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

In [3]:
```
movies=pd.read_csv("top10K-TMDB-movies.csv")
```

In [6]:
```
movies.head()
```

Out[6]:

	id	title	genre	original_language	overview	popularity	release
0	278	The Shawshank Redemption	Drama,Crime	en	Framed in the 1940s for the double murder of h...	94.075	1994
1	19404	Dilwale Dulhania Le Jayenge	Comedy,Drama,Romance	hi	Raj is a rich, carefree, happy-go-lucky second...	25.408	1995
2	238	The Godfather	Drama,Crime	en	Spanning the years 1945 to 1955, a chronicle o...	90.585	1972
3	424	Schindler's List	Drama,History,War	en	The true story of how businessman Oskar Schind...	44.761	1993
4	240	The Godfather: Part II	Drama,Crime	en	In the continuing saga of the Corleone crime f...	57.749	1974

In [8]:
```
movies.describe()
```

Out[8]:

	id	popularity	vote_average	vote_count
count	10000.000000	10000.000000	10000.000000	10000.000000
mean	161243.505000	34.697267	6.621150	1547.309400
std	211422.046043	211.684175	0.766231	2648.295789
min	5.000000	0.600000	4.600000	200.000000
25%	10127.750000	9.154750	6.100000	315.000000
50%	30002.500000	13.637500	6.600000	583.500000
75%	310133.500000	25.651250	7.200000	1460.000000
max	934761.000000	10436.917000	8.700000	31917.000000

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 9 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   id                 10000 non-null  int64  
 1   title              10000 non-null  object 
 2   genre              9997 non-null   object 
 3   original_language  10000 non-null  object 
 4   overview           9987 non-null   object 
 5   popularity         10000 non-null  float64
 6   release_date       10000 non-null  object 
 7   vote_average       10000 non-null  float64
 8   vote_count         10000 non-null  int64  
dtypes: float64(2), int64(2), object(5)
memory usage: 703.2+ KB
```

In [10]:
```
movies.isnull().sum()
```

Out[10]:
```
id                    0
title                 0
genre                 3
original_language     0
overview             13
popularity            0
release_date          0
vote_average          0
vote_count            0
dtype: int64
```

Feature Selection

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

Out[11]:
```
Index(['id', 'title', 'genre', 'original_language', 'overview', 'popularity',
       'release_date', 'vote_average', 'vote_count'],
      dtype='object')
```

In [12]:
```
movies=movies[['id','title','overview','genre']]
```

In [13]:

```
movies
```

Out[13]:

	id	title	overview	genre
0	278	The Shawshank Redemption	Framed in the 1940s for the double murder of h...	Drama,Crime
1	19404	Dilwale Dulhania Le Jayenge	Raj is a rich, carefree, happy-go-lucky second...	Comedy,Drama,Romance
2	238	The Godfather	Spanning the years 1945 to 1955, a chronicle o...	Drama,Crime
3	424	Schindler's List	The true story of how businessman Oskar Schind...	Drama,History,War
4	240	The Godfather: Part II	In the continuing saga of the Corleone crime f...	Drama,Crime
...	...	...	...	...
9995	10196	The Last Airbender	The story follows the adventures of Aang, a yo...	Action,Adventure,Fantasy
9996	331446	Sharknado 3: Oh Hell No!	The sharks take bite out of the East Coast whe...	Action,TV Movie,Science Fiction,Comedy,Adventure
9997	13995	Captain America	During World War II, a brave, patriotic Americ...	Action,Science Fiction,War
9998	2312	In the Name of the King: A Dungeon Siege Tale	A man named Farmer sets out to rescue his kidn...	Adventure,Fantasy,Action,Drama
9999	455957	Domino	Seeking justice for his partner’s murder by an...	Thriller,Action,Crime

10000 rows × 4 columns

content based recommendation system

In [14]:

```
movies['tags']=movies['overview']+movies['genre']
```

```
C:\Users\jishi\AppData\Local\Temp\ipykernel_17268\226370073.py:1: SettingWith
CopyWarning:
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/s
table/user_guide/indexing.html#returning-a-view-versus-a-copy (https://panda
s.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-ver
sus-a-copy)
  movies['tags']=movies['overview']+movies['genre']
```

In [15]:
```
movies
```

Out[15]:

	id	title	overview	genre	tags
0	278	The Shawshank Redemption	Framed in the 1940s for the double murder of h...	Drama,Crime	Framed in the 1940s for the double murder of h...
1	19404	Dilwale Dulhania Le Jayenge	Raj is a rich, carefree, happy-go-lucky second...	Comedy,Drama,Romance	Raj is a rich, carefree, happy-go-lucky second...
2	238	The Godfather	Spanning the years 1945 to 1955, a chronicle o...	Drama,Crime	Spanning the years 1945 to 1955, a chronicle o...
3	424	Schindler's List	The true story of how businessman Oskar Schind...	Drama,History,War	The true story of how businessman Oskar Schind...
4	240	The Godfather: Part II	In the continuing saga of the Corleone crime f...	Drama,Crime	In the continuing saga of the Corleone crime f...
...	...	...	...	...	...
9995	10196	The Last Airbender	The story follows the adventures of Aang, a yo...	Action,Adventure,Fantasy	The story follows the adventures of Aang, a yo...
9996	331446	Sharknado 3: Oh Hell No!	The sharks take bite out of the East Coast whe...	Action,TV Movie,Science Fiction,Comedy,Adventure	The sharks take bite out of the East Coast whe...
9997	13995	Captain America	During World War II, a brave, patriotic Americ...	Action,Science Fiction,War	During World War II, a brave, patriotic Americ...
9998	2312	In the Name of the King: A Dungeon Siege Tale	A man named Farmer sets out to rescue his kidn...	Adventure,Fantasy,Action,Drama	A man named Farmer sets out to rescue his kidn...
9999	455957	Domino	Seeking justice for his partner’s murder by an...	Thriller,Action,Crime	Seeking justice for his partner’s murder by an...

10000 rows × 5 columns

In [18]:
```
new_data=movies.drop(columns=['overview','genre'])
```

In [19]:
```
new_data
```

Out[19]:

	id	title	tags
0	278	The Shawshank Redemption	Framed in the 1940s for the double murder of h...
1	19404	Dilwale Dulhania Le Jayenge	Raj is a rich, carefree, happy-go-lucky second...
2	238	The Godfather	Spanning the years 1945 to 1955, a chronicle o...
3	424	Schindler's List	The true story of how businessman Oskar Schind...
4	240	The Godfather: Part II	In the continuing saga of the Corleone crime f...
...	...	...	...
9995	10196	The Last Airbender	The story follows the adventures of Aang, a yo...
9996	331446	Sharknado 3: Oh Hell No!	The sharks take bite out of the East Coast whe...
9997	13995	Captain America	During World War II, a brave, patriotic Americ...
9998	2312	In the Name of the King: A Dungeon Siege Tale	A man named Farmer sets out to rescue his kidn...
9999	455957	Domino	Seeking justice for his partner’s murder by an...

10000 rows × 3 columns

In [20]:
```
#1.Bag of word
#2.TFIDF
```

In [23]:
```
from sklearn.feature_extraction.text import CountVectorizer
```

In [24]:
```
cv=CountVectorizer(max_features=10000,stop_words='english')
```

In [25]:
```
cv
```

Out[25]:
```
CountVectorizer(max_features=10000, stop_words='english')
```

In [27]:
```
vector=cv.fit_transform(new_data['tags'].values.astype('U')).toarray()
```

In [28]:
```
vector.shape
```

Out[28]:
```
(10000, 10000)
```

In [29]:
```
from sklearn.metrics.pairwise import cosine_similarity
```

In [30]:
```
similarity=cosine_similarity(vector)
```

In [31]:
```
similarity
```

Out[31]:
```
array([[1.        , 0.05634362, 0.12888482, ..., 0.07559289, 0.11065667,
         0.06388766],
       [0.05634362, 1.        , 0.07624929, ..., 0.        , 0.03636965,
         0.        ],
       [0.12888482, 0.07624929, 1.        , ..., 0.02273314, 0.06655583,
         0.08645856],
       ...,
       [0.07559289, 0.        , 0.02273314, ..., 1.        , 0.03253   ,
         0.02817181],
       [0.11065667, 0.03636965, 0.06655583, ..., 0.03253   , 1.        ,
         0.0412393 ],
       [0.06388766, 0.        , 0.08645856, ..., 0.02817181, 0.0412393 ,
         1.        ]])
```

In [35]:
```
new_data[new_data['title']=="The Godfather"].index[0]
```

Out[35]:
```
2
```

In [36]:
```
new_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   id      10000 non-null  int64 
 1   title   10000 non-null  object
 2   tags    9985 non-null   object
dtypes: int64(1), object(2)
memory usage: 234.5+ KB
```

In [46]:
```
distance = sorted(list(enumerate(similarity[2])),reverse=True,key=lambda vecto
for i in distance[0:5]:
    print(new_data.iloc[i[0]].title)
```

```
The Godfather
The Godfather: Part II
Blood Ties
Joker
Bomb City
```

In [48]:
```
def recommand(movies):
    index=new_data[new_data['title']==movies].index[0]
    distance = sorted(list(enumerate(similarity[index])),reverse=True,key=lamb
    for i in distance[0:5]:
        print(new_data.iloc[i[0]].title)
```

In [49]:
```
recommand("Iron Man")
```

```
Iron Man
Iron Man 3
Guardians of the Galaxy Vol. 2
Avengers: Age of Ultron
Star Wars: Episode III - Revenge of the Sith
```

In [50]:
```
import pickle
```

In [55]:
```
pickle.dump(new_data,open('movies_list.pkl','wb'))
```

In [57]:
```
pickle.dump(similarity,open('similarity.pkl','wb'))
```

In [56]:
```
pickle.load(open('movies_list.pkl','rb'))
```

Out[56]:

	id	title	tags
0	278	The Shawshank Redemption	Framed in the 1940s for the double murder of h...
1	19404	Dilwale Dulhania Le Jayenge	Raj is a rich, carefree, happy-go-lucky second...
2	238	The Godfather	Spanning the years 1945 to 1955, a chronicle o...
3	424	Schindler's List	The true story of how businessman Oskar Schind...
4	240	The Godfather: Part II	In the continuing saga of the Corleone crime f...
...	...	...	...
9995	10196	The Last Airbender	The story follows the adventures of Aang, a yo...
9996	331446	Sharknado 3: Oh Hell No!	The sharks take bite out of the East Coast whe...
9997	13995	Captain America	During World War II, a brave, patriotic Americ...
9998	2312	In the Name of the King: A Dungeon Siege Tale	A man named Farmer sets out to rescue his kidn...
9999	455957	Domino	Seeking justice for his partner’s murder by an...

10000 rows × 3 columns

In [ ]:

In [ ]: