Movie Recommendation System

EDA and Prediction System (Notebook)

In [203]:
```
# Importing Libraries
import numpy as np
import pandas as pd
```

Load the datasets

In [204]:
```
# Reading Files
credits_df = pd.read_csv('tmdb_5000_credits.csv')
```

In [205]:
```
movies_df = pd.read_csv('tmdb_5000_movies.csv')
```

In [206]:
```
# Printing Top 5 Data
credits_df.head()
```

Out[206]:

	movie_id	title	cast	crew
0	19995	Avatar	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
1	285	Pirates of the Caribbean: At World's End	[{"cast_id": 4, "character": "Captain Jack Spa...	[{"credit_id": "52fe4232c3a36847f800b579", "de...
2	206647	Spectre	[{"cast_id": 1, "character": "James Bond", "cr...	[{"credit_id": "54805967c3a36829b5002c41", "de...
3	49026	The Dark Knight Rises	[{"cast_id": 2, "character": "Bruce Wayne / Ba...	[{"credit_id": "52fe4781c3a36847f81398c3", "de...
4	49529	John Carter	[{"cast_id": 5, "character": "John Carter", "c...	[{"credit_id": "52fe479ac3a36847f813eaa3", "de...

In [207]:
```
# Printing Top 5 Data
movies_df.head()
```

Out[207]:

	budget	genres	homepage	id	keywords	original_language	original_title	overview	popularity
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...	150.4375
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	Captain Barbossa, long believed to be dead, ha...	139.0826
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	A cryptic message from Bond's past sends him o...	107.3767
3	250000000	[{"id": 28, "name": "Action"}, {"id": 80, "nam...	http://www.thedarkknightrises.com/	49026	[{"id": 849, "name": "dc comics"}, {"id": 853,...	en	The Dark Knight Rises	Following the death of District Attorney Harve...	112.3129
4	260000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://movies.disney.com/john-carter	49529	[{"id": 818, "name": "based on novel"}, {"id":...	en	John Carter	John Carter is a war-weary, former military ca...	43.9269

In [208]:
```
# Merging two different datasets into one to work as a one dataframe based on 'title' or 'id'
merge_df = pd.merge(movies_df,credits_df,on='title')
```

In [209]:
```
# Top 5 Merge Data
merge_df.head()
```

Out[209]:

	budget	genres	homepage	id	keywords	original_language	original_title	overview	popularity
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...	150.4375
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	Captain Barbossa, long believed to be dead, ha...	139.0826
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	A cryptic message from Bond's past sends him o...	107.3767
3	250000000	[{"id": 28, "name": "Action"}, {"id": 80, "nam...	http://www.thedarkknightrises.com/	49026	[{"id": 849, "name": "dc comics"}, {"id": 853,...	en	The Dark Knight Rises	Following the death of District Attorney Harve...	112.3129
4	260000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://movies.disney.com/john-carter	49529	[{"id": 818, "name": "based on novel"}, {"id":...	en	John Carter	John Carter is a war-weary, former military ca...	43.9269

5 rows × 23 columns

Overview

In [210]:
```
print(merge_df.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
None
```

Insight:

- The dataset contains information about movies, encompassing various aspects such as budget, genres, production details, release date, and popularity, cast, crew.
- There are numeric columns for quantitative analysis, categorical columns for categorization, and text data for potential natural language processing tasks.

In [211]:
```
print(merge_df.shape)
```

```
(4809, 23)
```

Insight:

- There are in total 4809 movies with 23 features in it the dataframe

Selecting only require fields

In [212]:
```
features = (merge_df.columns.tolist())
print(features)
```

```
['budget', 'genres', 'homepage', 'id', 'keywords', 'original_language', 'original_title', 'overview', '
popularity', 'production_companies', 'production_countries', 'release_date', 'revenue', 'runtime', 'spo
ken_languages', 'status', 'tagline', 'title', 'vote_average', 'vote_count', 'movie_id', 'cast', 'crew']
```

In [213]:
```
# genres, id, keywords, title, overview, vote_average, cast, crew are importants for recommendation
selected_features = ['id','title','genres','keywords','overview','cast','crew']
movies = merge_df[selected_features]
movies.head()
```

Out[213]:

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

In [134]:
```
movies.shape
```

Out[134]:
```
(4809, 7)
```

Insight:

- There are 4807 movies with selecting only 7 features

In [214]:
```
movies.info()
```

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

In [215]:
```
movies.isnull().sum()
```

Out[215]:
```
id          0
title       0
genres      0
keywords    0
overview    3
cast        0
crew        0
dtype: int64
```

Insight

- There are three missing values in the overview column
- Different Imputation techniques could be performed to fill those values.
- For smaller amount of overview, we could remove it

In [216]:
```
# Removing the missing values rows
movies.dropna(inplace=True)
```

```
/var/folders/yb/v5f667y93j393zcsz2t_26qw0000gn/T/ipykernel_13438/3891106807.py:2: SettingWithCopyWarnin
g:
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 (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  movies.dropna(inplace=True)
```

In [217]:
```
# Rechecking for missing values
movies.isnull().sum()
```

Out[217]:
```
id          0
title       0
genres      0
keywords    0
overview    0
cast        0
crew        0
dtype: int64
```

In [218]:
```
# Checking for duplicates
movies.duplicated().sum()
```

Out[218]:
```
0
```

In [219]:
```
movies.iloc[0].genres
```

Out[219]:
```
'[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 878, "name": "Science Fiction"}]'
```

Insight:

- Since it is in form of list of dictionaries, we only need the name key's value for drawing the genres name associated with the movie.
- ast.literal_eval is used to safely evaluate the expression represented as a string containing a Python literal or container display

In [220]:
```
import ast
# Function to get the name key's value in a List
def convert(obj):
    L = []
    for i in ast.literal_eval(obj):
        L.append(i['name'])
    return L
```

In [221]:
```
# Applying the function and updating the columns
movies['genres'] = movies['genres'].apply(convert)
movies['keywords'] = movies['keywords'].apply(convert)
```

```
/var/folders/yb/v5f667y93j393zcsz2t_26qw0000gn/T/ipykernel_13438/2687346810.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 (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  movies['genres'] = movies['genres'].apply(convert)
/var/folders/yb/v5f667y93j393zcsz2t_26qw0000gn/T/ipykernel_13438/2687346810.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 (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
  movies['keywords'] = movies['keywords'].apply(convert)
```

In [222]:
```
# Checking for the changes
movies.head()
```

Out[222]:

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

In [223]:
```
# 0th index values
movies['cast'][0]
```

Out[223]:
```
'[{"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}, {"cast_id": 32, "character": "Private Fike", "credit_id": "52fe48009251416c750aca5b", "gender": 2, "id": 154153, "name": "Sean Anthon
```

In [224]:
```
# Function to get first 4 actor/actress' name
def convert4(obj):
    L = []
    count = 0;
    for i in ast.literal_eval(obj):
        if(count!=4):
            L.append(i['name'])
            count+=1
        else:
            break
    return L
```

In [225]:
```
# Updating the cast column
movies['cast'] = movies['cast'].apply(convert4)
```

```
/var/folders/yb/v5f667y93j393zcsz2t_26qw0000gn/T/ipykernel_13438/2674671810.py:2: SettingWithCopyWarnin
g:
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 (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.h
tml#returning-a-view-versus-a-copy)
  movies['cast'] = movies['cast'].apply(convert4)
```

In [226]:
```
# Checking for the first value
movies['cast'][0]
```

Out[226]:
```
['Sam Worthington', 'Zoe Saldana', 'Sigourney Weaver', 'Stephen Lang']
```

- In the crew section we only need to derive the Director name for "job" value

In [227]:
```
movies['crew'][0]
```

Out[227]:

```
'[{"credit_id": "52fe48009251416c750aca23", "department": "Editing", "gender": 0, "id": 1721, "job":
"Editor", "name": "Stephen E. Rivkin"}, {"credit_id": "539c47ecc3a36810e3001f87", "department": "Ar
t", "gender": 2, "id": 496, "job": "Production Design", "name": "Rick Carter"}, {"credit_id": "54491c
89c3a3680fb4001cf7", "department": "Sound", "gender": 0, "id": 900, "job": "Sound Designer", "name":
"Christopher Boyes"}, {"credit_id": "54491cb70e0a267480001bd0", "department": "Sound", "gender": 0, "
id": 900, "job": "Supervising Sound Editor", "name": "Christopher Boyes"}, {"credit_id": "539c4a4cc3a
36810c9002101", "department": "Production", "gender": 1, "id": 1262, "job": "Casting", "name": "Mali
Finn"}, {"credit_id": "5544ee3b925141499f0008fc", "department": "Sound", "gender": 2, "id": 1729, "jo
b": "Original Music Composer", "name": "James Horner"}, {"credit_id": "52fe48009251416c750ac9c3", "de
partment": "Directing", "gender": 2, "id": 2710, "job": "Director", "name": "James Cameron"}, {"credi
t_id": "52fe48009251416c750ac9d9", "department": "Writing", "gender": 2, "id": 2710, "job": "Writer",
"name": "James Cameron"}, {"credit_id": "52fe48009251416c750aca17", "department": "Editing", "gende
r": 2, "id": 2710, "job": "Editor", "name": "James Cameron"}, {"credit_id": "52fe48009251416c750aca2
9", "department": "Production", "gender": 2, "id": 2710, "job": "Producer", "name": "James Cameron"},
{"credit_id": "52fe48009251416c750aca3f", "department": "Writing", "gender": 2, "id": 2710, "job": "S
creenplay", "name": "James Cameron"}, {"credit_id": "539c4987c3a36810ba0021a4", "department": "Art",
"gender": 2, "id": 7236, "job": "Art Direction", "name": "Andrew Menzies"}, {"credit_id": "549598c3c3
a3686ae9004383", "department": "Visual Effects", "gender": 0, "id": 6690, "job": "Visual Effects Prod
ucer", "name": "Jill Brooks"}, {"credit_id": "52fe48009251416c750aca4b", "department": "Production",
```


In [228]:
```
# Only need the director name
def get_director(obj):
    L = []
    for i in ast.literal_eval(obj):
        if i['job'] == 'Director':
            L.append(i['name'])
            break
    return L
```

In [229]:
```
# Updating Crew
movies['crew'] = movies['crew'].apply(get_director)
```

```
/var/folders/yb/v5f667y93j393zcsz2t_26qw0000gn/T/ipykernel_13438/4113385147.py:2: SettingWithCopyWarnin
g:
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 (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.h
tml#returning-a-view-versus-a-copy)
  movies['crew'] = movies['crew'].apply(get_director)
```

In [230]:
```
# Checking for the crew
movies.head()
```

Out[230]:

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

In [231]:
```
# Converting text to the list
movies['overview'] = movies['overview'].apply(lambda x:x.split())
```

```
/var/folders/yb/v5f667y93j393zcsz2t_26qw0000gn/T/ipykernel_13438/1971172765.py:2: SettingWithCopyWarnin
g:
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 (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.h
tml#returning-a-view-versus-a-copy)
  movies['overview'] = movies['overview'].apply(lambda x:x.split())
```

In [232]:
```
movies.head()
```

Out[232]:

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

In [233]:
```
# Taking care of the blank spaces in the columns
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'] = movies['cast'].apply(lambda x:[i.replace(" ","")for i in x])
movies['crew'] = movies['crew'].apply(lambda x:[i.replace(" ","")for i in x])
```

```
/var/folders/yb/v5f667y93j393zcsz2t_26qw0000gn/T/ipykernel_13438/2267989278.py:2: SettingWithCopyWarnin
g:
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 (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.h
tml#returning-a-view-versus-a-copy)
  movies['genres'] = movies['genres'].apply(lambda x:[i.replace(" ","")for i in x])
/var/folders/yb/v5f667y93j393zcsz2t_26qw0000gn/T/ipykernel_13438/2267989278.py:3: SettingWithCopyWarnin
g:
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 (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.h
tml#returning-a-view-versus-a-copy)
  movies['keywords'] = movies['keywords'].apply(lambda x:[i.replace(" ","")for i in x])
/var/folders/yb/v5f667y93j393zcsz2t_26qw0000gn/T/ipykernel_13438/2267989278.py:4: SettingWithCopyWarnin
g:
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 (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.h
tml#returning-a-view-versus-a-copy)
  movies['cast'] = movies['cast'].apply(lambda x:[i.replace(" ","")for i in x])
/var/folders/yb/v5f667y93j393zcsz2t_26qw0000gn/T/ipykernel_13438/2267989278.py:5: SettingWithCopyWarnin
g:
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 (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.h
tml#returning-a-view-versus-a-copy)
  movies['crew'] = movies['crew'].apply(lambda x:[i.replace(" ","")for i in x])
```

In [234]:
```
movies.head()
```

Out[234]:

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

In [235]:
```
# Combining all the columns into the one columns
movies['tags'] = movies['overview'] + movies['genres'] + movies['keywords'] + movies['cast'] + movies[
```

```
/var/folders/yb/v5f667y93j393zcsz2t_26qw0000gn/T/ipykernel_13438/2167919987.py:2: SettingWithCopyWarnin
g:
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 (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.h
tml#returning-a-view-versus-a-copy)
  movies['tags'] = movies['overview'] + movies['genres'] + movies['keywords'] + movies['cast'] + movie
s['crew']
```

In [236]:
```
movies.head()
```

Out[236]:

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

spacetravel, p...	mili...	SamanthaMorton, Wi...	former, mili...

In [237]:
```
# Creating final Dataset
final_df = movies[['id','title','tags']]
```

In [238]:
```
final_df.head()
```

Out[238]:

	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...

In [239]:
```
final_df['tags'].apply(lambda x:" ".join(x))
```

Out[239]:
```
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 [240]:
```
final_df['tags'] = final_df['tags'].apply(lambda x:" ".join(x))
```

```
/var/folders/yb/v5f667y93j393zcsz2t_26qw0000gn/T/ipykernel_13438/2951742378.py:1: SettingWithCopyWarnin
g:
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 (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.h
tml#returning-a-view-versus-a-copy)
  final_df['tags'] = final_df['tags'].apply(lambda x:" ".join(x))
```

In [241]:
```
final_df.head()
```

Out[241]:

	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 [242]:
```
# Checking for the 0th index
final_df['tags'][0]
```

Out[242]:
```
'In the 22nd century, a paraplegic Marine is dispatched to the moon Pandora on a unique mission, but be
comes torn between following orders and protecting an alien civilization. Action Adventure Fantasy Scie
nceFiction cultureclash future spacewar spacecolony society spacetravel futuristic romance space alien
tribe alienplanet cgi marine soldier battle loveaffair antiwar powerrelations mindandsoul 3d SamWorthin
gton ZoeSaldana SigourneyWeaver StephenLang JamesCameron'
```

Converting into Lower Case

In [243]:
```
final_df['tags'] = final_df['tags'].apply(lambda x:x.lower())
```

```
/var/folders/yb/v5f667y93j393zcsz2t_26qw0000gn/T/ipykernel_13438/2213019244.py:1: SettingWithCopyWarnin
g:
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 (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.h
tml#returning-a-view-versus-a-copy)
  final_df['tags'] = final_df['tags'].apply(lambda x:x.lower())
```

In [244]:
```
final_df.head()
```

Out[244]:

	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 [245]:
```
!pip install nltk
```

```
Requirement already satisfied: nltk in /Users/varundesai/anaconda3/lib/python3.10/site-packages (3.7)
Requirement already satisfied: click in /Users/varundesai/anaconda3/lib/python3.10/site-packages (from
nltk) (8.0.4)
Requirement already satisfied: joblib in /Users/varundesai/anaconda3/lib/python3.10/site-packages (from
nltk) (1.1.1)
Requirement already satisfied: regex>=2021.8.3 in /Users/varundesai/anaconda3/lib/python3.10/site-packa
ges (from nltk) (2022.7.9)
Requirement already satisfied: tqdm in /Users/varundesai/anaconda3/lib/python3.10/site-packages (from n
ltk) (4.64.1)
```

In [246]:
```
# Importing Libraries
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import PorterStemmer
import string
```

In [247]:
```
# Removing Stopwords, punctuation marks, applying stemming operation
def pre_process(text):
    tokens = word_tokenize(text)
    tokens = [word for word in tokens if word.isalnum()]

    stop_words = set(stopwords.words('english'))
    tokens = [word for word in tokens if word.lower() not in stop_words]

    stemmer = PorterStemmer()
    tokens = [stemmer.stem(word) for word in tokens]

    processed_text = ' '.join(tokens)


    return processed_text
```

In [248]:
```
final_df['tags'] = final_df['tags'].apply(pre_process)
```

```
/var/folders/yb/v5f667y93j393zcsz2t_26qw0000gn/T/ipykernel_13438/3906571443.py:1: SettingWithCopyWarnin
g:
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 (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.h
tml#returning-a-view-versus-a-copy)
  final_df['tags'] = final_df['tags'].apply(pre_process)
```

In [249]:
```
final_df.head()
```

Out[249]:

	id	title	tags
0	19995	Avatar	22nd centuri parapleg marin dispatch moon pand...
1	285	Pirates of the Caribbean: At World's End	captain barbossa long believ dead come back li...
2	206647	Spectre	cryptic messag bond past send trail uncov sini...
3	49026	The Dark Knight Rises	follow death district attorney harvey dent bat...
4	49529	John Carter	john carter former militari captain inexplic t...

In [250]:
```
final_df['tags'][0]
```

Out[250]:
```
'22nd centuri parapleg marin dispatch moon pandora uniqu mission becom torn follow order protect alien
civil action adventur fantasi sciencefict cultureclash futur spacewar spacecoloni societi spacetravel f
uturist romanc space alien tribe alienplanet cgi marin soldier battl loveaffair antiwar powerrel mindan
dsoul 3d samworthington zoesaldana sigourneyweav stephenlang jamescameron'
```

Training

- Different techniques such as Bag of Words, IF-IDF, Word2Vec could be used to perform word embedding.

In [251]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
# Max_features are 5000
tfidf_vectorizer = TfidfVectorizer(max_features=5000)

# Creating Matrix for the vector values
tfidf_matrix = tfidf_vectorizer.fit_transform(final_df['tags'])

# Converting into the numpy array
tfidf_df = pd.DataFrame(tfidf_matrix.toarray(), columns=tfidf_vectorizer.get_feature_names_out())
```

In [252]:
```
tfidf_df.head()
```

Out[252]:

	007	10	100	11	12	13	15	16	17	17th	...	zhangyim	zhangziyi	zion	zoe	zoesaldana	zombi	zombieapocalyps	zone	zoo	zoo
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	0.207132	0.0	0.0	0.0	0.0	
1	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	0.000000	0.0	0.0	0.0	0.0	
2	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	0.000000	0.0	0.0	0.0	0.0	
3	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	0.000000	0.0	0.0	0.0	0.0	
4	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	0.000000	0.0	0.0	0.0	0.0	

5 rows × 5000 columns

- Sinces 5000 features have been selected now, we can calculate the Cosine Similarity, in order to find the similar vectors values

In [253]:
```
from sklearn.metrics.pairwise import cosine_similarity
cosine_similarity(tfidf_df).shape
```

Out[253]:
```
(4806, 4806)
```

In [254]:
```
similarity = cosine_similarity(tfidf_df)
```

In [255]:
```
similarity[0]
```

Out[255]:
```
array([1.        , 0.02218662, 0.02952472, ..., 0.04533449, 0.00551997,
       0.        ])
```

Insight:

- 0th index value is same as itself therefore values is 1
- 0th index is similar to the 1st index item with value of 0.02218..
- Higher the number, more similar it is to the item

In [256]:
```
# Creating Recommendation Function
def recommend(movie):
    movie_index = final_df[final_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(i[0])
```

In [257]:
```
recommend('Batman')
```

```
1363
210
```

```
428
3
1364
```

In [258]:
```
final_df.iloc[428].title
```

Out[258]:
```
'Batman Returns'
```

Insight:

- 428 indexed value has the title "Batman Returns" which is similar to the input "Batman"

In [259]:
```
# Storing the model in .pkl format
import pickle
pickle.dump(final_df,open('movies.pkl','wb'))
```

In [ ]:

In [ ]:

Type *Markdown* and LaTeX: α^2

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: