Preprocessing of Dataset:-

```
import numpy as np
import pandas as pd

movies = pd.read_csv('tmdb_5000_movies.csv')
credits = pd.read_csv('tmdb_5000_credits.csv')

movies.head()
```

```
      budget                                                     genres  \
0  237000000  [{"id": 28, "name": "Action"}, {"id": 12, "nam...
1  300000000  [{"id": 12, "name": "Adventure"}, {"id": 14, "...
2  245000000  [{"id": 28, "name": "Action"}, {"id": 12, "nam...
3  250000000  [{"id": 28, "name": "Action"}, {"id": 80, "nam...
4  260000000  [{"id": 28, "name": "Action"}, {"id": 12, "nam...

                                       homepage      id  \
0                   http://www.avatarmovie.com/   19995
1  http://disney.go.com/disneypictures/pirates/     285
2   http://www.sonypictures.com/movies/spectre/  206647
3            http://www.thedarkknightrises.com/   49026
4          http://movies.disney.com/john-carter   49529

                                            keywords original_language
\
0  [{"id": 1463, "name": "culture clash"}, {"id":...                 en

1  [{"id": 270, "name": "ocean"}, {"id": 726, "na...                 en

2  [{"id": 470, "name": "spy"}, {"id": 818, "name...                 en

3  [{"id": 849, "name": "dc comics"}, {"id": 853,...                 en

4  [{"id": 818, "name": "based on novel"}, {"id":...                 en


                             original_title  \
0                                    Avatar
1  Pirates of the Caribbean: At World's End
2                                   Spectre
3                     The Dark Knight Rises
4                                John Carter

                                            overview  popularity  \
0  In the 22nd century, a paraplegic Marine is di...  150.437577
1  Captain Barbossa, long believed to be dead, ha...  139.082615
2  A cryptic message from Bond’s past sends him o...  107.376788
3  Following the death of District Attorney Harve...  112.312950
```

```
4  John Carter is a war-weary, former military ca...   43.926995

                                production_companies  \
0  [{"name": "Ingenious Film Partners", "id": 289...
1  [{"name": "Walt Disney Pictures", "id": 2}, {"...
2  [{"name": "Columbia Pictures", "id": 5}, {"nam...
3  [{"name": "Legendary Pictures", "id": 923}, {"...
4        [{"name": "Walt Disney Pictures", "id": 2}]

                                production_countries release_date
revenue  \
0  [{"iso_3166_1": "US", "name": "United States o...   2009-12-10
2787965087
1  [{"iso_3166_1": "US", "name": "United States o...   2007-05-19
961000000
2  [{"iso_3166_1": "GB", "name": "United Kingdom"...   2015-10-26
880674609
3  [{"iso_3166_1": "US", "name": "United States o...   2012-07-16
1084939099
4  [{"iso_3166_1": "US", "name": "United States o...   2012-03-07
284139100

   runtime                                     spoken_languages
status  \
0    162.0  [{"iso_639_1": "en", "name": "English"}, {"iso...
Released
1    169.0           [{"iso_639_1": "en", "name": "English"}]
Released
2    148.0  [{"iso_639_1": "fr", "name": "Fran\u00e7ais"},...
Released
3    165.0           [{"iso_639_1": "en", "name": "English"}]
Released
4    132.0           [{"iso_639_1": "en", "name": "English"}]
Released

                                          tagline  \
0                     Enter the World of Pandora.
1  At the end of the world, the adventure begins.
2                           A Plan No One Escapes
3                                 The Legend Ends
4            Lost in our world, found in another.

                                     title  vote_average  vote_count

0                                   Avatar           7.2       11800

1  Pirates of the Caribbean: At World's End           6.9        4500

2                                  Spectre           6.3        4466
```

```
3                    The Dark Knight Rises           7.6        9106

4                               John Carter           6.1        2124


credits.head()

   movie_id                                         title  \
0     19995                                        Avatar
1       285  Pirates of the Caribbean: At World's End
2    206647                                       Spectre
3     49026                        The Dark Knight Rises
4     49529                                   John Carter

                                                cast  \
0  [{"cast_id": 242, "character": "Jake Sully", "...
1  [{"cast_id": 4, "character": "Captain Jack Spa...
2  [{"cast_id": 1, "character": "James Bond", "cr...
3  [{"cast_id": 2, "character": "Bruce Wayne / Ba...
4  [{"cast_id": 5, "character": "John Carter", "c...

                                                crew
0  [{"credit_id": "52fe48009251416c750aca23", "de...
1  [{"credit_id": "52fe4232c3a36847f800b579", "de...
2  [{"credit_id": "54805967c3a36829b5002c41", "de...
3  [{"credit_id": "52fe4781c3a36847f81398c3", "de...
4  [{"credit_id": "52fe479ac3a36847f813eaa3", "de...

movies = movies.merge(credits,on='title')

movies

         budget                                                     genres  \
0     237000000  [{"id": 28, "name": "Action"}, {"id": 12, "nam...
1     300000000  [{"id": 12, "name": "Adventure"}, {"id": 14, "...
2     245000000  [{"id": 28, "name": "Action"}, {"id": 12, "nam...
3     250000000  [{"id": 28, "name": "Action"}, {"id": 80, "nam...
4     260000000  [{"id": 28, "name": "Action"}, {"id": 12, "nam...
...         ...                                                ...
4804     220000  [{"id": 28, "name": "Action"}, {"id": 80, "nam...
4805       9000  [{"id": 35, "name": "Comedy"}, {"id": 10749, "...
4806          0  [{"id": 35, "name": "Comedy"}, {"id": 18, "nam...
4807          0                                                 []
4808          0               [{"id": 99, "name": "Documentary"}]

                                               homepage      id  \
0                         http://www.avatarmovie.com/   19995
1          http://disney.go.com/disneypictures/pirates/     285
2           http://www.sonypictures.com/movies/spectre/  206647
3                  http://www.thedarkknightrises.com/   49026
4                http://movies.disney.com/john-carter   49529
```

```
...                                                  ...     ...
4804                                                 NaN    9367
4805                                                 NaN   72766
4806  http://www.hallmarkchannel.com/signedsealeddel...  231617
4807                        http://shanghaicalling.com/  126186
4808                                                 NaN   25975

                                                keywords
original_language  \
0     [{"id": 1463, "name": "culture clash"}, {"id":...
en
1     [{"id": 270, "name": "ocean"}, {"id": 726, "na...
en
2     [{"id": 470, "name": "spy"}, {"id": 818, "name...
en
3     [{"id": 849, "name": "dc comics"}, {"id": 853,...
en
4     [{"id": 818, "name": "based on novel"}, {"id":...
en
...                                                 ...
...
4804  [{"id": 5616, "name": "united states\u2013mexi...
es
4805                                                 []
en
4806  [{"id": 248, "name": "date"}, {"id": 699, "nam...
en
4807                                                 []
en
4808  [{"id": 1523, "name": "obsession"}, {"id": 224...
en

                               original_title  \
0                                      Avatar
1     Pirates of the Caribbean: At World's End
2                                     Spectre
3                       The Dark Knight Rises
4                                  John Carter
...                                        ...
4804                               El Mariachi
4805                                 Newlyweds
4806                  Signed, Sealed, Delivered
4807                           Shanghai Calling
4808                         My Date with Drew

                                             overview  popularity  \
0     In the 22nd century, a paraplegic Marine is di...  150.437577
1     Captain Barbossa, long believed to be dead, ha...  139.082615
2     A cryptic message from Bond’s past sends him o...  107.376788
3     Following the death of District Attorney Harve...  112.312950
```

```
4     John Carter is a war-weary, former military ca...   43.926995
...                                                  ...         ...
4804  El Mariachi just wants to play his guitar and ...   14.269792
4805  A newlywed couple's honeymoon is upended by th...    0.642552
4806  "Signed, Sealed, Delivered" introduces a dedic...    1.444476
4807  When ambitious New York attorney Sam is sent t...    0.857008
4808  Ever since the second grade when he first saw ...    1.929883

                                   production_companies  ...
runtime  \
0     [{"name": "Ingenious Film Partners", "id": 289...  ...   162.0

1     [{"name": "Walt Disney Pictures", "id": 2}, {"...  ...   169.0

2     [{"name": "Columbia Pictures", "id": 5}, {"nam...  ...   148.0

3     [{"name": "Legendary Pictures", "id": 923}, {"...  ...   165.0

4           [{"name": "Walt Disney Pictures", "id": 2}]  ...   132.0

...                                                 ...  ...     ...

4804          [{"name": "Columbia Pictures", "id": 5}]  ...    81.0

4805                                                []  ...    85.0

4806  [{"name": "Front Street Pictures", "id": 3958}...  ...   120.0

4807                                                []  ...    98.0

4808  [{"name": "rusty bear entertainment", "id": 87...  ...    90.0


                                       spoken_languages    status  \
0     [{"iso_639_1": "en", "name": "English"}, {"iso...  Released
1              [{"iso_639_1": "en", "name": "English"}]  Released
2     [{"iso_639_1": "fr", "name": "Fran\u00e7ais"},...  Released
3              [{"iso_639_1": "en", "name": "English"}]  Released
4              [{"iso_639_1": "en", "name": "English"}]  Released
...                                                 ...       ...
4804      [{"iso_639_1": "es", "name": "Espa\u00f1ol"}]  Released
4805                                                []  Released
4806           [{"iso_639_1": "en", "name": "English"}]  Released
4807           [{"iso_639_1": "en", "name": "English"}]  Released
4808           [{"iso_639_1": "en", "name": "English"}]  Released

                                             tagline  \
0                          Enter the World of Pandora.
1        At the end of the world, the adventure begins.
2                                A Plan No One Escapes
3                                     The Legend Ends
```

```
4                  Lost in our world, found in another.
...                                                 ...
4804  He didn't come looking for trouble, but troubl...
4805  A newlywed couple's honeymoon is upended by th...
4806                                                NaN
4807                           A New Yorker in Shanghai
4808                                                NaN

                                       title vote_average vote_count
\
0                                     Avatar          7.2      11800

1     Pirates of the Caribbean: At World's End          6.9       4500

2                                    Spectre          6.3       4466

3                      The Dark Knight Rises          7.6       9106

4                                  John Carter          6.1       2124

...                                        ...          ...        ...

4804                               El Mariachi          6.6        238

4805                                 Newlyweds          5.9          5

4806                  Signed, Sealed, Delivered          7.0          6

4807                       Shanghai Calling          5.7          7

4808                      My Date with Drew          6.3         16


     movie_id                                               cast  \
0       19995  [{"cast_id": 242, "character": "Jake Sully", "...
1         285  [{"cast_id": 4, "character": "Captain Jack Spa...
2      206647  [{"cast_id": 1, "character": "James Bond", "cr...
3       49026  [{"cast_id": 2, "character": "Bruce Wayne / Ba...
4       49529  [{"cast_id": 5, "character": "John Carter", "c...
...       ...                                                ...
4804     9367  [{"cast_id": 1, "character": "El Mariachi", "c...
4805    72766  [{"cast_id": 1, "character": "Buzzy", "credit_...
4806   231617  [{"cast_id": 8, "character": "Oliver O\u2019To...
4807   126186  [{"cast_id": 3, "character": "Sam", "credit_id...
4808    25975  [{"cast_id": 3, "character": "Herself", "credi...

                                                    crew
0     [{"credit_id": "52fe48009251416c750aca23", "de...
1     [{"credit_id": "52fe4232c3a36847f800b579", "de...
2     [{"credit_id": "54805967c3a36829b5002c41", "de...
3     [{"credit_id": "52fe4781c3a36847f81398c3", "de...
```

```
4     [{"credit_id": "52fe479ac3a36847f813eaa3", "de...
...                                                 ...
4804  [{"credit_id": "52fe44eec3a36847f80b280b", "de...
4805  [{"credit_id": "52fe487dc3a368484e0fb013", "de...
4806  [{"credit_id": "52fe4df3c3a36847f8275ecf", "de...
4807  [{"credit_id": "52fe4ad9c3a368484e16a36b", "de...
4808  [{"credit_id": "58ce021b9251415a390165d9", "de...

[4809 rows x 23 columns]
```

Display all columns name:-

```
movies.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 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: 864.2+ KB
```

Separate the columns from total dataset:-

genres, id, keywords, title, overview, release_date, status, cast, crew:-

```
movies =
movies[['movie_id','title','overview','genres','keywords','release_dat
e','cast','crew','status']]

movies.head()
```

```
   movie_id                                     title  \
0     19995                                    Avatar
1       285  Pirates of the Caribbean: At World's End
2    206647                                   Spectre
3     49026                     The Dark Knight Rises
4     49529                               John Carter

                                            overview  \
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...

                                              genres  \
0  [{"id": 28, "name": "Action"}, {"id": 12, "nam...
1  [{"id": 12, "name": "Adventure"}, {"id": 14, "...
2  [{"id": 28, "name": "Action"}, {"id": 12, "nam...
3  [{"id": 28, "name": "Action"}, {"id": 80, "nam...
4  [{"id": 28, "name": "Action"}, {"id": 12, "nam...

                                            keywords release_date  \
0  [{"id": 1463, "name": "culture clash"}, {"id":...   2009-12-10
1  [{"id": 270, "name": "ocean"}, {"id": 726, "na...   2007-05-19
2  [{"id": 470, "name": "spy"}, {"id": 818, "name...   2015-10-26
3  [{"id": 849, "name": "dc comics"}, {"id": 853,...   2012-07-16
4  [{"id": 818, "name": "based on novel"}, {"id":...   2012-03-07

                                                cast  \
0  [{"cast_id": 242, "character": "Jake Sully", "...
1  [{"cast_id": 4, "character": "Captain Jack Spa...
2  [{"cast_id": 1, "character": "James Bond", "cr...
3  [{"cast_id": 2, "character": "Bruce Wayne / Ba...
4  [{"cast_id": 5, "character": "John Carter", "c...

                                                crew    status
0  [{"credit_id": "52fe48009251416c750aca23", "de...  Released
1  [{"credit_id": "52fe4232c3a36847f800b579", "de...  Released
2  [{"credit_id": "54805967c3a36829b5002c41", "de...  Released
3  [{"credit_id": "52fe4781c3a36847f81398c3", "de...  Released
4  [{"credit_id": "52fe479ac3a36847f813eaa3", "de...  Released
```

```
movies.isnull().sum()
```

```
movie_id         0
title            0
overview         3
genres           0
keywords         0
release_date     1
cast             0
crew             0
status           0
dtype: int64
```

Remove missing Data form Dataset:-

```
movies.dropna(inplace=True)
```

```
C:\Users\subha\AppData\Local\Temp\ipykernel_20920\3786870272.py:1:
SettingWithCopyWarning:
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
  movies.dropna(inplace=True)
```

```
movies.isnull().sum()
```

```
movie_id         0
title            0
overview         0
genres           0
keywords         0
release_date     0
cast             0
crew             0
status           0
dtype: int64
```

Calculate duplicate Data form Dataset:-

```
movies.duplicated().sum
```

```
<bound method Series.sum of 0       False
1       False
2       False
3       False
4       False
        ...
4804    False
4805    False
4806    False
```

```
4807    False
4808    False
Length: 4805, dtype: bool>
```

Change the formate of genres, keywords, cast, crew columns:-

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

movies['genres'] = movies['genres'].apply(convert)
```

```
C:\Users\subha\AppData\Local\Temp\ipykernel_20920\1189516248.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
  movies['genres'] = movies['genres'].apply(convert)
```

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

```
C:\Users\subha\AppData\Local\Temp\ipykernel_20920\3328512696.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
  movies['keywords'] = movies['keywords'].apply(convert)
```

```
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


movies['cast'] = movies['cast'].apply(convert3)
```

```
C:\Users\subha\AppData\Local\Temp\ipykernel_20920\3593664571.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
  movies['cast'] = movies['cast'].apply(convert3)
```

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

movies['crew'] = movies['crew'].apply(fatch_director)
```

```
C:\Users\subha\AppData\Local\Temp\ipykernel_20920\1823090780.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
  movies['crew'] = movies['crew'].apply(fatch_director)
```

After change the formate of columns, the tables are:-

```
movies.head()
```

```
   movie_id                                     title  \
0     19995                                    Avatar
1       285  Pirates of the Caribbean: At World's End
2    206647                                   Spectre
3     49026                     The Dark Knight Rises
4     49529                               John Carter

                                            overview  \
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...

                                        genres  \
0  [Action, Adventure, Fantasy, Science Fiction]
```

```
1                    [Adventure, Fantasy, Action]
2                      [Action, Adventure, Crime]
3                 [Action, Crime, Drama, Thriller]
4             [Action, Adventure, Science Fiction]
```

```
                                                keywords release_date  \
0  [culture clash, future, space war, space colon...   2009-12-10
1  [ocean, drug abuse, exotic island, east india ...   2007-05-19
2  [spy, based on novel, secret agent, sequel, mi...   2015-10-26
3  [dc comics, crime fighter, terrorist, secret i...   2012-07-16
4  [based on novel, mars, medallion, space travel...   2012-03-07

                                               cast
crew  \
0  [Sam Worthington, Zoe Saldana, Sigourney Weaver]      [James
Cameron]
1     [Johnny Depp, Orlando Bloom, Keira Knightley]     [Gore
Verbinski]
2      [Daniel Craig, Christoph Waltz, Léa Seydoux]         [Sam
Mendes]
3      [Christian Bale, Michael Caine, Gary Oldman]  [Christopher
Nolan]
4    [Taylor Kitsch, Lynn Collins, Samantha Morton]     [Andrew
Stanton]

     status
0  Released
1  Released
2  Released
3  Released
4  Released
```

Remove space between two words from keywords, genres, cast, crew:-

```
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])
```

```
C:\Users\subha\AppData\Local\Temp\ipykernel_20920\324806532.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
  movies['genres'] = movies['genres'].apply(lambda x:[i.replace("
","") for i in x])
C:\Users\subha\AppData\Local\Temp\ipykernel_20920\324806532.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
  movies['keywords'] = movies['keywords'].apply(lambda x:[i.replace("
","") for i in x])
C:\Users\subha\AppData\Local\Temp\ipykernel_20920\324806532.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
  movies['cast'] = movies['cast'].apply(lambda x:[i.replace(" ","")
for i in x])
C:\Users\subha\AppData\Local\Temp\ipykernel_20920\324806532.py:4:
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
  movies['crew'] = movies['crew'].apply(lambda x:[i.replace(" ","")
for i in x])

movies.head()

   movie_id                                     title  \
0     19995                                    Avatar
1       285  Pirates of the Caribbean: At World's End
2    206647                                   Spectre
3     49026                     The Dark Knight Rises
4     49529                               John Carter

                                            overview  \
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...
```

```
                                        genres  \
0  [Action, Adventure, Fantasy, ScienceFiction]
1                 [Adventure, Fantasy, Action]
2                   [Action, Adventure, Crime]
3              [Action, Crime, Drama, Thriller]
4           [Action, Adventure, ScienceFiction]

                                        keywords release_date  \
0  [cultureclash, future, spacewar, spacecolony, ...   2009-12-10
1  [ocean, drugabuse, exoticisland, eastindiatrad...   2007-05-19
2  [spy, basedonnovel, secretagent, sequel, mi6, ...   2015-10-26
3  [dccomics, crimefighter, terrorist, secretiden...   2012-07-16
4  [basedonnovel, mars, medallion, spacetravel, p...   2012-03-07

                                        cast                crew
status
0  [SamWorthington, ZoeSaldana, SigourneyWeaver]      [JamesCameron]
Released
1     [JohnnyDepp, OrlandoBloom, KeiraKnightley]     [GoreVerbinski]
Released
2      [DanielCraig, ChristophWaltz, LéaSeydoux]         [SamMendes]
Released
3      [ChristianBale, MichaelCaine, GaryOldman]  [ChristopherNolan]
Released
4    [TaylorKitsch, LynnCollins, SamanthaMorton]     [AndrewStanton]
Released
```

```
import re
def combine_columns(row):
    return [row['overview']] + row['genres'] + row['keywords'] +
row['cast'] + row['crew']


movies['tags'] = movies.apply(combine_columns, axis=1)
```

```
C:\Users\subha\AppData\Local\Temp\ipykernel_20920\1577708755.py:6:
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
  movies['tags'] = movies.apply(combine_columns, axis=1)
```

```
movies
```

```
      movie_id                                         title  \
0        19995                                        Avatar
1          285  Pirates of the Caribbean: At World's End
2       206647                                       Spectre
```

```
3        49026                       The Dark Knight Rises
4        49529                                 John Carter
...        ...                                         ...
4804      9367                                 El Mariachi
4805     72766                                   Newlyweds
4806    231617                 Signed, Sealed, Delivered
4807    126186                           Shanghai Calling
4808     25975                          My Date with Drew

                                                  overview  \
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 ...

                                               genres  \
0     [Action, Adventure, Fantasy, ScienceFiction]
1                     [Adventure, Fantasy, Action]
2                       [Action, Adventure, Crime]
3                  [Action, Crime, Drama, Thriller]
4               [Action, Adventure, ScienceFiction]
...                                             ...
4804                      [Action, Crime, Thriller]
4805                             [Comedy, Romance]
4806               [Comedy, Drama, Romance, TVMovie]
4807                                            []
4808                                 [Documentary]

                                               keywords
release_date  \
0     [cultureclash, future, spacewar, spacecolony, ...   2009-12-10

1     [ocean, drugabuse, exoticisland, eastindiatrad...   2007-05-19

2     [spy, basedonnovel, secretagent, sequel, mi6, ...   2015-10-26

3     [dccomics, crimefighter, terrorist, secretiden...   2012-07-16

4     [basedonnovel, mars, medallion, spacetravel, p...   2012-03-07

...                                                  ...          ...

4804  [unitedstates—mexicobarrier, legs, arms, paper...   1992-09-04
```

```
4805                                                  []   2011-12-26

4806  [date, loveatfirstsight, narration, investigat...   2013-10-13

4807                                                  []   2012-05-03

4808           [obsession, camcorder, crush, dreamgirl]   2005-08-05


                                                    cast
crew  \
0      [SamWorthington, ZoeSaldana, SigourneyWeaver]
[JamesCameron]
1         [JohnnyDepp, OrlandoBloom, KeiraKnightley]
[GoreVerbinski]
2          [DanielCraig, ChristophWaltz, LéaSeydoux]
[SamMendes]
3          [ChristianBale, MichaelCaine, GaryOldman]
[ChristopherNolan]
4        [TaylorKitsch, LynnCollins, SamanthaMorton]
[AndrewStanton]
...                                                ...                  .
..
4804  [CarlosGallardo, JaimedeHoyos, PeterMarquardt]
[RobertRodriguez]
4805       [EdwardBurns, KerryBishé, MarshaDietlein]
[EdwardBurns]
4806         [EricMabius, KristinBooth, CrystalLowe]
[ScottSmith]
4807          [DanielHenney, ElizaCoupe, BillPaxton]
[DanielHsia]
4808  [DrewBarrymore, BrianHerzlinger, CoreyFeldman]
[BrianHerzlinger]

        status                                                tags
0     Released  [In the 22nd century, a paraplegic Marine is d...
1     Released  [Captain Barbossa, long believed to be dead, h...
2     Released  [A cryptic message from Bond’s past sends him ...
3     Released  [Following the death of District Attorney Harv...
4     Released  [John Carter is a war-weary, former military c...
...        ...                                                ...
4804  Released  [El Mariachi just wants to play his guitar and...
4805  Released  [A newlywed couple's honeymoon is upended by t...
4806  Released  ["Signed, Sealed, Delivered" introduces a dedi...
4807  Released  [When ambitious New York attorney Sam is sent ...
4808  Released  [Ever since the second grade when he first saw...

[4805 rows x 10 columns]
```

Convert the "movies" dataset in a new dataset "new_df":-

```
new_df = movies[['movie_id','title','tags','release_date','status']]

new_df
```

```
      movie_id                                     title  \
0        19995                                    Avatar
1          285  Pirates of the Caribbean: At World's End
2       206647                                   Spectre
3        49026                     The Dark Knight Rises
4        49529                               John Carter
...        ...                                       ...
4804      9367                               El Mariachi
4805     72766                                 Newlyweds
4806    231617                  Signed, Sealed, Delivered
4807    126186                           Shanghai Calling
4808     25975                         My Date with Drew

                                                   tags release_date
status
0     [In the 22nd century, a paraplegic Marine is d...   2009-12-10
Released
1     [Captain Barbossa, long believed to be dead, h...   2007-05-19
Released
2     [A cryptic message from Bond’s past sends him ...   2015-10-26
Released
3     [Following the death of District Attorney Harv...   2012-07-16
Released
4     [John Carter is a war-weary, former military c...   2012-03-07
Released
...                                                 ...          ...
...
4804  [El Mariachi just wants to play his guitar and...   1992-09-04
Released
4805  [A newlywed couple's honeymoon is upended by t...   2011-12-26
Released
4806  ["Signed, Sealed, Delivered" introduces a dedi...   2013-10-13
Released
4807  [When ambitious New York attorney Sam is sent ...   2012-05-03
Released
4808  [Ever since the second grade when he first saw...   2005-08-05
Released

[4805 rows x 5 columns]
```

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

```
C:\Users\subha\AppData\Local\Temp\ipykernel_20920\3089450492.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
  new_df['tags'] = new_df['tags'].apply(lambda x:" ".join(x))
```

```
new_df
```

```
      movie_id                                         title  \
0        19995                                        Avatar
1          285  Pirates of the Caribbean: At World's End
2       206647                                       Spectre
3        49026                         The Dark Knight Rises
4        49529                                   John Carter
...        ...                                           ...
4804      9367                                   El Mariachi
4805     72766                                     Newlyweds
4806    231617                      Signed, Sealed, Delivered
4807    126186                               Shanghai Calling
4808     25975                             My Date with Drew

                                                   tags release_date
status
0     In the 22nd century, a paraplegic Marine is di...   2009-12-10
Released
1     Captain Barbossa, long believed to be dead, ha...   2007-05-19
Released
2     A cryptic message from Bond’s past sends him o...   2015-10-26
Released
3     Following the death of District Attorney Harve...   2012-07-16
Released
4     John Carter is a war-weary, former military ca...   2012-03-07
Released
...                                                 ...          ...
...
4804  El Mariachi just wants to play his guitar and ...   1992-09-04
Released
4805  A newlywed couple's honeymoon is upended by th...   2011-12-26
Released
4806  "Signed, Sealed, Delivered" introduces a dedic...   2013-10-13
Released
4807  When ambitious New York attorney Sam is sent t...   2012-05-03
Released
4808  Ever since the second grade when he first saw ...   2005-08-05
Released

[4805 rows x 5 columns]
```

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

```
C:\Users\subha\AppData\Local\Temp\ipykernel_20920\3214958533.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
  new_df['tags'] = new_df['tags'].apply(lambda x:x.lower())
```

Vectorization:-

```
import nltk
!pip install nltk
```

```
Requirement already satisfied: nltk in c:\users\subha\anaconda3\lib\
site-packages (3.8.1)
Requirement already satisfied: click in c:\users\subha\anaconda3\lib\
site-packages (from nltk) (8.1.7)
Requirement already satisfied: joblib in c:\users\subha\anaconda3\lib\
site-packages (from nltk) (1.2.0)
Requirement already satisfied: regex>=2021.8.3 in c:\users\subha\
anaconda3\lib\site-packages (from nltk) (2023.10.3)
Requirement already satisfied: tqdm in c:\users\subha\anaconda3\lib\
site-packages (from nltk) (4.65.0)
Requirement already satisfied: colorama in c:\users\subha\anaconda3\
lib\site-packages (from click->nltk) (0.4.6)
```

```
from nltk.stem.porter import PorterStemmer
ps = PorterStemmer()
```

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

```
new_df['tags'].apply(stem)
```

```
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: 4805, dtype: object
```

```
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(max_features=5000,stop_words='english')

vectors = cv.fit_transform(new_df['tags']).toarray()
```

Calculate Cosine distance of all vectors:-

```
from sklearn.metrics.pairwise import cosine_similarity

similarity = cosine_similarity(vectors)

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)

import pickle
pickle.dump(new_df.to_dict(),open('movies_dict.pkl','wb'))

pickle.dump(similarity,open('similarity.pkl','wb'))
```