



```
[1]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
[2]: import pandas as pd
movies = pd.read_csv("top10K-TMDB-movies.csv")
movies.head(5)
```

[2]:	id	title	genre	original_language	overview	popularity	release_date	vote_average	vote_count
0	278	The Shawshank Redemption	Drama,Crime	en	Framed in the 1940s for the double murder of h...	94.075	1994-09-23	8.7	21862
1	19404	Dilwale Dulhania Le Jayenge	Comedy,Drama,Romance	hi	Raj is a rich, carefree, happy-go-lucky second...	25.408	1995-10-19	8.7	3731
2	238	The Godfather	Drama,Crime	en	Spanning the years 1945 to 1955, a chronicle o...	90.585	1972-03-14	8.7	16280
3	424	Schindler's List	Drama,History,War	en	The true story of how businessman Oskar Schind...	44.761	1993-12-15	8.6	12959
4	240	The Godfather: Part II	Drama,Crime	en	In the continuing saga of the Corleone crime f...	57.749	1974-12-20	8.6	9811

```
[3]: movies.describe()
```

```
[3]:
```

	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

```
[4]: 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
```

```
[5]: movies.isnull().sum()
```

```
[5]: 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 part

```
[6]: movies.columns
```

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

```
[7]: movies = movies[["id","title","overview","genre"]]
```

```
[7]: movies = movies[["id","title","overview","genre"]]
```

```
[8]: movies
```

[8]:	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

```
[9]: movies["tags"] = movies["overview"]+movies["genre"]
```

```
[10]: movies
```

[10]:	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...
...

```
[11]: new_data = movies.drop(columns=["overview","genre"])    #iam drop the overview & genre columns
```

```
[12]: new_data.head(20)
```

```
[12]:
```

	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...
5	667257	Impossible Things	Matilde is a woman who, after the death of her...
6	129	Spirited Away	A young girl, Chihiro, becomes trapped in a st...
7	730154	Your Eyes Tell	A tragic accident lead to Kaori's blindness, b...
8	372754	Dou kyu sei – Classmates	Rihito Sajo, an honor student with a perfect s...

```
[13]: from sklearn.feature_extraction.text import CountVectorizer          # convert the tags data into bag to word
```

```
[14]: cv = CountVectorizer(max_features=10000,stop_words = "english")
```

```
[15]: cv
```

```
[15]: ▼ CountVectorizer ⓘ ?
```

```
▶ Parameters
```



```
[16]: vector = cv.fit_transform(new_data["tags"].values.astype("U")).toarray()
```

```
[17]: vector.shape
```

```
[17]: (10000, 10000)
```

```
[18]: from sklearn.metrics.pairwise import cosine_similarity
```

```
[19]: similarity = cosine_similarity(vector)
```

```
[20]: similarity
```

```
[20]: array([[1.          , 0.05634362, 0.13041013, ..., 0.07559289, 0.11065667,
          0.06900656],
          [0.05634362, 1.          , 0.07715167, ..., 0.          , 0.03636965,
          0.          ],
          [0.13041013, 0.07715167, 1.          , ..., 0.02300219, 0.0673435 ,
          0.09449112],
          ...,
          [0.07559289, 0.          , 0.02300219, ..., 1.          , 0.03253   ,
          0.03042903],
          [0.11065667, 0.03636965, 0.0673435 , ..., 0.03253   , 1.          ,
          0.04454354],
          [0.06900656, 0.          , 0.09449112, ..., 0.03042903, 0.04454354,
          1.          ]], shape=(10000, 10000))
```

```
[21]: new_data[new_data["title"]=="The Godfather"].index[0]
```

```
[21]: np.int64(2)
```

[22]: `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
```

[23]: `distance = sorted(list(enumerate(similarity[2])), reverse=True, key=lambda vector: vector[1])`
`for i in distance[0:5]:` *#top 5 value*
 `print(new_data.iloc[i[0]].title)`

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

[2]: `def recommand(movies):`
 `index = new_data[new_data["title"] == movies].index[0]`
 `distance = sorted(list(enumerate(similarity[index])), reverse=True, key=lambda vector: vector[1])`
 `for i in distance[0:5]:` *# top 5 recommendations*
 `print(new_data.iloc[i[0]].title)`


```
[25]: recommend("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

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:
```