18th',
 '19',
 '1930s',
 '1940s',
 '1950',
 '1950s',
 '1960s',
```

In [81]:
```
!pip install nltk
```

```
Requirement already satisfied: nltk in c:\users\v mouli\anaconda3\lib\site
-packages (3.7)
Requirement already satisfied: regex>=2021.8.3 in c:\users\v mouli\anacond
a3\lib\site-packages (from nltk) (2022.3.15)
Requirement already satisfied: tqdm in c:\users\v mouli\anaconda3\lib\site
-packages (from nltk) (4.64.0)
Requirement already satisfied: click in c:\users\v mouli\anaconda3\lib\sit
e-packages (from nltk) (8.0.4)
Requirement already satisfied: joblib in c:\users\v mouli\anaconda3\lib\si
te-packages (from nltk) (1.4.0)
Requirement already satisfied: colorama in c:\users\v mouli\anaconda3\lib
\site-packages (from click->nltk) (0.4.4)
```

In [82]:
```
from nltk.stem.porter import PorterStemmer
ps=PorterStemmer()
```

In [86]:
```
def stem(text):
    y=[]

    for i in text.split():
        y.append(ps.stem(i))
    return " ".join(y)
```

In [85]:
```
ps.stem('dancing')
```

Out[85]:
```
'danc'
```

In [87]:
```
stem('in the 22nd century, a paraplegic marine is dispatched to the moon pa
```

Out[87]:
```
'in the 22nd 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. action adventur fantasi sciencefict cultureclash futur space
war spacecoloni societi spacetravel futurist romanc space alien tribe alie
nplanet cgi marin soldier battl loveaffair antiwar powerrel mindandsoul 3d
samworthington zoesaldana sigourneyweav jamescameron'
```

In [89]:
```
new_df['tags']=new_df['tags'].apply(stem)
```

```
C:\Users\v mouli\AppData\Local\Temp\ipykernel_11484\3514595201.py:1: Setti
ngWithCopyWarning:
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-doc
s/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://
pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-
view-versus-a-copy)
  new_df['tags']=new_df['tags'].apply(stem)
```

In [90]:
```
from sklearn.metrics.pairwise import cosine_similarity
```

In [92]:
```
similarity=cosine_similarity(vectors)
```

In [107]:
```
sorted(list(enumerate(similarity[0])),reverse=True,key=lambda x:x[1])[1:6]
```

Out[107]:
```
[(539, 0.26089696604360174),
 (1194, 0.2581988897471611),
 (507, 0.25302403842552984),
 (260, 0.25110592822973776),
 (1216, 0.24944382578492943)]
```

In [130]:
```
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
    for i in movies_list:
        print(new_df.iloc[i[0]].title)
```

In [122]:
```
recommend('Batman Begins')
```

```
The Dark Knight
The Dark Knight Rises
Batman
Batman & Robin
Batman
```

In [115]:
```
new_df.iloc[1216].title
```

Out[115]:
```
'Aliens vs Predator: Requiem'
```

In [123]:
```
import pickle
```

In [124]:
```
pickle.dump(new_df,open('movies.pkl','wb'))
```

In [126]:
```
new_df['title'].values
```

Out[126]:
```
array(['Avatar', "Pirates of the Caribbean: At World's End", 'Spectre',
       ..., 'Signed, Sealed, Delivered', 'Shanghai Calling',
       'My Date with Drew'], dtype=object)
```

In [128]:
```
pickle.dump(new_df.to_dict(),open('movie_dict.pkl','wb'))
```

In [129]:
```
pickle.dump(similarity,open('similarity.pkl','wb'))
```

In [ ]:
```

```