```
In [1]: import numpy as np
         import pandas as pd
In [2]: movies = pd.read_csv('tmdb_5000_movies.csv')
         credits = pd.read csv('tmdb 5000 credits.csv')
        movies = movies.merge(credits, on='title')
In [4]: movies.head(1)
Out[4]:
                                                            id keywords original_language original_title overview
                                                                                                                popularity production_companies
                budget
                         genres
                                              homepage
                                                                   [{"id":
                                                                                                         In the
                       [{"id": 28,
                                                                                                          22nd
                                                                   1463,
                         "name":
                                                                                                                              [{"name": "Ingenious
                                                                 "name":
                                                                                                      century, a
          0 237000000 "Action"},
                                                                                                               150.437577
                                http://www.avatarmovie.com/ 19995
                                                                                               Avatar
                                                                                                                              Film Partners", "id"
                                                                 "culture
                                                                                                      paraplegic
                                                                                                                                         289...
                        {"id": 12,
                                                                 clash"},
                                                                                                      Marine is
                         "nam...
                                                                  {"id":...
                                                                                                           di...
         1 rows × 23 columns
In [5]: # removing not needed columns from the data set
         # say, for a content based movie recommender system 'budget' attribute won't have any affect
         # removing only such columns for more ex: like; homepage, original language, original title (title is already there)..
In [6]: | movies = movies[['movie id', 'title', 'overview', 'genres', 'keywords', 'cast', 'crew']]
```

In [7]: movies.head(5)

Out[7]:

Out[10]: 0

04.0[7].		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 [8]:	mov	ies.isnu	ll().sum()					
Out[8]:	tit ove gen key cas cre	rview res words t	0 0 3 0 0 0 0					
In [9]:	mov	ies.drop	na(inplace = T	rue)				
In [10]:	mov:	ies.dupl	icated().sum()					

```
In [11]: # preprocessing genres
         movies.iloc[0].genres
Out[11]: '[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 878, "name":
         "Science Fiction"}]'
In [12]: # found stringified list of dictionaries
         # firstly have to convert the string into a pure list using 'ast' module
         import ast
         # now will use ast.literal eval(string) where ever needed.
In [13]: def convert(string):
             L = []
             for x in ast.literal eval(string):
                 L.append(x['name'])
             return L
In [14]: movies['genres'] = movies['genres'].apply(convert)
In [15]: # same can be applied on the keywords column
         movies['keywords'] = movies['keywords'].apply(convert)
```

```
In [16]:

def convertcast(string):
    L = []
    counter = 0
    for x in ast.literal_eval(string):
        if counter != 3:
              L.append(x['name'])
              counter += 1
        else:
              break

return L
```

```
In [17]: movies['cast'] = movies['cast'].apply(convertcast)
```

In [18]: movies.head(5)

Out[18]:

crew	cast	keywords	genres	overview	title	movie_id	
[{"credit_id": "52fe48009251416c750aca23", "de	[Sam Worthington, Zoe Saldana, Sigourney Weaver]	[culture clash, future, space war, space colon	[Action, Adventure, Fantasy, Science Fiction]	In the 22nd century, a paraplegic Marine is di	Avatar	19995	0
[{"credit_id": "52fe4232c3a36847f800b579", "de	[Johnny Depp, Orlando Bloom, Keira Knightley]	[ocean, drug abuse, exotic island, east india	[Adventure, Fantasy, Action]	Captain Barbossa, long believed to be dead, ha	Pirates of the Caribbean: At World's End	285	1
[{"credit_id": "54805967c3a36829b5002c41", "de	[Daniel Craig, Christoph Waltz, Léa Seydoux]	[spy, based on novel, secret agent, sequel, mi	[Action, Adventure, Crime]	A cryptic message from Bond's past sends him o	Spectre	206647	2
[{"credit_id": "52fe4781c3a36847f81398c3", "de	[Christian Bale, Michael Caine, Gary Oldman]	[dc comics, crime fighter, terrorist, secret i	[Action, Crime, Drama, Thriller]	Following the death of District Attorney Harve	The Dark Knight Rises	49026	3
[{"credit_id": "52fe479ac3a36847f813eaa3", "de	[Taylor Kitsch, Lynn Collins, Samantha Morton]	[based on novel, mars, medallion, space travel	[Action, Adventure, Science Fiction]	John Carter is a war-weary, former military ca	John Carter	49529	4

```
In [19]: def fetchdirector(string):
                 L = []
                 for x in ast.literal eval(string):
                      if x['job'] == 'Director':
                            L.append(x['name'])
                            break
                 return L
In [20]:
           movies['crew'] = movies['crew'].apply(fetchdirector)
In [21]:
            movies.head(5)
Out[21]:
                movie id
                                           title
                                                                overview
                                                                                                                keywords
                                                                                          genres
                                                                                                                                                cast
                                                                                                                                                               crew
                                                                                                                               [Sam Worthington, Zoe
                                                                               [Action, Adventure,
                                                                                                      [culture clash, future,
                                                     In the 22nd century, a
                                                                                                                                                              [James
             0
                    19995
                                                                                 Fantasy, Science
                                                                                                         space war, space
                                                                                                                                   Saldana, Sigourney
                                         Avatar
                                                   paraplegic Marine is di...
                                                                                                                                                           Cameron]
                                                                                          Fiction]
                                                                                                                   colon...
                                                                                                                                             Weaver1
                                   Pirates of the
                                                                                                       [ocean, drug abuse,
                                                   Captain Barbossa, Iong
                                                                              [Adventure, Fantasy,
                                                                                                                               [Johnny Depp, Orlando
                                                                                                                                                               [Gore
             1
                      285
                           Caribbean: At World's
                                                                                                    exotic island, east india
                                                  believed to be dead, ha...
                                                                                          Action]
                                                                                                                               Bloom, Keira Knightley]
                                                                                                                                                           Verbinski]
                                            End
                                                   A cryptic message from
                                                                                                      [spy, based on novel,
                                                                               [Action, Adventure,
                                                                                                                               [Daniel Craig, Christoph
                                                                                                                                                       [Sam Mendes]
             2
                   206647
                                                    Bond's past sends him
                                                                                                      secret agent, sequel,
                                        Spectre
                                                                                                                                  Waltz, Léa Seydoux]
                                                                                          Crime]
                                                                                                                     mi...
                                                                      0...
                                 The Dark Knight
                                                                            [Action, Crime, Drama,
                                                                                                   [dc comics, crime fighter,
                                                                                                                               [Christian Bale, Michael
                                                                                                                                                         [Christopher
                                                     Following the death of
             3
                    49026
                                          Rises
                                                   District Attorney Harve...
                                                                                          Thriller]
                                                                                                         terrorist, secret i...
                                                                                                                                 Caine, Gary Oldman]
                                                                                                                                                              Nolan1
                                                      John Carter is a war-
                                                                                                     [based on novel, mars,
                                                                               [Action, Adventure,
                                                                                                                            [Taylor Kitsch, Lynn Collins,
                                                                                                                                                             [Andrew
                    49529
                                     John Carter
                                                     weary, former military
                                                                                                          medallion, space
                                                                                  Science Fiction1
                                                                                                                                    Samantha Mortonl
                                                                                                                                                             Stanton]
                                                                                                                  travel...
           # converting overview from 'string' to 'list'
In [22]:
            movies['overview'] = movies['overview'].apply(lambda x:x.split())
```

In [23]: movies.head(5)

Out[23]:

	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 [24]: # removing spaces from between words like 'Sam Worthington' & 'Sam Mendes',
         # if not removed may result in alternate responses
         movies['genres'] = movies['genres'].apply(lambda x:[i.replace(" ","") for i in x])
         movies['cast'] = movies['cast'].apply(lambda x:[i.replace(" ","") for i in x])
         movies['keywords'] = movies['keywords'].apply(lambda x:[i.replace(" ","") for i in x])
         movies['crew'] = movies['crew'].apply(lambda x:[i.replace(" ","") for i in x])
```

In [25]: movies.head()

Out[25]:

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

```
In [26]: # combining all columns into a single column named 'tags'
movies['tags'] = movies['overview'] + movies['genres'] + movies['keywords'] + movies['cast'] + movies['crew']
```

In [27]: movies.head()

Out[27]:

crew	cast	keywords	genres	overview	title	movie_id	
[JamesCameron]	[SamWorthington, ZoeSaldana, SigourneyWeaver]	[cultureclash, future, spacewar, spacecolony,	[Action, Adventure, Fantasy, ScienceFiction]	[In, the, 22nd, century,, a, paraplegic, Marin	Avatar	19995	0
[GoreVerbinski]	[JohnnyDepp, OrlandoBloom, KeiraKnightley]	[ocean, drugabuse, exoticisland, eastindiatrad	[Adventure, Fantasy, Action]	[Captain, Barbossa,, long, believed, to, be, d	Pirates of the Caribbean: At World's End	285	1
[SamMendes]	[DanielCraig, ChristophWaltz, LéaSeydoux]	[spy, basedonnovel, secretagent, sequel, mi6,	[Action, Adventure, Crime]	[A, cryptic, message, from, Bond's, past, send	Spectre	206647	2
[ChristopherNolan]	[ChristianBale, MichaelCaine, GaryOldman]	[dccomics, crimefighter, terrorist, secretiden	[Action, Crime, Drama, Thriller]	[Following, the, death, of, District, Attorney	The Dark Knight Rises	49026	3
[AndrewStanton]	[TaylorKitsch, LynnCollins, SamanthaMorton]	[basedonnovel, mars, medallion, spacetravel, p	[Action, Adventure, ScienceFiction]	[John, Carter, is, a, war-weary,, former, mili	John Carter	49529	4
]	[GoreVerbinski] [SamMendes] [ChristopherNolan]	ZoeSaldana, [JamesCameron] SigourneyWeaver] [JohnnyDepp, OrlandoBloom, KeiraKnightley] [DanielCraig, ChristophWaltz, LéaSeydoux] [ChristianBale, MichaelCaine, GaryOldman] [TaylorKitsch, LynnCollins, [AndrewStanton]	[cultureclash, future, spacewar, spacewar, spacecolony, SigourneyWeaver] [ocean, drugabuse, exoticisland, eastindiatrad KeiraKnightley] [spy, basedonnovel, secretagent, sequel, mi6, [DanielCraig, ChristophWaltz, LéaSeydoux] [dccomics, crimefighter, terrorist, secretiden GaryOldman] [basedonnovel, mars, medallion, LynnCollins, [AndrewStanton]	[Action, Adventure, Fantasy, Spacewar, ScienceFiction] Spacecolony, SigourneyWeaver] [Adventure, Fantasy, Action] Spacecolony, SigourneyWeaver] [Action, Adventure, Crime] Spy, basedonnovel, Secretagent, sequel, Drama, Thriller] Crimesighter, terrorist, Secretiden ScienceFiction] Spacecolony, SigourneyWeaver] [Action, Adventure, Crime, Drama, Thriller] ScienceFiction] Spacecolony, ScienceFiction] Spacecolony, Spacecolon	[In, the, 22nd, century,, a, paraplegic, Marin [Cultureclash, future, spacewar, spacewar, spacewar, spacewar, spacewar, sigourneyWeaver] [JohnnyDepp, OrlandoBloom, keiraKnightley] [GoreVerbinski keiraKnightley] [GoreVerbinski keiraKnightley] [GoreVerbinski keiraKnightley] [SamWorthington, ZoeSaldana, SigourneyWeaver] [JohnnyDepp, OrlandoBloom, keiraKnightley] [GoreVerbinski	Avatar Cartury, a paraplegic, Marin Fantasy, ScienceFiction SigourneyWeaver Fantasy, ScienceFiction Fantasy,	1995 Avatar

```
In [28]:
```

```
In [29]: # converting the (list)tags into (string)tags
         new_df['tags'] = new_df['tags'].apply(lambda x:" ".join(x))
```

C:\Users\Vishwas\AppData\Local\Temp\ipykernel_3504\160216261.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#returnin g-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user guide/indexing.html#returning-a-view-versu s-a-copy)

```
new df['tags'] = new df['tags'].apply(lambda x:" ".join(x))
```

In [30]: new_df.head()

Out[30]:

tags	title	movie_id	
In the 22nd century, a paraplegic Marine is di	Avatar	19995	0
Captain Barbossa, long believed to be dead, ha	Pirates of the Caribbean: At World's End	285	1
A cryptic message from Bond's past sends him o	Spectre	206647	2
Following the death of District Attorney Harve	The Dark Knight Rises	49026	3
John Carter is a war-weary, former military ca	John Carter	49529	4

```
In [31]: new_df['tags'][0]
```

Out[31]: 'In the 22nd century, a paraplegic Marine is dispatched to the moon Pandora on a unique mission, but becomes torn be tween following orders and protecting an alien civilization. Action Adventure Fantasy ScienceFiction cultureclash fu ture spacewar spacecolony society spacetravel futuristic romance space alien tribe alienplanet cgi marine soldier ba ttle loveaffair antiwar powerrelations mindandsoul 3d SamWorthington ZoeSaldana SigourneyWeaver JamesCameron'

```
In [32]: new_df['tags'] = new_df['tags'].apply(lambda x: x.lower())
```

C:\Users\Vishwas\AppData\Local\Temp\ipykernel_3504\1380776331.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/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)

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

```
new df.head()
In [33]:
```

Out[33]:

```
movie_id
                                                   title
                                                                                                      tags
0
       19995
                                                             in the 22nd century, a paraplegic marine is di...
                                                 Avatar
         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...
                                 The Dark Knight Rises
3
      49026
                                                              following the death of district attorney harve...
      49529
                                           John Carter
                                                             john carter is a war-weary, former military ca...
```

```
In [34]: import nltk
         from nltk.stem.porter import PorterStemmer
         ps = PorterStemmer()
```

```
In [35]: def stem(text):
             y = []
             for i in text.split():
                 y.append(ps.stem(i))
             return " ".join(y)
```

```
new_df['tags'] = new_df['tags'].apply(stem)
In [36]:
```

C:\Users\Vishwas\AppData\Local\Temp\ipykernel_3504\3213734980.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/indexing.html#returnin g-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versu s-a-copy)

```
new df['tags'] = new df['tags'].apply(stem)
```

```
In [37]: # performing vectorization: text --> vector
         # using 'bag of words';
         from sklearn.feature extraction.text import CountVectorizer
         cv = CountVectorizer(max features = 5000, stop words = 'english')
In [38]: vectors = cv.fit transform(new df['tags']).toarray()
In [39]: vectors
Out[39]: array([[0, 0, 0, ..., 0, 0, 0],
                 [0, 0, 0, \ldots, 0, 0, 0],
                 [0, 0, 0, \ldots, 0, 0, 0],
                [0, 0, 0, \ldots, 0, 0, 0],
                [0, 0, 0, \ldots, 0, 0, 0],
                [0, 0, 0, ..., 0, 0, 0]], dtype=int64)
In [40]: |vectors[0]
Out[40]: array([0, 0, 0, ..., 0, 0, 0], dtype=int64)
In [41]: cv.get feature names out()
Out[41]: array(['000', '007', '10', ..., 'zone', 'zoo', 'zooeydeschanel'],
               dtype=object)
In [42]: new_df['tags'][0]
Out[42]: 'in the 22nd century, a parapleg marin is dispatch to the moon pandora on a uniqu mission, but becom torn between fo
         llow order and protect an alien civilization. action adventur fantasi sciencefict cultureclash futur spacewar spacec
         oloni societi spacetravel futurist romanc space alien tribe alienplanet cgi marin soldier battl loveaffair antiwar p
         owerrel mindandsoul 3d samworthington zoesaldana sigourneyweav jamescameron'
```

```
In [43]: # checking for similarities
         from sklearn.metrics.pairwise import cosine similarity
In [44]: similarity = cosine_similarity(vectors)
In [45]: similarity[1]
                                      , 0.06063391, ..., 0.02378257, 0.
Out[45]: array([0.08346223, 1.
                0.02615329])
In [46]: # recommending Logic
         def recommend(movie):
             # finding movie index in data
             movie_index = new_df[new_df['title'] == movie].index[0]
             # distances with other movies
             distances = similarity[movie_index]
             # gitting top 5 similar movies
             movies list = sorted(list(enumerate(distances)), reverse = True, key=lambda x:x[1])[1:6]
             for i in movies_list:
                 print(new_df.iloc[i[0]].title)
In [47]: recommend('Batman Begins')
         The Dark Knight
         Batman
         Batman
         The Dark Knight Rises
         10th & Wolf
In [48]: import pickle
```

```
In [49]: pickle.dump(new_df, open('movies.pkl', 'wb'))
In [50]: pickle.dump(new_df.to_dict(), open('movie_dict.pkl', 'wb'))
In [51]: pickle.dump(similarity, open('similarity.pkl', 'wb'))
In []:
```