In [1]:

```
pip install numpy
```

```
Requirement already satisfied: numpy in c:\users\adiso\appdata\local\programs\python\python39\lib\site-packages
(1.22.3)
Note: you may need to restart the kernel to use updated packages.
```

In [2]:

```
pip install pandas
```

```
Requirement already satisfied: pandas in c:\users\adiso\appdata\local\programs\python\python39\lib\site-package
s (1.4.1)Note: you may need to restart the kernel to use updated packages.
Requirement already satisfied: pytz>=2020.1 in c:\users\adiso\appdata\local\programs\python\python39\lib\site-p
ackages (from pandas) (2021.3)
Requirement already satisfied: numpy>=1.18.5 in c:\users\adiso\appdata\local\programs\python\python39\lib\site-
packages (from pandas) (1.22.3)
Requirement already satisfied: python-dateutil>=2.8.1 in c:\users\adiso\appdata\local\programs\python\python39\
lib\site-packages (from pandas) (2.8.2)
Requirement already satisfied: six>=1.5 in c:\users\adiso\appdata\local\programs\python\python39\lib\site-packa
ges (from python-dateutil>=2.8.1->pandas) (1.15.0)
```

In [3]:

```
import pandas as pd
import numpy as np
```

In [4]:

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

In [5]:

```
movies.head(1)
```

Out[5]:

	budget	genres	homepage	id	keywords	original_language	original_title	overview	popularity	production_comp
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...	150.437577	[{"name": "Ing Film Partners

In [6]:

```
credits.head(1)
```

Out[6]:

	movie_id	title	cast	crew
0	19995	Avatar	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...

In [7]:

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

Out[7]:

	budget	genres	homepage	id	keywords	original_language	original_title
0	237000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.avatarmovie.com/	19995	[{"id": 1463, "name": "culture clash"}, {"id":...	en	Avatar
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
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
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
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
...	...	...	...	...	...	...	...
4804	220000	[{"id": 28, "name": "Action"}, {"id": 80, "nam...	NaN	9367	[{"id": 5616, "name": "united states\u2013mexi...	es	El Mariachi
4805	9000	[{"id": 35, "name": "Comedy"}, {"id": 10749, "...	NaN	72766	[]	en	Newlyweds
4806	0	[{"id": 35, "name": "Comedy"}, {"id": 18, "nam...	http://www.hallmarkchannel.com/signedsealeddel...	231617	[{"id": 248, "name": "date"}, {"id": 699, "nam...	en	Signed, Sealed, Delivered
4807	0	[]	http://shanghaicalling.com/	126186	[]	en	Shanghai Calling
4808	0	[{"id": 99, "name": "Documentary"}]	NaN	25975	[{"id": 1523, "name": "obsession"}, {"id": 224...	en	My Date with Drew

4809 rows × 23 columns

In [8]:

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

Out[8]:

```
(4809, 23)
```

In [9]:

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

In [10]:

```
movies.head(1)
```

Out[10]:

	budget	genres	homepage	id	keywords	original_language	original_title	overview	popularity	production_comp
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...	150.437577	[{"name": "Ing Film Partners

1 rows × 23 columns

In [11]:

```
#only keep important data and truncate reduntant data
```

```
# movie_id
# cast
# crew --> Director Only
# genres
# title
# overview
# keywords

movies[['movie_id', 'title', 'overview', 'genres', 'keywords', 'cast', 'crew']]
```

Out[11]:

	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...
...	...	...	...	...	...	...	...
4804	9367	El Mariachi	El Mariachi just wants to play his guitar and ...	[{"id": 28, "name": "Action"}, {"id": 80, "nam...	[{"id": 5616, "name": "united states\u2013mexi...	[{"cast_id": 1, "character": "El Mariachi", "c...	[{"credit_id": "52fe44eec3a36847f80b280b", "de...
4805	72766	Newlyweds	A newlywed couple's honeymoon is upended by th...	[{"id": 35, "name": "Comedy"}, {"id": 10749, "...	[]	[{"cast_id": 1, "character": "Buzzy", "credit_...	[{"credit_id": "52fe487dc3a368484e0fb013", "de...
4806	231617	Signed, Sealed, Delivered	"Signed, Sealed, Delivered" introduces a dedic...	[{"id": 35, "name": "Comedy"}, {"id": 18, "nam...	[{"id": 248, "name": "date"}, {"id": 699, "nam...	[{"cast_id": 8, "character": "Oliver O\u2019To...	[{"credit_id": "52fe4df3c3a36847f8275ecf", "de...
4807	126186	Shanghai Calling	When ambitious New York attorney Sam is sent t...	[]	[]	[{"cast_id": 3, "character": "Sam", "credit_id...	[{"credit_id": "52fe4ad9c3a368484e16a36b", "de...
4808	25975	My Date with Drew	Ever since the second grade when he first saw ...	[{"id": 99, "name": "Documentary"}]	[{"id": 1523, "name": "obsession"}, {"id": 224...	[{"cast_id": 3, "character": "Herself", "credi...	[{"credit_id": "58ce021b9251415a390165d9", "de...

4809 rows × 7 columns

In [12]:

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

In [13]:

```
movies.head()
```

Out[13]:

	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 [14]:

```
#now check if there is any sort of missing data

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

Out[14]:
```
movie_id    0
title       0
overview    3
genres      0
keywords    0
cast        0
crew        0
dtype: int64
```

In [15]:
```
#now drop those 3 movies whose overview is unknown

movies.dropna(inplace=True)
```

In [16]:
```
movies.isnull().sum()
```

Out[16]:
```
movie_id    0
title       0
overview    0
genres      0
keywords    0
cast        0
crew        0
dtype: int64
```

In [17]:
```
#now there is no missing data in movies data frame
```

In [18]:
```
#now check for any duplicate data
movies.duplicated().sum()
```

Out[18]:
```
0
```

EDITING GENRE COLUMN

In [19]:
```
#check the format of genre column
movies.iloc[0].genres
```

Out[19]:
```
'[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 878, "na
me": "Science Fiction"}]'
```

In [20]:
```
# the genres are in "List of Dictionary" format
# convert it into a single list --> ['Action', 'Adventure', 'SciFi', 'Romance'...]

# first check type of genre list
type(movies.iloc[0].genres)
```

Out[20]:
```
str
```

In [21]:
```
# it is string so convert it into integer using --> 'ast' module

import ast

#helper function to return list of genres, keywords etc.
def convert(obj):
    List = []
    for i in ast.literal_eval(obj):
        List.append(i['name'])
    return List
```

In [22]:
```
# reverify the type of genre list
type(convert(movies.iloc[0].genres))
```

Out[22]:
```
list
```

In [23]:
```
# list of genres obtained :)

# store list in movies['genres']
movies['genres'] = movies['genres'].apply(convert)
```

In [24]:
```
movies.head()
```

Out[24]:

	movie_id	title	overview	genres	keywords	cast	crew
0	19995	Avatar	In the 22nd century, a paraplegic Marine is di...	[Action, Adventure, Fantasy, Science Fiction]	[{"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...	[Adventure, Fantasy, Action]	[{"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...	[Action, Adventure, Crime]	[{"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...	[Action, Crime, Drama, Thriller]	[{"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...	[Action, Adventure, Science Fiction]	[{"id": 818, "name": "based on novel"}, {"id":...	[{"cast_id": 5, "character": "John Carter", "c...	[{"credit_id": "52fe479ac3a36847f813eaa3", "de...

EDITING KEYWORDS COLUMN

In [25]:

```
# repeat same step on keywords column and convert list of dicts into list of strings
type(movies.iloc[0].keywords)
```

Out[25]:

```
str
```

In [26]:

```
type(convert(movies.iloc[0].keywords))
```

Out[26]:

```
list
```

In [27]:

```
#store result in movies['keywords']
movies['keywords'] = movies['keywords'].apply(convert)
```

In [28]:

```
movies.head()
```

Out[28]:

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

EDITING CAST COLUMN

In [29]:

```
#helper function to return first 3 cast names of each movie

def convertCast(obj):
    List = []
    counter = 0;
    for i in ast.literal_eval(obj):
        if counter < 3:
            List.append(i['name'])
            counter += 1
        else:
            break
    return List
```

In [30]:

```
#store result in movies['cast']
movies['cast'] = movies['cast'].apply(convertCast)
```

EDITING CREW COLUMN

In [31]:

```
#view crew column format
movies['crew'][0]
```

Out[31]:

```
'[{"credit_id": "52fe48009251416c750aca23", "department": "Editing", "gender": 0, "id": 1721, "job": "Editor",
"name": "Stephen E. Rivkin"}, {"credit_id": "539c47ecc3a36810e3001f87", "department": "Art", "gender": 2, "id":
496, "job": "Production Design", "name": "Rick Carter"}, {"credit_id": "54491c89c3a3680fb4001cf7", "department"
: "Sound", "gender": 0, "id": 900, "job": "Sound Designer", "name": "Christopher Boyes"}, {"credit_id": "54491c
b70e0a267480001bd0", "department": "Sound", "gender": 0, "id": 900, "job": "Supervising Sound Editor", "name":
"Christopher Boyes"}, {"credit_id": "539c4a4cc3a36810c9002101", "department": "Production", "gender": 1, "id":
1262, "job": "Casting", "name": "Mali Finn"}, {"credit_id": "5544ee3b925141499f0008fc", "department": "Sound",
"gender": 2, "id": 1729, "job": "Original Music Composer", "name": "James Horner"}, {"credit_id": "52fe48009251
416c750ac9c3", "department": "Directing", "gender": 2, "id": 2710, "job": "Director", "name": "James Cameron"},
{"credit_id": "52fe48009251416c750ac9d9", "department": "Writing", "gender": 2, "id": 2710, "job": "Writer", "n
ame": "James Cameron"}, {"credit_id": "52fe48009251416c750aca17", "department": "Editing", "gender": 2, "id": 2
710, "job": "Editor", "name": "James Cameron"}, {"credit_id": "52fe48009251416c750aca29", "department": "Produc
tion", "gender": 2, "id": 2710, "job": "Producer", "name": "James Cameron"}, {"credit_id": "52fe48009251416c750
aca3f", "department": "Writing", "gender": 2, "id": 2710, "job": "Screenplay", "name": "James Cameron"}, {"cred
it_id": "539c4987c3a36810ba0021a4", "department": "Art", "gender": 2, "id": 7236, "job": "Art Direction", "name
": "Andrew Menzies"}, {"credit_id": "549598c3c3a3686ae9004383", "department": "Visual Effects", "gender": 0, "i
d": 6690, "job": "Visual Effects Producer", "name": "Jill Brooks"}, {"credit_id": "52fe48009251416c750aca4b", "
department": "Production", "gender": 1, "id": 6347, "job": "Casting", "name": "Margery Simkin"}, {"credit_id":
"570b6f419251417da70032fe", "department": "Art", "gender": 2, "id": 6878, "job": "Supervising Art Director", "n
ame": "Kevin Ishioka"}, {"credit_id": "5495a0fac3a3686ae9004468", "department": "Sound", "gender": 0, "id": 688
3, "job": "Music Editor", "name": "Dick Bernstein"}, {"credit_id": "54959706c3a3686af3003e81", "department": "S
ound", "gender": 0, "id": 8159, "job": "Sound Effects Editor", "name": "Shannon Mills"}, {"credit_id": "54491d5
8c3a3680fb1001ccb", "department": "Sound", "gender": 0, "id": 8160, "job": "Foley", "name": "Dennie Thorpe"}, {
"credit_id": "54491d6cc3a3680fa5001b2c", "department": "Sound", "gender": 0, "id": 8163, "job": "Foley", "name"
: "Jana Vance"}, {"credit_id": "52fe48009251416c750aca57", "department": "Costume & Make-Up", "gender": 1, "id"
: 8527, "job": "Costume Design", "name": "Deborah Lynn Scott"}, {"credit_id": "52fe48009251416c750aca2f", "depa
rtment": "Production", "gender": 2, "id": 8529, "job": "Producer", "name": "Jon Landau"}, {"credit_id": "539c49
37c3a36810ba002194", "department": "Art", "gender": 0, "id": 9618, "job": "Art Direction", "name": "Sean Hawort
h"}, {"credit_id": "539c49b6c3a36810c10020e6", "department": "Art", "gender": 1, "id": 12653, "job": "Set Decor
ation", "name": "Kim Sinclair"}, {"credit_id": "570b6f2f9251413a0e00020d", "department": "Art", "gender": 1, "i
d": 12653, "job": "Supervising Art Director", "name": "Kim Sinclair"}, {"credit_id": "54491a6c0e0a26748c001b19"
, "department": "Art", "gender": 2, "id": 14350, "job": "Set Designer", "name": "Richard F. Mays"}, {"credit_id
": "56928cf4c3a3684cff0025c4", "department": "Production", "gender": 1, "id": 20294, "job": "Executive Producer
", "name": "Laeta Kalogridis"}, {"credit_id": "52fe48009251416c750aca51", "department": "Costume & Make-Up", "g
ender": 0, "id": 17675, "job": "Costume Design", "name": "Mayes C. Rubeo"}, {"credit_id": "52fe48009251416c750a
ca11", "department": "Camera", "gender": 2, "id": 18265, "job": "Director of Photography", "name": "Mauro Fiore
"}, {"credit_id": "5449194d0e0a26748f001b39", "department": "Art", "gender": 0, "id": 42281, "job": "Set Design
er", "name": "Scott Herbertson"}, {"credit_id": "52fe48009251416c750aca05", "department": "Crew", "gender": 0,
"id": 42288, "job": "Stunts", "name": "Woody Schultz"}, {"credit_id": "5592aefb92514152de0010f5", "department":
"Costume & Make-Up", "gender": 0, "id": 29067, "job": "Makeup Artist", "name": "Linda DeVetta"}, {"credit_id":
"5592afa492514152de00112c", "department": "Costume & Make-Up", "gender": 0, "id": 29067, "job": "Hairstylist",
"name": "Linda DeVetta"}, {"credit_id": "54959ed592514130fc002e5d", "department": "Camera", "gender": 2, "id":
33302, "job": "Camera Operator", "name": "Richard Bluck"}, {"credit_id": "539c4891c3a36810ba002147", "departmen
t": "Art", "gender": 2, "id": 33303, "job": "Art Direction", "name": "Simon Bright"}, {"credit_id": "54959c0692
51417a81001f3a", "department": "Visual Effects", "gender": 0, "id": 113145, "job": "Visual Effects Supervisor",
"name": "Richard Martin"}, {"credit_id": "54959a0dc3a3680ff5002c8d", "department": "Crew", "gender": 2, "id": 5
8188, "job": "Visual Effects Editor", "name": "Steve R. Moore"}, {"credit_id": "52fe48009251416c750aca1d", "dep
artment": "Editing", "gender": 2, "id": 58871, "job": "Editor", "name": "John Refoua"}, {"credit_id": "54491a4d
c3a3680fc30018ca", "department": "Art", "gender": 0, "id": 92359, "job": "Set Designer", "name": "Karl J. Marti
n"}, {"credit_id": "52fe48009251416c750aca35", "department": "Camera", "gender": 1, "id": 72201, "job": "Direct
or of Photography", "name": "Chiling Lin"}, {"credit_id": "52fe48009251416c750ac9ff", "department": "Crew", "ge
nder": 0, "id": 89714, "job": "Stunts", "name": "Ilram Choi"}, {"credit_id": "54959c529251416e2b004394", "depar
tment": "Visual Effects", "gender": 2, "id": 93214, "job": "Visual Effects Supervisor", "name": "Steven Quale"}
, {"credit_id": "54491edf0e0a267489001c37", "department": "Crew", "gender": 1, "id": 122607, "job": "Dialect Co
ach", "name": "Carla Meyer"}, {"credit_id": "539c485bc3a368653d001a3a", "department": "Art", "gender": 2, "id":
132585, "job": "Art Direction", "name": "Nick Bassett"}, {"credit_id": "539c4903c3a368653d001a74", "department"
: "Art", "gender": 0, "id": 132596, "job": "Art Direction", "name": "Jill Cormack"}, {"credit_id": "539c4967c3a
368653d001a94", "department": "Art", "gender": 0, "id": 132604, "job": "Art Direction", "name": "Andy McLaren"}
, {"credit_id": "52fe48009251416c750aca45", "department": "Crew", "gender": 0, "id": 236696, "job": "Motion Cap
ture Artist", "name": "Terry Notary"}, {"credit_id": "54959e02c3a3680fc60027d2", "department": "Crew", "gender"
: 2, "id": 956198, "job": "Stunt Coordinator", "name": "Garrett Warren"}, {"credit_id": "54959ca3c3a3686ae30043
8c", "department": "Visual Effects", "gender": 2, "id": 957874, "job": "Visual Effects Supervisor", "name": "Jo
nathan Rothbart"}, {"credit_id": "570b6f519251412c74001b2f", "department": "Art", "gender": 0, "id": 957889, "j
ob": "Supervising Art Director", "name": "Stefan Dechant"}, {"credit_id": "570b6f62c3a3680b77007460", "departme
nt": "Art", "gender": 2, "id": 959555, "job": "Supervising Art Director", "name": "Todd Cherniawsky"}, {"credit
_id": "539c4a3ac3a36810da0021cc", "department": "Production", "gender": 0, "id": 1016177, "job": "Casting", "na
me": "Miranda Rivers"}, {"credit_id": "539c482cc3a36810c1002062", "department": "Art", "gender": 0, "id": 10325
36, "job": "Production Design", "name": "Robert Stromberg"}, {"credit_id": "539c4b65c3a36810c9002125", "departm
ent": "Costume & Make-Up", "gender": 2, "id": 1071680, "job": "Costume Design", "name": "John Harding"}, {"cred
it_id": "54959e6692514130fc002e4e", "department": "Camera", "gender": 0, "id": 1177364, "job": "Steadicam Opera
tor", "name": "Roberto De Angelis"}, {"credit_id": "539c49f1c3a368653d001aac", "department": "Costume & Make-Up
", "gender": 2, "id": 1202850, "job": "Makeup Department Head", "name": "Mike Smithson"}, {"credit_id": "549599
9ec3a3686ae100460c", "department": "Visual Effects", "gender": 0, "id": 1204668, "job": "Visual Effects Produce
r", "name": "Alain Lalanne"}, {"credit_id": "54959cdfc3a3681153002729", "department": "Visual Effects", "gender
": 0, "id": 1206410, "job": "Visual Effects Supervisor", "name": "Lucas Salton"}, {"credit_id": "54959623925141
7a81001eae", "department": "Crew", "gender": 0, "id": 1234266, "job": "Post Production Supervisor", "name": "Ja
nace Tashjian"}, {"credit_id": "54959c859251416e1e003efe", "department": "Visual Effects", "gender": 0, "id": 1
271932, "job": "Visual Effects Supervisor", "name": "Stephen Rosenbaum"}, {"credit_id": "5592af28c3a368775a0010
5f", "department": "Costume & Make-Up", "gender": 0, "id": 1310064, "job": "Makeup Artist", "name": "Frankie Ka
rena"}, {"credit_id": "539c4adfc3a36810e300203b", "department": "Costume & Make-Up", "gender": 1, "id": 1319844
, "job": "Costume Supervisor", "name": "Lisa Lovaas"}, {"credit_id": "54959b579251416e2b004371", "department":
"Visual Effects", "gender": 0, "id": 1327028, "job": "Visual Effects Supervisor", "name": "Jonathan Fawkner"},
{"credit_id": "539c48a7c3a36810b5001fa7", "department": "Art", "gender": 0, "id": 1330561, "job": "Art Directio
n", "name": "Robert Bavin"}, {"credit_id": "539c4a71c3a36810da0021e0", "department": "Costume & Make-Up", "gend
```

```
er": 0, "id": 1330567, "job": "Costume Supervisor", "name": "Anthony Almaraz"}, {"credit_id": "539c4a8ac3a36810
ba0021e4", "department": "Costume & Make-Up", "gender": 0, "id": 1330570, "job": "Costume Supervisor", "name":
"Carolyn M. Fenton"}, {"credit_id": "539c4ab6c3a36810da0021f0", "department": "Costume & Make-Up", "gender": 0,
"id": 1330574, "job": "Costume Supervisor", "name": "Beth Koenigsberg"}, {"credit_id": "54491ab70e0a267480001ba
2", "department": "Art", "gender": 0, "id": 1336191, "job": "Set Designer", "name": "Sam Page"}, {"credit_id":
"544919d9c3a3680fc30018bd", "department": "Art", "gender": 0, "id": 1339441, "job": "Set Designer", "name": "Te
x Kadonaga"}, {"credit_id": "54491cf50e0a267483001b0c", "department": "Editing", "gender": 0, "id": 1352422, "j
ob": "Dialogue Editor", "name": "Kim Foscato"}, {"credit_id": "544919f40e0a26748c001b09", "department": "Art",
"gender": 0, "id": 1352962, "job": "Set Designer", "name": "Tammy S. Lee"}, {"credit_id": "5495a115c3a3680ff500
2d71", "department": "Crew", "gender": 0, "id": 1357070, "job": "Transportation Coordinator", "name": "Denny Ca
ira"}, {"credit_id": "5495a12f92514130fc002e94", "department": "Crew", "gender": 0, "id": 1357071, "job": "Tran
sportation Coordinator", "name": "James Waitkus"}, {"credit_id": "5495976fc3a36811530026b0", "department": "Sou
nd", "gender": 0, "id": 1360103, "job": "Supervising Sound Editor", "name": "Addison Teague"}, {"credit_id": "5
4491837c3a3680fb1001c5a", "department": "Art", "gender": 2, "id": 1376887, "job": "Set Designer", "name": "C. S
cott Baker"}, {"credit_id": "54491878c3a3680fb4001c9d", "department": "Art", "gender": 0, "id": 1376888, "job":
"Set Designer", "name": "Luke Caska"}, {"credit_id": "544918dac3a3680fa5001ae0", "department": "Art", "gender":
0, "id": 1376889, "job": "Set Designer", "name": "David Chow"}, {"credit_id": "544919110e0a267486001b68", "depa
rtment": "Art", "gender": 0, "id": 1376890, "job": "Set Designer", "name": "Jonathan Dyer"}, {"credit_id": "544
91967c3a3680faa001b5e", "department": "Art", "gender": 0, "id": 1376891, "job": "Set Designer", "name": "Joseph
Hiura"}, {"credit_id": "54491997c3a3680fb1001c8a", "department": "Art", "gender": 0, "id": 1376892, "job": "Art
Department Coordinator", "name": "Rebecca Jellie"}, {"credit_id": "544919ba0e0a26748f001b42", "department": "Ar
t", "gender": 0, "id": 1376893, "job": "Set Designer", "name": "Robert Andrew Johnson"}, {"credit_id": "54491b1
dc3a3680faa001b8c", "department": "Art", "gender": 0, "id": 1376895, "job": "Assistant Art Director", "name": "
Mike Stassi"}, {"credit_id": "54491b79c3a3680fbb001826", "department": "Art", "gender": 0, "id": 1376897, "job"
: "Construction Coordinator", "name": "John Villarino"}, {"credit_id": "54491baec3a3680fb4001ce6", "department"
: "Art", "gender": 2, "id": 1376898, "job": "Assistant Art Director", "name": "Jeffrey Wisniewski"}, {"credit_i
d": "54491d2fc3a3680fb4001d07", "department": "Editing", "gender": 0, "id": 1376899, "job": "Dialogue Editor",
"name": "Cheryl Nardi"}, {"credit_id": "54491d86c3a3680fa5001b2f", "department": "Editing", "gender": 0, "id":
1376901, "job": "Dialogue Editor", "name": "Marshall Winn"}, {"credit_id": "54491d9dc3a3680faa001bb0", "departm
ent": "Sound", "gender": 0, "id": 1376902, "job": "Supervising Sound Editor", "name": "Gwendolyn Yates Whittle"
}, {"credit_id": "54491dc10e0a267486001bce", "department": "Sound", "gender": 0, "id": 1376903, "job": "Sound R
e-Recording Mixer", "name": "William Stein"}, {"credit_id": "54491f500e0a26747c001c07", "department": "Crew", "
gender": 0, "id": 1376909, "job": "Choreographer", "name": "Lula Washington"}, {"credit_id": "549599239251412c4
e002a2e", "department": "Visual Effects", "gender": 0, "id": 1391692, "job": "Visual Effects Producer", "name":
"Chris Del Conte"}, {"credit_id": "54959d54c3a36831b8001d9a", "department": "Visual Effects", "gender": 2, "id"
: 1391695, "job": "Visual Effects Supervisor", "name": "R. Christopher White"}, {"credit_id": "54959bdf9251412c
4e002a66", "department": "Visual Effects", "gender": 0, "id": 1394070, "job": "Visual Effects Supervisor", "nam
e": "Dan Lemmon"}, {"credit_id": "5495971d92514132ed002922", "department": "Sound", "gender": 0, "id": 1394129,
"job": "Sound Effects Editor", "name": "Tim Nielsen"}, {"credit_id": "5592b25792514152cc0011aa", "department":
"Crew", "gender": 0, "id": 1394286, "job": "CG Supervisor", "name": "Michael Mulholland"}, {"credit_id": "54959
a329251416e2b004355", "department": "Crew", "gender": 0, "id": 1394750, "job": "Visual Effects Editor", "name":
"Thomas Nittmann"}, {"credit_id": "54959d6dc3a3686ae9004401", "department": "Visual Effects", "gender": 0, "id"
: 1394755, "job": "Visual Effects Supervisor", "name": "Edson Williams"}, {"credit_id": "5495a08fc3a3686ae30044
1c", "department": "Editing", "gender": 0, "id": 1394953, "job": "Digital Intermediate", "name": "Christine Car
r"}, {"credit_id": "55402d659251413d6d000249", "department": "Visual Effects", "gender": 0, "id": 1395269, "job
": "Visual Effects Supervisor", "name": "John Bruno"}, {"credit_id": "54959e7b9251416e1e003f3e", "department":
"Camera", "gender": 0, "id": 1398970, "job": "Steadicam Operator", "name": "David Emmerichs"}, {"credit_id": "5
4959734c3a3686ae10045e0", "department": "Sound", "gender": 0, "id": 1400906, "job": "Sound Effects Editor", "na
me": "Christopher Scarabosio"}, {"credit_id": "549595dd92514130fc002d79", "department": "Production", "gender":
0, "id": 1401784, "job": "Production Supervisor", "name": "Jennifer Teves"}, {"credit_id": "549596009251413af70
028cc", "department": "Production", "gender": 0, "id": 1401785, "job": "Production Manager", "name": "Brigitte
Yorke"}, {"credit_id": "549596e892514130fc002d99", "department": "Sound", "gender": 0, "id": 1401786, "job": "S
ound Effects Editor", "name": "Ken Fischer"}, {"credit_id": "549598229251412c4e002a1c", "department": "Crew", "
gender": 0, "id": 1401787, "job": "Special Effects Coordinator", "name": "Iain Hutton"}, {"credit_id": "5495983
49251416e2b00432b", "department": "Crew", "gender": 0, "id": 1401788, "job": "Special Effects Coordinator", "na
me": "Steve Ingram"}, {"credit_id": "54959905c3a3686ae3004324", "department": "Visual Effects", "gender": 0, "i
d": 1401789, "job": "Visual Effects Producer", "name": "Joyce Cox"}, {"credit_id": "5495994b92514132ed002951",
"department": "Visual Effects", "gender": 0, "id": 1401790, "job": "Visual Effects Producer", "name": "Jenny Fo
ster"}, {"credit_id": "549599cbc3a3686ae1004613", "department": "Crew", "gender": 0, "id": 1401791, "job": "Vis
ual Effects Editor", "name": "Christopher Marino"}, {"credit_id": "549599f2c3a3686ae100461e", "department": "Cr
ew", "gender": 0, "id": 1401792, "job": "Visual Effects Editor", "name": "Jim Milton"}, {"credit_id": "54959a51
c3a3686af3003eb5", "department": "Visual Effects", "gender": 0, "id": 1401793, "job": "Visual Effects Producer"
, "name": "Cyndi Ochs"}, {"credit_id": "54959a7cc3a36811530026f4", "department": "Crew", "gender": 0, "id": 140
1794, "job": "Visual Effects Editor", "name": "Lucas Putnam"}, {"credit_id": "54959b91c3a3680ff5002cb4", "depar
tment": "Visual Effects", "gender": 0, "id": 1401795, "job": "Visual Effects Supervisor", "name": "Anthony \'Ma
x\' Ivins"}, {"credit_id": "54959bb69251412c4e002a5f", "department": "Visual Effects", "gender": 0, "id": 14017
96, "job": "Visual Effects Supervisor", "name": "John Knoll"}, {"credit_id": "54959cbbc3a3686ae3004391", "depar
tment": "Visual Effects", "gender": 2, "id": 1401799, "job": "Visual Effects Supervisor", "name": "Eric Saindon
"}, {"credit_id": "54959d06c3a3686ae90043f6", "department": "Visual Effects", "gender": 0, "id": 1401800, "job"
: "Visual Effects Supervisor", "name": "Wayne Stables"}, {"credit_id": "54959d259251416e1e003f11", "department"
: "Visual Effects", "gender": 0, "id": 1401801, "job": "Visual Effects Supervisor", "name": "David Stinnett"},
{"credit_id": "54959db49251413af7002975", "department": "Visual Effects", "gender": 0, "id": 1401803, "job": "V
isual Effects Supervisor", "name": "Guy Williams"}, {"credit_id": "54959de4c3a3681153002750", "department": "Cr
ew", "gender": 0, "id": 1401804, "job": "Stunt Coordinator", "name": "Stuart Thorp"}, {"credit_id": "54959ef2c3
a3680fc60027f2", "department": "Lighting", "gender": 0, "id": 1401805, "job": "Best Boy Electric", "name": "Gil
es Coburn"}, {"credit_id": "54959f07c3a3680fc60027f9", "department": "Camera", "gender": 2, "id": 1401806, "job
": "Still Photographer", "name": "Mark Fellman"}, {"credit_id": "54959f47c3a3681153002774", "department": "Ligh
ting", "gender": 0, "id": 1401807, "job": "Lighting Technician", "name": "Scott Sprague"}, {"credit_id": "54959
f8cc3a36831b8001df2", "department": "Visual Effects", "gender": 0, "id": 1401808, "job": "Animation Director",
"name": "Jeremy Hollobon"}, {"credit_id": "54959fa0c3a36831b8001dfb", "department": "Visual Effects", "gender":
0, "id": 1401809, "job": "Animation Director", "name": "Orlando Meunier"}, {"credit_id": "54959fb6c3a3686af3003
f54", "department": "Visual Effects", "gender": 0, "id": 1401810, "job": "Animation Director", "name": "Taisuke
Tanimura"}, {"credit_id": "54959fd2c3a36831b8001e02", "department": "Costume & Make-Up", "gender": 0, "id": 140
1812, "job": "Set Costumer", "name": "Lilia Mishel Acevedo"}, {"credit_id": "54959ff9c3a3686ae300440c", "depart
ment": "Costume & Make-Up", "gender": 0, "id": 1401814, "job": "Set Costumer", "name": "Alejandro M. Hernandez"
}, {"credit_id": "5495a0ddc3a3686ae10046fe", "department": "Editing", "gender": 0, "id": 1401815, "job": "Digit
al Intermediate", "name": "Marvin Hall"}, {"credit_id": "5495a1f7c3a3686ae3004443", "department": "Production",
"gender": 0, "id": 1401816, "job": "Publicist", "name": "Judy Alley"}, {"credit_id": "5592b29fc3a36869d100002f"
, "department": "Crew", "gender": 0, "id": 1418381, "job": "CG Supervisor", "name": "Mike Perry"}, {"credit_id"
```

```
: "5592b23a9251415df8001081", "department": "Crew", "gender": 0, "id": 1426854, "job": "CG Supervisor", "name":
"Andrew Morley"}, {"credit_id": "55491e1192514104c40002d8", "department": "Art", "gender": 0, "id": 1438901, "j
ob": "Conceptual Design", "name": "Seth Engstrom"}, {"credit_id": "5525d5809251417276002b06", "department": "Cr
ew", "gender": 0, "id": 1447362, "job": "Visual Effects Art Director", "name": "Eric Oliver"}, {"credit_id": "5
54427ca925141586500312a", "department": "Visual Effects", "gender": 0, "id": 1447503, "job": "Modeling", "name"
: "Matsune Suzuki"}, {"credit_id": "551906889251415aab001c88", "department": "Art", "gender": 0, "id": 1447524,
"job": "Art Department Manager", "name": "Paul Tobin"}, {"credit_id": "5592af8492514152cc0010de", "department":
"Costume & Make-Up", "gender": 0, "id": 1452643, "job": "Hairstylist", "name": "Roxane Griffin"}, {"credit_id":
"553d3c109251415852001318", "department": "Lighting", "gender": 0, "id": 1453938, "job": "Lighting Artist", "na
me": "Arun Ram-Mohan"}, {"credit_id": "5592af4692514152d5001355", "department": "Costume & Make-Up", "gender":
0, "id": 1457305, "job": "Makeup Artist", "name": "Georgia Lockhart-Adams"}, {"credit_id": "5592b2eac3a36877470
012a5", "department": "Crew", "gender": 0, "id": 1466035, "job": "CG Supervisor", "name": "Thrain Shadbolt"}, {
"credit_id": "5592b032c3a36877450015f1", "department": "Crew", "gender": 0, "id": 1483220, "job": "CG Superviso
r", "name": "Brad Alexander"}, {"credit_id": "5592b05592514152d80012f6", "department": "Crew", "gender": 0, "id
": 1483221, "job": "CG Supervisor", "name": "Shadi Almassizadeh"}, {"credit_id": "5592b090c3a36877570010b5", "d
epartment": "Crew", "gender": 0, "id": 1483222, "job": "CG Supervisor", "name": "Simon Clutterbuck"}, {"credit_
id": "5592b0dbc3a368774b00112c", "department": "Crew", "gender": 0, "id": 1483223, "job": "CG Supervisor", "nam
e": "Graeme Demmocks"}, {"credit_id": "5592b0fe92514152db0010c1", "department": "Crew", "gender": 0, "id": 1483
224, "job": "CG Supervisor", "name": "Adrian Fernandes"}, {"credit_id": "5592b11f9251415df8001059", "department
": "Crew", "gender": 0, "id": 1483225, "job": "CG Supervisor", "name": "Mitch Gates"}, {"credit_id": "5592b15dc
3a3687745001645", "department": "Crew", "gender": 0, "id": 1483226, "job": "CG Supervisor", "name": "Jerry Kung
"}, {"credit_id": "5592b18e925141645a0004ae", "department": "Crew", "gender": 0, "id": 1483227, "job": "CG Supe
rvisor", "name": "Andy Lomas"}, {"credit_id": "5592b1bfc3a368775d0010e7", "department": "Crew", "gender": 0, "i
d": 1483228, "job": "CG Supervisor", "name": "Sebastian Marino"}, {"credit_id": "5592b2049251415df8001078", "de
partment": "Crew", "gender": 0, "id": 1483229, "job": "CG Supervisor", "name": "Matthias Menz"}, {"credit_id":
"5592b27b92514152d800136a", "department": "Crew", "gender": 0, "id": 1483230, "job": "CG Supervisor", "name": "
Sergei Nevshupov"}, {"credit_id": "5592b2c3c3a36869e800003c", "department": "Crew", "gender": 0, "id": 1483231,
"job": "CG Supervisor", "name": "Philippe Rebours"}, {"credit_id": "5592b317c3a36877470012af", "department": "C
rew", "gender": 0, "id": 1483232, "job": "CG Supervisor", "name": "Michael Takarangi"}, {"credit_id": "5592b345
c3a36877470012bb", "department": "Crew", "gender": 0, "id": 1483233, "job": "CG Supervisor", "name": "David Wei
tzberg"}, {"credit_id": "5592b37cc3a368775100113b", "department": "Crew", "gender": 0, "id": 1483234, "job": "C
G Supervisor", "name": "Ben White"}, {"credit_id": "573c8e2f9251413f5d000094", "department": "Crew", "gender":
1, "id": 1621932, "job": "Stunts", "name": "Min Windle"}]'
```

In [32]:

```python
#we only need director for our purpose

#helper function to fetch director
def fetch_director(obj):
    List = []
    for i in ast.literal_eval(obj):
        if i['job'] == "Director":
            List.append(i['name'])
            break
    return List
```

In [33]:

```python
movies['crew'] = movies['crew'].apply(fetch_director)
```

In [34]:

```python
movies.head()
```

Out[34]:

	movie_id	title	overview	genres	keywords	cast	crew
0	19995	Avatar	In the 22nd century, a paraplegic Marine is di...	[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 dead, ha...	[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 sends him o...	[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 Harve...	[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 military ca...	[Action, Adventure, Science Fiction]	[based on novel, mars, medallion, space travel...	[Taylor Kitsch, Lynn Collins, Samantha Morton]	[Andrew Stanton]

EDITING OVERVIEW COLUMN

In [35]:

```python
# check overview data format
movies['overview'][0]
```

Out[35]:

```
'In the 22nd century, a paraplegic Marine is dispatched to the moon Pandora on a unique mission, but becomes to
rn between following orders and protecting an alien civilization.'
```

In [36]:

```python
# it is in string format --> convert it into a list and store in movies['overview']
movies['overview'] = movies['overview'].apply(lambda x: x.split())
```

In [37]:

```python
#verify type
type(movies['overview'][0])
```

Out[37]:

```
list
```

In [38]:

```
movies.head()
```

Out[38]:

	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]

Remove spaces between words of same entity

In [39]:

```
# repeat this step for genres keywords cast and 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])
```

In [40]:

```
movies.head()
```

Out[40]:

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

Create a final Data Frame Composed only of 3 columns

movie_id, title & tags

In [41]:

```
# here tag column contains every detail of overview, genre, keywords, cast & crew

movies['tags'] = movies['overview'] + movies['genres'] + movies['keywords'] + movies['cast'] + movies['crew']
```

In [42]:

```
# remove redundant columns and create a new data frame

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

In [43]:

```
# convert lists in tags column into strings

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

```
C:\Users\adiso\AppData\Local\Temp\ipykernel_11548\2487449930.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#ret
urning-a-view-versus-a-copy
  new_df['tags'] = new_df['tags'].apply(lambda x: " ".join(x))
```

In [44]:

```
new_df['tags'][0]
```

Out[44]:

```
'In the 22nd century, a paraplegic Marine is dispatched to the moon Pandora on a unique mission, but becomes to
rn between following orders and protecting an alien civilization. Action Adventure Fantasy ScienceFiction cultu
reclash future spacewar spacecolony society spacetravel futuristic romance space alien tribe alienplanet cgi ma
rine soldier battle loveaffair antiwar powerrelations mindandsoul 3d SamWorthington ZoeSaldana SigourneyWeaver
JamesCameron'
```

In [45]:

```
# convert tags strings into Lower Case --> RECOMMENDED

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

```
C:\Users\adiso\AppData\Local\Temp\ipykernel_11548\1211957509.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#ret
urning-a-view-versus-a-copy
  new_df['tags'] = new_df['tags'].apply(lambda x: x.lower())
```

In [46]:

```
# OUR FINAL DATA FRAME

new_df.head()
```

Out[46]:

	movie_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...

Applying text vectorization on tags column using "Bag of Words" technique

In [47]:

```
pip install sklearn
```

```
Requirement already satisfied: sklearn in c:\users\adiso\appdata\local\programs\python\python39\lib\site-packag
es (0.0)
Requirement already satisfied: scikit-learn in c:\users\adiso\appdata\local\programs\python\python39\lib\site-p
ackages (from sklearn) (1.0.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\adiso\appdata\local\programs\python\python39\li
b\site-packages (from scikit-learn->sklearn) (3.1.0)
Requirement already satisfied: joblib>=0.11 in c:\users\adiso\appdata\local\programs\python\python39\lib\site-p
ackages (from scikit-learn->sklearn) (1.1.0)
Requirement already satisfied: numpy>=1.14.6 in c:\users\adiso\appdata\local\programs\python\python39\lib\site-
packages (from scikit-learn->sklearn) (1.22.3)
Requirement already satisfied: scipy>=1.1.0 in c:\users\adiso\appdata\local\programs\python\python39\lib\site-p
ackages (from scikit-learn->sklearn) (1.8.0)
Note: you may need to restart the kernel to use updated packages.
```

In [48]:

```
# import countVectorizer of scikit learn
from sklearn.feature_extraction.text import CountVectorizer

# create a countvectorizer object to set no. of words & drop stop words
cv = CountVectorizer(max_features=5000, stop_words='english')
```

In [51]:

```
# convert sparse matrix returned by countVectorizer into numpy array

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

In [52]:

```
vectors
```

Out[52]:

```
array([[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],
       [0, 0, 0, ..., 0, 0, 0]], dtype=int64)
```

Apply Stemming to reduce infected words & convert them to stem

Apply Stemming to reduce inflected words & convert them to stem word

In [54]:
```
pip install nltk
```

```
Collecting nltk
  Downloading nltk-3.7-py3-none-any.whl (1.5 MB)
     ---------------------------------------- 1.5/1.5 MB 5.3 MB/s eta 0:00:00
Collecting click
  Downloading click-8.0.4-py3-none-any.whl (97 kB)
     ---------------------------------------- 97.5/97.5 KB 2.8 MB/s eta 0:00:00
Requirement already satisfied: tqdm in c:\users\adiso\appdata\local\programs\python\python39\lib\site-packages
(from nltk) (4.63.0)
Requirement already satisfied: joblib in c:\users\adiso\appdata\local\programs\python\python39\lib\site-package
s (from nltk) (1.1.0)
Collecting regex>=2021.8.3
  Downloading regex-2022.3.15-cp39-cp39-win_amd64.whl (274 kB)
     -------------------------------------- 274.4/274.4 KB 5.6 MB/s eta 0:00:00
Requirement already satisfied: colorama in c:\users\adiso\appdata\local\programs\python\python39\lib\site-packa
ges (from click->nltk) (0.4.4)
Installing collected packages: regex, click, nltk
Successfully installed click-8.0.4 nltk-3.7 regex-2022.3.15
Note: you may need to restart the kernel to use updated packages.
```

In [55]:
```
import nltk
```

In [56]:
```
from nltk.stem.porter import PorterStemmer
ps = PorterStemmer()
```

In [57]:
```
def stem(text):
    List = []
    for i in text.split():
        List.append(ps.stem(i))
    string = " ".join(List)
    return string
```

In [60]:
```
# testing stem function
new_df['tags'][0]
```

Out[60]:
```
'in the 22nd century, a paraplegic marine is dispatched to the moon pandora on a unique mission, but becomes to
rn between following orders and protecting an alien civilization. action adventure fantasy sciencefiction cultu
reclash future spacewar spacecolony society spacetravel futuristic romance space alien tribe alienplanet cgi ma
rine soldier battle loveaffair antiwar powerrelations mindandsoul 3d samworthington zoesaldana sigourneyweaver
jamescameron'
```

In [62]:
```
# after stemming
stem('in the 22nd century, a paraplegic marine is dispatched to the moon pandora on a unique mission, but becom
```

Out[62]:
```
'in the 22nd century, a parapleg marin is dispatch to the moon pandora on a uniqu mission, but becom torn betwe
en follow order and protect an alien civilization. action adventur fantasi sciencefict cultureclash futur space
war spacecoloni societi spacetravel futurist romanc space alien tribe alienplanet cgi marin soldier battl lovea
ffair antiwar powerrel mindandsoul 3d samworthington zoesaldana sigourneyweav jamescameron'
```

In [64]:
```
# store it in new_df['tags']
new_df['tags'] = new_df['tags'].apply(stem)
```

```
C:\Users\adiso\AppData\Local\Temp\ipykernel_11548\1324723276.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#ret
urning-a-view-versus-a-copy
  new_df['tags'] = new_df['tags'].apply(stem)
```

Calcuate Cosine Similarity for every movie

In [65]:
```
from sklearn.metrics.pairwise import cosine_similarity
```

In [66]:
```
new_df['tags'].shape
```

Out[66]:
```
(4806,)
```

In [67]:
```
# calculate cosine distance of every movie with another

similarity = cosine_similarity(vectors)
```

In [68]:
```
similarity.shape
```

Out[68]:
```
(4806, 4806)
```

In [69]:
```
# eg. shape[0] denotes distance between first and first, second, third ... 4806th movie
```

```
similarity[0]
```

Out[69]:

```
array([1.        , 0.08964215, 0.05976143, ..., 0.02519763, 0.02817181,
       0.        ])
```

In [102…

```
# create a helper function which returns top 10 movies when a movie is provided
def recommend(movie):
    # fetch movie index
    movie_index = new_df[new_df['title'] == movie].index[0]

    # obtain distances of this movie with every other movie including itself
    distances = similarity[movie_index]

    # reverse sort distances array on the basis of distance to obtain nearest movies
    # by keeping indices intact
    movie_list = sorted(list(enumerate(similarity[movie_index])), reverse=True, key = lambda x: x[1])

    # store recommeded list of top 10 similar movies
    recommended_list = movie_list[1: 11]

    for i in recommended_list:
        # simply print on the basis of title of the movie
        print(new_df.iloc[i[0]].title)
```

In [95]:

```
new_df[new_df['title'] == 'Spectre'].index
```

Out[95]:

```
Int64Index([2], dtype='int64')
```

In [104…

```
recommend('X-Men')
```

```
X2
X-Men: The Last Stand
X-Men: First Class
X-Men: Apocalypse
The Wolverine
X-Men: Days of Future Past
Iron Man 3
X-Men Origins: Wolverine
Fantastic Four
Iron Man 2
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js