In [3]:
```python
import pandas as pd
import numpy as np
```

In [4]:
```python
movies=pd.read_csv("C:/Users/zoaah/Downloads/archive/tmdb_5000_movies.csv")
credits=pd.read_csv("C:/Users/zoaah/Downloads/archive/tmdb_5000_credits.csv")
```

In [5]:
```python
movies.head(3)
```

Out[5]:

	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	c pa
1	300000000	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	http://disney.go.com/disneypictures/pirates/	285	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	en	Pirates of the Caribbean: At World's End	E
2	245000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.sonypictures.com/movies/spectre/	206647	[{"id": 470, "name": "spy"}, {"id": 818, "name...	en	Spectre	s

In [6]:
```python
credits.head(1)
```

Out[6]:

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

In [7]:
```python
credits.head(1)['cast']
```

Out[7]:
```
0    [{"cast_id": 242, "character": "Jake Sully", "...
Name: cast, dtype: object
```

In [8]:
```python
credits.head(1)['cast'].values
```

Out[8]:
array(['[{"cast_id": 242, "character": "Jake Sully", "credit_id": "5602a8a7c3a3685532001c9a", "gender": 2, "id": 65731, "name": "Sam Worthington", "order": 0}, {"cast_id": 3, "character": "Neytiri", "credit_id": "52fe48009251416c750ac9cb", "gender": 1, "id": 8691, "name": "Zoe Saldana", "order": 1}, {"cast_id": 25, "character": "Dr. Grace Augustine", "credit_id": "52fe48009251416c750aca39", "gender": 1, "id": 10205, "name": "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", "gender": 2, "id": 1771, "name": "Giovanni Ribisi", "order": 5}, {"cast_id": 7, "character": "Norm Spellman", "credit_id": "52fe48009251416c750ac9dd", "gender": 2, "id": 59231, "name": "Joel David Moore", "order": 6}, {"cast_id": 9, "character": "Moat", "credit_id": "52fe48009251416c750ac9e5", "gender": 1, "id": 30485, "name": "CCH Pounder", "order": 7}, {"cast_id": 11, "character": "Eytukan", "credit_id": "52fe48009251416c750ac9ed", "gender": 2, "id": 15853, "name": "Wes Studi", "order": 8}, {"cast_id": 10, "character": "Tsu\'Tey", "credit_id": "52fe48009251416c750ac9e9", "gender": 2, "id": 10964, "name": "Laz Alonso", "order": 9}, {"cast_id": 12, "character": "Dr. Max Patel", "credit_id": "52fe48009251416c750ac9f1", "gender": 2, "id": 95697, "name": "Dileep Rao", "order": 10}, {"cast_id": 13, "character": "Lyle Wainfleet", "credit_id": "52fe48009251416c750ac9f5", "gender": 2, "id": 98215, "name": "Matt Gerald", "order": 11},

In [9]:
```python
#merging of two Dataframes on the basis of title
movies.merge(credits,on='title')
```

Out[9]:

	budget	genres	homepage	id	keywords	original_langua
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":...	
...	...	...	...	...	...	
4804	220000	[{"id": 28, "name": "Action"}, {"id": 80, "nam...	NaN	9367	[{"id": 5616, "name": "united states\u2013mexi...	
4805	9000	[{"id": 35, "name": "Comedy"}, {"id": 10749, "...	NaN	72766	[]	
4806	0	[{"id": 35, "name": "Comedy"}, {"id": 18, "nam...	http://www.hallmarkchannel.com/signedsealeddel...	231617	[{"id": 248, "name": "date"}, {"id": 699, "nam...	
4807	0	[]	http://shanghaicalling.com/	126186	[]	
4808	0	[{"id": 99, "name": "Documentary"}]	NaN	25975	[{"id": 1523, "name": "obsession"}, {"id": 224...	

4809 rows × 23 columns

In [10]:
```python
movies.merge(credits,on='title') .shape
```

Out[10]: (4809, 23)

In [11]:
```
movies.shape
```

Out[11]: (4803, 20)

In [12]:
```
credits.shape
```

Out[12]: (4803, 4)

In [13]:
```
#new dataframe of movies
movies=movies.merge(credits,on='title')
```

In [14]:
```
movies.head(1)
```

Out[14]:

	budget	genres	homepage	id	keywords	original_language	original_title	overview	popularit
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...	150.43757

1 rows × 23 columns

In [15]:
```python
#keeping certain columns and removing the columns that are not necessary for our analysis
#genres
#id
#title
#keywords
#overview
#cast
#crew
movies[['movie_id','title','overview','genres','keywords','cast','crew']]
```

Out[15]:

	movie_id	title	overview	genres	keywords	cast	crew
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", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
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...	[{"credit_id": "52fe4232c3a36847f800b579", "de...
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...	[{"credit_id": "54805967c3a36829b5002c41", "de...
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...	[{"credit_id": "52fe4781c3a36847f81398c3", "de...
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...	[{"credit_id": "52fe479ac3a36847f813eaa3", "de...
...	...	...	...	...	...	...	...
4804	9367	El Mariachi	El Mariachi just wants to play his guitar and ...	[{"id": 28, "name": "Action"}, {"id": 80, "nam...	[{"id": 5616, "name": "united states\u2013mexi...	[{"cast_id": 1, "character": "El Mariachi", "c...	[{"credit_id": "52fe44eec3a36847f80b280b", "de...
4805	72766	Newlyweds	A newlywed couple's honeymoon is upended by th...	[{"id": 35, "name": "Comedy"}, {"id": 10749, "...	[]	[{"cast_id": 1, "character": "Buzzy", "credit_...	[{"credit_id": "52fe487dc3a368484e0fb013", "de...
4806	231617	Signed, Sealed, Delivered	"Signed, Sealed, Delivered" introduces a dedic...	[{"id": 35, "name": "Comedy"}, {"id": 18, "nam...	[{"id": 248, "name": "date"}, {"id": 699, "nam...	[{"cast_id": 8, "character": "Oliver O\u2019To...	[{"credit_id": "52fe4df3c3a36847f8275ecf", "de...
4807	126186	Shanghai Calling	When ambitious New York attorney Sam is sent t...	[]	[]	[{"cast_id": 3, "character": "Sam", "credit_id...	[{"credit_id": "52fe4ad9c3a368484e16a36b", "de...
4808	25975	My Date with Drew	Ever since the second grade when he first saw ...	[{"id": 99, "name": "Documentary"}]	[{"id": 1523, "name": "obsession"}, {"id": 224...	[{"cast_id": 3, "character": "Herself", "credi...	[{"credit_id": "58ce021b9251415a390165d9", "de...

4809 rows × 7 columns

In [16]:
```python
# final new dataframe for the recommender system
movies=movies[['movie_id','title','overview','genres','keywords','cast','crew']]
```

In [17]:
```python
movies.head()
```

Out[17]:

	movie_id	title	overview	genres	keywords	cast	crew
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", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
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...	[{"credit_id": "52fe4232c3a36847f800b579", "de...
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...	[{"credit_id": "54805967c3a36829b5002c41", "de...
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...	[{"credit_id": "52fe4781c3a36847f81398c3", "de...
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...	[{"credit_id": "52fe479ac3a36847f813eaa3", "de...

In [18]:
```python
#detecting missing values
movies.isnull().sum()
```

Out[18]:
```
movie_id    0
title       0
overview    3
genres      0
keywords    0
cast        0
crew        0
dtype: int64
```

In [19]:
```python
movies.dropna(inplace=True)
```

In [20]:
```python
#removed the movies that had missing values in overview column
movies.isnull().sum()
```

Out[20]:
```
movie_id    0
title       0
overview    0
genres      0
keywords    0
cast        0
crew        0
dtype: int64
```

In [21]:
```python
#to check for duplicate data
movies.duplicated().sum()
```

Out[21]: 0

In [22]:
```python
#reformating of columns
#here the format is not properly shown
movies.iloc[0].genres
```

Out[22]: '[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 878, "name": "Science Fiction"}]'

In [23]:
```python
#converting string of list into list
import ast
ast.literal_eval('[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name"
)
```

Out[23]:
```
[{'id': 28, 'name': 'Action'},
 {'id': 12, 'name': 'Adventure'},
 {'id': 14, 'name': 'Fantasy'},
 {'id': 878, 'name': 'Science Fiction'}]
```

In [24]:
```python
def convert(obj):
    L=[]
    for i in ast.literal_eval(obj):
        L.append(i['name'])
    return L
```

In [25]:
```python
movies['genres'].apply(convert)
```

Out[25]:
```
0          [Action, Adventure, Fantasy, Science Fiction]
1                         [Adventure, Fantasy, Action]
2                            [Action, Adventure, Crime]
3                       [Action, Crime, Drama, Thriller]
4               [Action, Adventure, Science Fiction]
                            ...
4804                         [Action, Crime, Thriller]
4805                                  [Comedy, Romance]
4806              [Comedy, Drama, Romance, TV Movie]
4807                                                []
4808                                     [Documentary]
Name: genres, Length: 4806, dtype: object
```

In [26]:
```python
movies['genres']=movies['genres'].apply(convert)
```

In [27]:
```python
movies.head()
```

Out[27]:

	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]	[{"id": 1463, "name": "culture clash"}, {"id":...	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
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...	[{"credit_id": "52fe4232c3a36847f800b579", "de...
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...	[{"credit_id": "54805967c3a36829b5002c41", "de...
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...	[{"credit_id": "52fe4781c3a36847f81398c3", "de...
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...	[{"credit_id": "52fe479ac3a36847f813eaa3", "de...

In [28]: `movies['keywords'].apply(convert)`

Out[28]:
```
0       [culture clash, future, space war, space colon...
1       [ocean, drug abuse, exotic island, east india ...
2       [spy, based on novel, secret agent, sequel, mi...
3       [dc comics, crime fighter, terrorist, secret i...
4       [based on novel, mars, medallion, space travel...
                              ...
4804    [united states-mexico barrier, legs, arms, pap...
4805                                                   []
4806    [date, love at first sight, narration, investi...
4807                                                   []
4808             [obsession, camcorder, crush, dream girl]
Name: keywords, Length: 4806, dtype: object
```

In [29]: `movies['keywords']=movies['keywords'].apply(convert)`

In [30]: `movies.head()`

Out[30]:

	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...	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
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...	[{"credit_id": "52fe4232c3a36847f800b579", "de...
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...	[{"credit_id": "54805967c3a36829b5002c41", "de...
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...	[{"credit_id": "52fe4781c3a36847f81398c3", "de...
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...	[{"credit_id": "52fe479ac3a36847f813eaa3", "de...

In [31]:
```python
#cast column for first movie
movies['cast'][0]
```

Out[31]: '[{"cast_id": 242, "character": "Jake Sully", "credit_id": "5602a8a7c3a3685532001c9a", "gender": 2, "id": 65731, "name": "Sam Worthington", "order": 0}, {"cast_id": 3, "character": "Neytiri", "credit_id": "52fe48009251416c750ac9cb", "gender": 1, "id": 8691, "name": "Zoe Saldana", "order": 1}, {"cast_id": 25, "character": "Dr. Grace Augustine", "credit_id": "52fe48009251416c750aca39", "gender": 1, "id": 10205, "name": "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", "gender": 2, "id": 1771, "name": "Giovanni Ribisi", "order": 5}, {"cast_id": 7, "character": "Norm Spellman", "credit_id": "52fe48009251416c750ac9dd", "gender": 2, "id": 59231, "name": "Joel David Moore", "order": 6}, {"cast_id": 9, "character": "Moat", "credit_id": "52fe48009251416c750ac9e5", "gender": 1, "id": 30485, "name": "CCH Pounder", "order": 7}, {"cast_id": 11, "character": "Eytukan", "credit_id": "52fe48009251416c750ac9ed", "gender": 2, "id": 15853, "name": "Wes Studi", "order": 8}, {"cast_id": 10, "character": "Tsu\'Tey", "credit_id": "52fe48009251416c750ac9e9", "gender": 2, "id": 10964, "name": "Laz Alonso", "order": 9}, {"cast_id": 12, "character": "Dr. Max Patel", "credit_id": "52fe48009251416c750ac9f1", "gender": 2, "id": 95697, "name": "Dileep Rao", "order": 10}, {"cast_id": 13, "character": "Lyle Wainfleet", "credit_id": "52fe48009251416c750ac9f5", "gender": 2, "id": 98215, "name": "Matt Gerald", "order": 11}, {"c

In [32]:
```python
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 [33]:
```python
movies['cast'].apply(convert3)
```

Out[33]:
```
0          [Sam Worthington, Zoe Saldana, Sigourney Weaver]
1            [Johnny Depp, Orlando Bloom, Keira Knightley]
2            [Daniel Craig, Christoph Waltz, Léa Seydoux]
3            [Christian Bale, Michael Caine, Gary Oldman]
4          [Taylor Kitsch, Lynn Collins, Samantha Morton]
                              ...
4804     [Carlos Gallardo, Jaime de Hoyos, Peter Marqua...
4805        [Edward Burns, Kerry Bishé, Marsha Dietlein]
4806         [Eric Mabius, Kristin Booth, Crystal Lowe]
4807          [Daniel Henney, Eliza Coupe, Bill Paxton]
4808     [Drew Barrymore, Brian Herzlinger, Corey Feldman]
Name: cast, Length: 4806, dtype: object
```

In [34]:
```python
movies['cast']=movies['cast'].apply(convert3)
```

In [35]:
```python
movies.head()
```

Out[35]:

	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]	[{"credit_id": "52fe48009251416c750aca23", "de...
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]	[{"credit_id": "52fe4232c3a36847f800b579", "de...
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]	[{"credit_id": "54805967c3a36829b5002c41", "de...
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]	[{"credit_id": "52fe4781c3a36847f81398c3", "de...
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]	[{"credit_id": "52fe479ac3a36847f813eaa3", "de...

In [36]:
```python
movies['cast'][0]
```

Out[36]:
```
['Sam Worthington', 'Zoe Saldana', 'Sigourney Weaver']
```

In [37]:
```python
#cast column for first movie
movies['crew'][0]
```

Out[37]: '[{"credit_id": "52fe48009251416c750aca23", "department": "Editing", "gender": 0, "id": 1721, "job": "Editor", "name": "Stephen E. Rivkin"}, {"credit_id": "539c47ecc3a36810e3001f87", "depa rtment": "Art", "gender": 2, "id": 496, "job": "Production Design", "name": "Rick Carter"}, {"credit_id": "54491c89c3a3680fb4001cf7", "department": "Sound", "gender": 0, "id": 900, "jo b": "Sound Designer", "name": "Christopher Boyes"}, {"credit_id": "54491cb70e0a267480001bd0", "department": "Sound", "gender": 0, "id": 900, "job": "Supervising Sound Editor", "name": "Chr istopher Boyes"}, {"credit_id": "539c4a4cc3a36810c9002101", "department": "Production", "gende r": 1, "id": 1262, "job": "Casting", "name": "Mali Finn"}, {"credit_id": "5544ee3b925141499f00 08fc", "department": "Sound", "gender": 2, "id": 1729, "job": "Original Music Composer", "nam e": "James Horner"}, {"credit_id": "52fe48009251416c750ac9c3", "department": "Directing", "gen der": 2, "id": 2710, "job": "Director", "name": "James Cameron"}, {"credit_id": "52fe480092514 16c750ac9d9", "department": "Writing", "gender": 2, "id": 2710, "job": "Writer", "name": "Jame s Cameron"}, {"credit_id": "52fe48009251416c750aca17", "department": "Editing", "gender": 2, "id": 2710, "job": "Editor", "name": "James Cameron"}, {"credit_id": "52fe48009251416c750aca2 9", "department": "Production", "gender": 2, "id": 2710, "job": "Producer", "name": "James Cam eron"}, {"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 Menzie s"}, {"credit_id": "549598c3c3a3686ae9004383", "department": "Visual Effects", "gender": 0, "i

In [39]:
```python
def fetch_director(obj):
    L=[]
    for i in ast.literal_eval(obj):
        if i ['job']=='Director':
            L.append(i['name'])
            break
    return L
```

In [40]:
```python
movies['crew'].apply(fetch_director)
```

Out[40]:
```
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: 4806, dtype: object
```

In [41]:
```python
movies['crew']=movies['crew'].apply(fetch_director)
```

In [42]:
```python
movies.head()
```

Out[42]:

	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 [43]:
```python
movies['overview'][0] #its a string
```

Out[43]:
```
'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.'
```

In [44]:
```python
#to convert string into list
movies['overview'].apply(lambda x:x.split())
```

Out[44]:
```
0       [In, the, 22nd, century,, a, paraplegic, Marin...
1       [Captain, Barbossa,, long, believed, to, be, d...
2       [A, cryptic, message, from, Bond's, past, send...
3       [Following, the, death, of, District, Attorney...
4       [John, Carter, is, a, war-weary,, former, mili...
                              ...
4804    [El, Mariachi, just, wants, to, play, his, gui...
4805    [A, newlywed, couple's, honeymoon, is, upended...
4806    ["Signed,, Sealed,, Delivered", introduces, a,...
4807    [When, ambitious, New, York, attorney, Sam, is...
4808    [Ever, since, the, second, grade, when, he, fi...
Name: overview, Length: 4806, dtype: object
```

In [45]:
```python
movies['overview']=movies['overview'].apply(lambda x:x.split())
```

In [46]:
```python
movies.head()
```

Out[46]:

	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 [47]:
```python
movies['tags']=movies['overview']+movies['genres']+movies['keywords']+movies['cast']+movies['crew'
```

In [48]:
```python
movies.head()
```

Out[48]:

	movie_id	title	overview	genres	keywords	cast	crew	tags
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]	[In, the, 22nd, century,, a, paraplegic, Marin...
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]	[Captain, Barbossa,, long, believed, to, be, d...
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]	[A, cryptic, message, from, Bond's, past, send...
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]	[Following, the, death, of, District, Attorney...
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]	[John, Carter, is, a, war-weary,, former, mili...

In [51]:
```python
new_df=movies[['movie_id','title','tags']]
```

In [52]:
```python
new_df
```

Out[52]:

	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 [53]:
```python
new_df['tags'].apply(lambda x:" ".join(x))
```

Out[53]:
```
0       In the 22nd century, a paraplegic Marine is di...
1       Captain Barbossa, long believed to be dead, ha...
2       A cryptic message from Bond's past sends him o...
3       Following the death of District Attorney Harve...
4       John Carter is a war-weary, former military ca...
                              ...
4804    El Mariachi just wants to play his guitar and ...
4805    A newlywed couple's honeymoon is upended by th...
4806    "Signed, Sealed, Delivered" introduces a dedic...
4807    When ambitious New York attorney Sam is sent t...
4808    Ever since the second grade when he first saw ...
Name: tags, Length: 4806, dtype: object
```

In [54]:
```python
new_df['tags']=new_df['tags'].apply(lambda x:" ".join(x))
```

```
C:\Users\zoaah\AppData\Local\Temp\ipykernel_10508\487797088.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/ind
exing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guid
e/indexing.html#returning-a-view-versus-a-copy)
  new_df['tags']=new_df['tags'].apply(lambda x:" ".join(x))
```

In [55]:
```python
new_df.head()
```

Out[55]:

	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 [65]:
```python
new_df['tags'][0]
```

Out[65]: 'in the 22nd century, a parapleg marin is dispatch to the moon pandora on a uniqu mission, but be com torn between follow order and protect an alien civilization. action adventur fantasi scienc f iction cultur clash futur space war space coloni societi space travel futurist romanc space alien tribe alien planet cgi marin soldier battl love affair anti war power relat mind and soul 3d sam worthington zoe saldana sigourney weaver jame cameron'

In [66]:
```python
new_df['tags'][1]
```

Out[66]: "captain barbossa, long believ to be dead, ha come back to life and is head to the edg of the ear th with will turner and elizabeth swann. but noth is quit as it seems. adventur fantasi action oc ean drug abus exot island east india trade compani love of one' life traitor shipwreck strong wom an ship allianc calypso afterlif fighter pirat swashbuckl aftercreditssting johnni depp orlando b loom keira knightley gore verbinski"

In [56]:
```python
!pip install nltk
```

```
Requirement already satisfied: nltk in c:\users\zoaah\anaconda4\lib\site-packages (3.7)
Requirement already satisfied: click in c:\users\zoaah\anaconda4\lib\site-packages (from nltk)
(8.0.4)
Requirement already satisfied: joblib in c:\users\zoaah\anaconda4\lib\site-packages (from nltk)
(1.1.1)
Requirement already satisfied: tqdm in c:\users\zoaah\anaconda4\lib\site-packages (from nltk) (4.
64.1)
Requirement already satisfied: regex>=2021.8.3 in c:\users\zoaah\anaconda4\lib\site-packages (fro
m nltk) (2022.7.9)
Requirement already satisfied: colorama in c:\users\zoaah\anaconda4\lib\site-packages (from click
->nltk) (0.4.6)
```

In [75]:
```python
import nltk
```

In [76]:
```python
from nltk.stem.porter import PorterStemmer
ps=PorterStemmer()
```

In [77]:
```python
def stem(text):
    y=[]
    for i in text.split():
        y.append(ps.stem(i))
    return " ".join(y)
```

In [78]:
```python
new_df['tags'].apply(stem)
```

Out[78]:
```
0          in the 22nd century, a parapleg marin is dispa...
1          captain barbossa, long believ to be dead, ha c...
2          a cryptic messag from bond' past send him on a...
3          follow the death of district attorney harvey d...
4          john carter is a war-weary, former militari ca...
                                 ...
4804       el mariachi just want to play hi guitar and ca...
4805       a newlyw couple' honeymoon is upend by the arr...
4806       "signed, sealed, delivered" introduc a dedic q...
4807       when ambiti new york attorney sam is sent to s...
4808       ever sinc the second grade when he first saw h...
Name: tags, Length: 4806, dtype: object
```

In [79]:
```python
new_df['tags']=new_df['tags'].apply(stem)
```

```
C:\Users\zoaah\AppData\Local\Temp\ipykernel_10508\3514595201.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/ind
exing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guid
e/indexing.html#returning-a-view-versus-a-copy)
  new_df['tags']=new_df['tags'].apply(stem)
```

In [80]:
```python
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(max_features=5000,stop_words='english')
```

In [81]:
```python
cv.fit_transform(new_df['tags']).toarray()
```

Out[81]:
```
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 [82]:
```python
cv.fit_transform(new_df['tags']).toarray().shape
```

Out[82]:
```
(4806, 5000)
```

In [83]:
```python
vector = cv.fit_transform(new_df['tags']).toarray()
```

In [84]:
```python
vector
```

Out[84]:
```
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 [85]:
```python
vector[0]
```

Out[85]: `array([0, 0, 0, ..., 0, 0, 0], dtype=int64)`

In [86]:
```python
cv.get_feature_names_out()
```

Out[86]: `array(['000', '007', '10', ..., 'zooey', 'zucker', 'zwick'], dtype=object)`

In [87]:
```python
len(cv.get_feature_names_out())
```

Out[87]: `5000`

In [88]:
```python
from sklearn.metrics.pairwise import cosine_similarity
```

In [89]:
```python
cosine_similarity(vector)
```

Out[89]:
```
array([[1.        , 0.06692745, 0.06554735, ..., 0.04442617, 0.03121953,
        0.        ],
       [0.06692745, 1.        , 0.04167571, ..., 0.03766218, 0.        ,
        0.01756008],
       [0.06554735, 0.04167571, 1.        , ..., 0.01844278, 0.07776158,
        0.        ],
       ...,
       [0.04442617, 0.03766218, 0.01844278, ..., 1.        , 0.03513642,
        0.03108349],
       [0.03121953, 0.        , 0.07776158, ..., 0.03513642, 1.        ,
        0.06552976],
       [0.        , 0.01756008, 0.        , ..., 0.03108349, 0.06552976,
        1.        ]])
```

In [90]:
```python
cosine_similarity(vector).shape
```

Out[90]: `(4806, 4806)`

In [91]:
```python
similarity = cosine_similarity(vector)
```

In [92]:
```python
similarity
```

Out[92]:
```
array([[1.        , 0.06692745, 0.06554735, ..., 0.04442617, 0.03121953,
        0.        ],
       [0.06692745, 1.        , 0.04167571, ..., 0.03766218, 0.        ,
        0.01756008],
       [0.06554735, 0.04167571, 1.        , ..., 0.01844278, 0.07776158,
        0.        ],
       ...,
       [0.04442617, 0.03766218, 0.01844278, ..., 1.        , 0.03513642,
        0.03108349],
       [0.03121953, 0.        , 0.07776158, ..., 0.03513642, 1.        ,
        0.06552976],
       [0.        , 0.01756008, 0.        , ..., 0.03108349, 0.06552976,
        1.        ]])
```

In [95]:
```python
similarity[1]
```

Out[95]:
```
array([0.06692745, 1.        , 0.04167571, ..., 0.03766218, 0.        ,
       0.01756008])
```

```
In [96]: sorted(list(enumerate(similarity[0])),reverse=True,key = lambda x: x[1])[1:6]
```

```
Out[96]: [(2409, 0.4935676473963735),
          (4336, 0.4113063728303113),
          (1537, 0.3880712182748949),
          (3162, 0.3856962663899156),
          (373, 0.3804429551263411)]
```

```
In [97]: def recommend(movie):
             movie_index = new_df[new_df['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(i[0])
```

```
In [98]: recommend('Avatar')
```

```
2409
4336
1537
3162
373
```

```
In [99]: new_df.iloc[1216]
```

```
Out[99]: movie_id                                      440
         title                 Aliens vs Predator: Requiem
         tags         a sequel to 2004' alien vs. predator, the icon...
         Name: 1216, dtype: object
```

```
In [100]: new_df.iloc[1216].title
```

```
Out[100]: 'Aliens vs Predator: Requiem'
```

```
In [101]: new_df.iloc[4336].title
```

```
Out[101]: 'Silent Running'
```

```
In [102]: def recommend(movie):
              movie_index = new_df[new_df['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(new_df.iloc[i[0]].title)
                  =
```

```
In [103]: recommend('Avatar')
```

```
Aliens
Silent Running
Moonraker
Alien
Mission to Mars
```

```
In [104]: recommend('Batman Begins')
```

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

In [105]: ```python
import pickle
```

In [106]: ```python
pickle.dump(new_df,open('movies.pkl','wb'))
```

In [107]: ```python
new_df
```

Out[107]:

| | 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... |
| ... | ... | ... | ... |
| 4804 | 9367 | El Mariachi | el mariachi just want to play hi guitar and ca... |
| 4805 | 72766 | Newlyweds | a newlyw couple' honeymoon is upend by the arr... |
| 4806 | 231617 | Signed, Sealed, Delivered | "signed, sealed, delivered" introduc a dedic q... |
| 4807 | 126186 | Shanghai Calling | when ambiti new york attorney sam is sent to s... |
| 4808 | 25975 | My Date with Drew | ever sinc the second grade when he first saw h... |

4806 rows × 3 columns

In [108]: ```python
new_df['title'].values
```

Out[108]: ```
array(['Avatar', "Pirates of the Caribbean: At World's End", 'Spectre',
       ..., 'Signed, Sealed, Delivered', 'Shanghai Calling',
       'My Date with Drew'], dtype=object)
```

In [109]: ```python
new_df.to_dict()
```

Out[109]: ```
{'movie_id': {0: 19995,
  1: 285,
  2: 206647,
  3: 49026,
  4: 49529,
  5: 559,
  6: 38757,
  7: 99861,
  8: 767,
  9: 209112,
  10: 1452,
  11: 10764,
  12: 58,
  13: 57201,
  14: 49521,
  15: 2454,
  16: 24428,
  17: 1865,
  18: 41154,
```

In [110]: ```python
pickle.dump(new_df.to_dict(),open('movies_dict.pkl','wb'))
```

In [111]: ```python
pickle.dump(similarity,open('similarity.pkl','wb'))
```

In [ ]: