In [2]:

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

In [3]:

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

In [4]:

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

In [5]:

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

In [6]:

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

In [7]:

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

Out[7]:

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

In [8]:

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

In [9]:

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

Out[9]:

```
0
```

In [10]:

```
movies.iloc[0].genres
```

Out[10]:

```
'[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 1
4, "name": "Fantasy"}, {"id": 878, "name": "Science Fiction"}]'
```

In [11]:

```
import ast
```

In [12]:

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

In [13]:

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

In [14]:

```
movies.iloc[0].keywords
```

Out[14]:

```
'[{"id": 1463, "name": "culture clash"}, {"id": 2964, "name": "future"},
{"id": 3386, "name": "space war"}, {"id": 3388, "name": "space colony"},
{"id": 3679, "name": "society"}, {"id": 3801, "name": "space travel"}, {"i
d": 9685, "name": "futuristic"}, {"id": 9840, "name": "romance"}, {"id": 9
882, "name": "space"}, {"id": 9951, "name": "alien"}, {"id": 10148, "nam
e": "tribe"}, {"id": 10158, "name": "alien planet"}, {"id": 10987, "name":
"cgi"}, {"id": 11399, "name": "marine"}, {"id": 13065, "name": "soldier"},
{"id": 14643, "name": "battle"}, {"id": 14720, "name": "love affair"}, {"i
d": 165431, "name": "anti war"}, {"id": 193554, "name": "power relation
s"}, {"id": 206690, "name": "mind and soul"}, {"id": 209714, "name": "3
d"}]'
```

In [15]:

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

In [16]:

```
movies.head()
```

Out[16]:

	movie_id	title	overview	genres	keywords	cast	
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", "...	[{"cre "52fe48009251416c750a
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...	[{"cre "52fe4232c3a36847f800
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...	[{"cre "54805967c3a36829b500
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...	[{"cre "52fe4781c3a36847f813
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...	[{"cre "52fe479ac3a36847f813

In [17]:

```
movies.iloc[0].cast
```

```
Out[17]:
```

```
'[{"cast_id": 242, "character": "Jake Sully", "credit_id": "5602a8a7c3a368
5532001c9a", "gender": 2, "id": 65731, "name": "Sam Worthington", "order":
0}, {"cast_id": 3, "character": "Neytiri", "credit_id": "52fe48009251416c7
50ac9cb", "gender": 1, "id": 8691, "name": "Zoe Saldana", "order": 1}, {"c
ast_id": 25, "character": "Dr. Grace Augustine", "credit_id": "52fe4800925
1416c750aca39", "gender": 1, "id": 10205, "name": "Sigourney Weaver", "ord
er": 2}, {"cast_id": 4, "character": "Col. Quaritch", "credit_id": "52fe48
009251416c750ac9cf", "gender": 2, "id": 32747, "name": "Stephen Lang", "or
der": 3}, {"cast_id": 5, "character": "Trudy Chacon", "credit_id": "52fe48
009251416c750ac9d3", "gender": 1, "id": 17647, "name": "Michelle Rodrigue
z", "order": 4}, {"cast_id": 8, "character": "Selfridge", "credit_id": "52
fe48009251416c750ac9e1", "gender": 2, "id": 1771, "name": "Giovanni Ribis
i", "order": 5}, {"cast_id": 7, "character": "Norm Spellman", "credit_id":
"52fe48009251416c750ac9dd", "gender": 2, "id": 59231, "name": "Joel David
Moore", "order": 6}, {"cast_id": 9, "character": "Moat", "credit_id": "52f
e48009251416c750ac9e5", "gender": 1, "id": 30485, "name": "CCH Pounder",
"order": 7}, {"cast_id": 11, "character": "Eytukan", "credit_id": "52fe480
09251416c750ac9ed", "gender": 2, "id": 15853, "name": "Wes Studi", "orde
r": 8}, {"cast_id": 10, "character": "Tsu\'Tey", "credit_id": "52fe4800925
1416c750ac9e9", "gender": 2, "id": 10964, "name": "Laz Alonso", "order":
9}, {"cast_id": 12, "character": "Dr. Max Patel", "credit_id": "52fe480092
51416c750ac9f1", "gender": 2, "id": 95697, "name": "Dileep Rao", "order":
10}, {"cast_id": 13, "character": "Lyle Wainfleet", "credit_id": "52fe4800
9251416c750ac9f5", "gender": 2, "id": 98215, "name": "Matt Gerald", "orde
r": 11}, {"cast_id": 32, "character": "Private Fike", "credit_id": "52fe48
009251416c750aca5b", "gender": 2, "id": 154153, "name": "Sean Anthony Mora
n", "order": 12}, {"cast_id": 33, "character": "Cryo Vault Med Tech", "cre
dit_id": "52fe48009251416c750aca5f", "gender": 2, "id": 397312, "name": "J
ason Whyte", "order": 13}, {"cast_id": 34, "character": "Venture Star Crew
Chief", "credit_id": "52fe48009251416c750aca63", "gender": 2, "id": 42317,
"name": "Scott Lawrence", "order": 14}, {"cast_id": 35, "character": "Lock
Up Trooper", "credit_id": "52fe48009251416c750aca67", "gender": 2, "id": 9
86734, "name": "Kelly Kilgour", "order": 15}, {"cast_id": 36, "character":
"Shuttle Pilot", "credit_id": "52fe48009251416c750aca6b", "gender": 0, "i
d": 1207227, "name": "James Patrick Pitt", "order": 16}, {"cast_id": 37,
"character": "Shuttle Co-Pilot", "credit_id": "52fe48009251416c750aca6f",
"gender": 0, "id": 1180936, "name": "Sean Patrick Murphy", "order": 17},
{"cast_id": 38, "character": "Shuttle Crew Chief", "credit_id": "52fe48009
251416c750aca73", "gender": 2, "id": 1019578, "name": "Peter Dillon", "ord
er": 18}, {"cast_id": 39, "character": "Tractor Operator / Troupe", "credi
t_id": "52fe48009251416c750aca77", "gender": 0, "id": 91443, "name": "Kevi
n Dorman", "order": 19}, {"cast_id": 40, "character": "Dragon Gunship Pilo
t", "credit_id": "52fe48009251416c750aca7b", "gender": 2, "id": 173391, "n
ame": "Kelson Henderson", "order": 20}, {"cast_id": 41, "character": "Drag
on Gunship Gunner", "credit_id": "52fe48009251416c750aca7f", "gender": 0,
"id": 1207236, "name": "David Van Horn", "order": 21}, {"cast_id": 42, "ch
aracter": "Dragon Gunship Navigator", "credit_id": "52fe48009251416c750aca
83", "gender": 0, "id": 215913, "name": "Jacob Tomuri", "order": 22}, {"ca
st_id": 43, "character": "Suit #1", "credit_id": "52fe48009251416c750aca8
7", "gender": 0, "id": 143206, "name": "Michael Blain-Rozgay", "order": 2
3}, {"cast_id": 44, "character": "Suit #2", "credit_id": "52fe48009251416c
750aca8b", "gender": 2, "id": 169676, "name": "Jon Curry", "order": 24},
{"cast_id": 46, "character": "Ambient Room Tech", "credit_id": "52fe480092
51416c750aca8f", "gender": 0, "id": 1048610, "name": "Luke Hawker", "orde
r": 25}, {"cast_id": 47, "character": "Ambient Room Tech / Troupe", "credi
t_id": "52fe48009251416c750aca93", "gender": 0, "id": 42288, "name": "Wood
y Schultz", "order": 26}, {"cast_id": 48, "character": "Horse Clan Leade
r", "credit_id": "52fe48009251416c750aca97", "gender": 2, "id": 68278, "na
me": "Peter Mensah", "order": 27}, {"cast_id": 49, "character": "Link Room
Tech", "credit_id": "52fe48009251416c750aca9b", "gender": 0, "id": 120724
7, "name": "Sonia Yee", "order": 28}, {"cast_id": 50, "character": "Basket
```

```
ball Avatar / Troupe", "credit_id": "52fe48009251416c750aca9f", "gender":
1, "id": 1207248, "name": "Jahnel Curfman", "order": 29}, {"cast_id": 51,
"character": "Basketball Avatar", "credit_id": "52fe48009251416c750acaa3",
"gender": 0, "id": 89714, "name": "Ilram Choi", "order": 30}, {"cast_id":
52, "character": "Na\'vi Child", "credit_id": "52fe48009251416c750acaa7",
"gender": 0, "id": 1207249, "name": "Kyla Warren", "order": 31}, {"cast_i
d": 53, "character": "Troupe", "credit_id": "52fe48009251416c750acaab", "g
ender": 0, "id": 1207250, "name": "Lisa Roumain", "order": 32}, {"cast_i
d": 54, "character": "Troupe", "credit_id": "52fe48009251416c750acaaf", "g
ender": 1, "id": 83105, "name": "Debra Wilson", "order": 33}, {"cast_id":
57, "character": "Troupe", "credit_id": "52fe48009251416c750acabb", "gende
r": 0, "id": 1207253, "name": "Chris Mala", "order": 34}, {"cast_id": 55,
"character": "Troupe", "credit_id": "52fe48009251416c750acab3", "gender":
0, "id": 1207251, "name": "Taylor Kibby", "order": 35}, {"cast_id": 56, "c
haracter": "Troupe", "credit_id": "52fe48009251416c750acab7", "gender": 0,
"id": 1207252, "name": "Jodie Landau", "order": 36}, {"cast_id": 58, "char
acter": "Troupe", "credit_id": "52fe48009251416c750acabf", "gender": 0, "i
d": 1207254, "name": "Julie Lamm", "order": 37}, {"cast_id": 59, "characte
r": "Troupe", "credit_id": "52fe48009251416c750acac3", "gender": 0, "id":
1207257, "name": "Cullen B. Madden", "order": 38}, {"cast_id": 60, "charac
ter": "Troupe", "credit_id": "52fe48009251416c750acac7", "gender": 0, "i
d": 1207259, "name": "Joseph Brady Madden", "order": 39}, {"cast_id": 61,
"character": "Troupe", "credit_id": "52fe48009251416c750acacb", "gender":
0, "id": 1207262, "name": "Frankie Torres", "order": 40}, {"cast_id": 62,
"character": "Troupe", "credit_id": "52fe48009251416c750acacf", "gender":
1, "id": 1158600, "name": "Austin Wilson", "order": 41}, {"cast_id": 63,
"character": "Troupe", "credit_id": "52fe48019251416c750acad3", "gender":
1, "id": 983705, "name": "Sara Wilson", "order": 42}, {"cast_id": 64, "cha
racter": "Troupe", "credit_id": "52fe48019251416c750acad7", "gender": 0,
"id": 1207263, "name": "Tamica Washington-Miller", "order": 43}, {"cast_i
d": 65, "character": "Op Center Staff", "credit_id": "52fe48019251416c750a
cadb", "gender": 1, "id": 1145098, "name": "Lucy Briant", "order": 44},
{"cast_id": 66, "character": "Op Center Staff", "credit_id": "52fe48019251
416c750acadf", "gender": 2, "id": 33305, "name": "Nathan Meister", "orde
r": 45}, {"cast_id": 67, "character": "Op Center Staff", "credit_id": "52f
e48019251416c750acae3", "gender": 0, "id": 1207264, "name": "Gerry Blair",
"order": 46}, {"cast_id": 68, "character": "Op Center Staff", "credit_id":
"52fe48019251416c750acae7", "gender": 2, "id": 33311, "name": "Matthew Cha
mberlain", "order": 47}, {"cast_id": 69, "character": "Op Center Staff",
"credit_id": "52fe48019251416c750acaeb", "gender": 0, "id": 1207265, "nam
e": "Paul Yates", "order": 48}, {"cast_id": 70, "character": "Op Center Du
ty Officer", "credit_id": "52fe48019251416c750acaef", "gender": 0, "id": 1
207266, "name": "Wray Wilson", "order": 49}, {"cast_id": 71, "character":
"Op Center Staff", "credit_id": "52fe48019251416c750acaf3", "gender": 2,
"id": 54492, "name": "James Gaylyn", "order": 50}, {"cast_id": 72, "charac
ter": "Dancer", "credit_id": "52fe48019251416c750acaf7", "gender": 0, "i
d": 1207267, "name": "Melvin Leno Clark III", "order": 51}, {"cast_id": 7
3, "character": "Dancer", "credit_id": "52fe48019251416c750acafb", "gende
r": 0, "id": 1207268, "name": "Carvon Futrell", "order": 52}, {"cast_id":
74, "character": "Dancer", "credit_id": "52fe48019251416c750acaff", "gende
r": 0, "id": 1207269, "name": "Brandon Jelkes", "order": 53}, {"cast_id":
75, "character": "Dancer", "credit_id": "52fe48019251416c750acb03", "gende
r": 0, "id": 1207270, "name": "Micah Moch", "order": 54}, {"cast_id": 76,
"character": "Dancer", "credit_id": "52fe48019251416c750acb07", "gender":
0, "id": 1207271, "name": "Hanniyah Muhammad", "order": 55}, {"cast_id": 7
7, "character": "Dancer", "credit_id": "52fe48019251416c750acb0b", "gende
r": 0, "id": 1207272, "name": "Christopher Nolen", "order": 56}, {"cast_i
d": 78, "character": "Dancer", "credit_id": "52fe48019251416c750acb0f", "g
ender": 0, "id": 1207273, "name": "Christa Oliver", "order": 57}, {"cast_i
d": 79, "character": "Dancer", "credit_id": "52fe48019251416c750acb13", "g
ender": 0, "id": 1207274, "name": "April Marie Thomas", "order": 58}, {"ca
```

```
st_id": 80, "character": "Dancer", "credit_id": "52fe48019251416c750acb1
7", "gender": 0, "id": 1207275, "name": "Bravita A. Threatt", "order": 5
9}, {"cast_id": 81, "character": "Mining Chief (uncredited)", "credit_id":
"52fe48019251416c750acb1b", "gender": 0, "id": 1207276, "name": "Colin Ble
asdale", "order": 60}, {"cast_id": 82, "character": "Veteran Miner (uncred
ited)", "credit_id": "52fe48019251416c750acb1f", "gender": 0, "id": 10796
9, "name": "Mike Bodnar", "order": 61}, {"cast_id": 83, "character": "Rich
ard (uncredited)", "credit_id": "52fe48019251416c750acb23", "gender": 0,
"id": 1207278, "name": "Matt Clayton", "order": 62}, {"cast_id": 84, "char
acter": "Nav\'i (uncredited)", "credit_id": "52fe48019251416c750acb27", "g
ender": 1, "id": 147898, "name": "Nicole Dionne", "order": 63}, {"cast_i
d": 85, "character": "Trooper (uncredited)", "credit_id": "52fe48019251416
c750acb2b", "gender": 0, "id": 1207280, "name": "Jamie Harrison", "order":
64}, {"cast_id": 86, "character": "Trooper (uncredited)", "credit_id": "52
fe48019251416c750acb2f", "gender": 0, "id": 1207281, "name": "Allan Henr
y", "order": 65}, {"cast_id": 87, "character": "Ground Technician (uncredi
ted)", "credit_id": "52fe48019251416c750acb33", "gender": 2, "id": 120728
2, "name": "Anthony Ingruber", "order": 66}, {"cast_id": 88, "character":
"Flight Crew Mechanic (uncredited)", "credit_id": "52fe48019251416c750acb3
7", "gender": 0, "id": 1207283, "name": "Ashley Jeffery", "order": 67},
{"cast_id": 14, "character": "Samson Pilot", "credit_id": "52fe48009251416
c750ac9f9", "gender": 0, "id": 98216, "name": "Dean Knowsley", "order": 6
8}, {"cast_id": 89, "character": "Trooper (uncredited)", "credit_id": "52f
e48019251416c750acb3b", "gender": 0, "id": 1201399, "name": "Joseph Mika-H
unt", "order": 69}, {"cast_id": 90, "character": "Banshee (uncredited)",
"credit_id": "52fe48019251416c750acb3f", "gender": 0, "id": 236696, "nam
e": "Terry Notary", "order": 70}, {"cast_id": 91, "character": "Soldier (u
ncredited)", "credit_id": "52fe48019251416c750acb43", "gender": 0, "id": 1
207287, "name": "Kai Pantano", "order": 71}, {"cast_id": 92, "character":
"Blast Technician (uncredited)", "credit_id": "52fe48019251416c750acb47",
"gender": 0, "id": 1207288, "name": "Logan Pithyou", "order": 72}, {"cast_
id": 93, "character": "Vindum Raah (uncredited)", "credit_id": "52fe480192
51416c750acb4b", "gender": 0, "id": 1207289, "name": "Stuart Pollock", "or
der": 73}, {"cast_id": 94, "character": "Hero (uncredited)", "credit_id":
"52fe48019251416c750acb4f", "gender": 0, "id": 584868, "name": "Raja", "or
"credit_id": "52fe48019251416c750acb53", "gender": 0, "id": 1207290, "nam
e": "Gareth Ruck", "order": 75}, {"cast_id": 96, "character": "Engineer (u
ncredited)", "credit_id": "52fe48019251416c750acb57", "gender": 0, "id": 1
062463, "name": "Rhian Sheehan", "order": 76}, {"cast_id": 97, "characte
r": "Col. Quaritch\'s Mech Suit (uncredited)", "credit_id": "52fe480192514
16c750acb5b", "gender": 0, "id": 60656, "name": "T. J. Storm", "order": 7
7}, {"cast_id": 98, "character": "Female Marine (uncredited)", "credit_i
d": "52fe48019251416c750acb5f", "gender": 0, "id": 1207291, "name": "Jodie
Taylor", "order": 78}, {"cast_id": 99, "character": "Ikran Clan Leader (un
credited)", "credit_id": "52fe48019251416c750acb63", "gender": 1, "id": 11
86027, "name": "Alicia Vela-Bailey", "order": 79}, {"cast_id": 100, "chara
cter": "Geologist (uncredited)", "credit_id": "52fe48019251416c750acb67",
"gender": 0, "id": 1207292, "name": "Richard Whiteside", "order": 80}, {"c
ast_id": 101, "character": "Na\'vi (uncredited)", "credit_id": "52fe480192
51416c750acb6b", "gender": 0, "id": 103259, "name": "Nikie Zambo", "orde
r": 81}, {"cast_id": 102, "character": "Ambient Room Tech / Troupe", "cred
it_id": "52fe48019251416c750acb6f", "gender": 1, "id": 42286, "name": "Jul
ene Renee", "order": 82}]
```

In [18]:

```
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
```

In [19]:

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

In [20]:

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

Out[20]:

	movie_id	title	overview	genres	keywords	cast	cre
0	19995	Avatar	In the 22nd century, a paraplegic Marine is...	[Action, Adventure, Fantasy, Science Fiction]	[culture clash, future, space war, space colon...	[Sam Worthington, Zoe Saldana, Sigourney Weaver]	[{"credit_id": "52fe48009251416c750aca23", "de

In [21]:

```
movies['crew'][0]
```

Out[21]:

```
'[{"credit_id": "52fe48009251416c750aca23", "department": "Editing", "g
ender": 0, "id": 1721, "job": "Editor", "name": "Stephen E. Rivkin"},
{"credit_id": "539c47ecc3a36810e3001f87", "department": "Art", "gende
r": 2, "id": 496, "job": "Production Design", "name": "Rick Carter"},
{"credit_id": "54491c89c3a3680fb4001cf7", "department": "Sound", "gende
r": 0, "id": 900, "job": "Sound Designer", "name": "Christopher Boye
s"}, {"credit_id": "54491cb70e0a267480001bd0", "department": "Sound",
"gender": 0, "id": 900, "job": "Supervising Sound Editor", "name": "Chr
istopher Boyes"}, {"credit_id": "539c4a4cc3a36810c9002101", "departmen
t": "Production", "gender": 1, "id": 1262, "job": "Casting", "name": "M
ali Finn"}, {"credit_id": "5544ee3b925141499f0008fc", "department": "So
und", "gender": 2, "id": 1729, "job": "Original Music Composer", "nam
e": "James Horner"}, {"credit_id": "52fe48009251416c750ac9c3", "departm
ent": "Directing", "gender": 2, "id": 2710, "job": "Director", "name":
"James Cameron"}, {"credit_id": "52fe48009251416c750ac9d9", "departmen
t": "Writing", "gender": 2, "id": 2710, "job": "Writer", "name": "James
Cameron"}, {"credit_id": "52fe48009251416c750aca17", "department": "Edi
ting", "gender": 2, "id": 2710, "job": "Editor", "name": "James Camero
```

In [22]:

```
def fetch_director(obj):
    L=[]
    counter=0;
    for i in ast.literal_eval(obj):
        if  i['job']=='Director':
            L.append(i['name'])
            break;

    return L
```

In [23]:

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

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]	[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]

In [25]:

```
movies['overview'][0]
```

Out[25]:

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

In [26]:

```
#convert to list
movies['overview']=movies['overview'].apply(lambda x:x.split())
```

In [27]:

```
# all the 4 cols are converted to tags ,
movies.head()
```

Out[27]:

	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 [28]:

```
# removes spaces in 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 [29]:

```
movies.head()
```

Out[29]:

	movie_id	title	overview	genres	keywords	cast	
0	19995	Avatar	[In, the, 22nd, century,, a, paraplegic, Marin...	[Action, Adventure, Fantasy, ScienceFiction]	[cultureclash, future, spacewar, spacecolony, ...	[SamWorthington, ZoeSaldana, SigourneyWeaver]	[James
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]	[Gore
2	206647	Spectre	[A, cryptic, message, from, Bond's, past, send...	[Action, Adventure, Crime]	[spy, basedonnovel, secretagent, sequel, mi6, ...	[DanielCraig, ChristophWaltz, LéaSeydoux]	[San
3	49026	The Dark Knight Rises	[Following, the, death, of, District, Attorney...	[Action, Crime, Drama, Thriller]	[dccomics, crimefighter, terrorist, secretiden...	[ChristianBale, MichaelCaine, GaryOldman]	[Christop
4	49529	John Carter	[John, Carter, is, a, war-weary,, former, mili...	[Action, Adventure, ScienceFiction]	[basedonnovel, mars, medallion, spacetravel, p...	[TaylorKitsch, LynnCollins, SamanthaMorton]	[Andre

In [30]:

```
#creating a new taggs col
movies['tags']=movies['overview']+movies['genres']+movies['keywords']+movies['cast']+mov
```

In [31]:

```
movies.head()
```

Out[31]:

	movie_id	title	overview	genres	keywords	cast	
0	19995	Avatar	[In, the, 22nd, century,, a, paraplegic, Marin...	[Action, Adventure, Fantasy, ScienceFiction]	[cultureclash, future, spacewar, spacecolony, ...	[SamWorthington, ZoeSaldana, SigourneyWeaver]	[James
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]	[Gore
2	206647	Spectre	[A, cryptic, message, from, Bond's, past, send...	[Action, Adventure, Crime]	[spy, basedonnovel, secretagent, sequel, mi6, ...	[DanielCraig, ChristophWaltz, LéaSeydoux]	[San
3	49026	The Dark Knight Rises	[Following, the, death, of, District, Attorney...	[Action, Crime, Drama, Thriller]	[dccomics, crimefighter, terrorist, secretiden...	[ChristianBale, MichaelCaine, GaryOldman]	[Christop
4	49529	John Carter	[John, Carter, is, a, war-weary,, former, mili...	[Action, Adventure, ScienceFiction]	[basedonnovel, mars, medallion, spacetravel, p...	[TaylorKitsch, LynnCollins, SamanthaMorton]	[Andre

In [32]:

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

In [33]:

```
# convert list into string with space
new_df['tags'] =new_df['tags'].apply(lambda x: " ".join(x))
```

```
C:\Users\sharv\AppData\Local\Temp\ipykernel_14220\4012028172.py:2: Setting
WithCopyWarning:
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-doc
s/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: " ".join(x))
```

In [34]:

```
new_df.head()
```

Out[34]:

	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...

In [35]:

```
# convert all case to lower
new_df['tags']=new_df['tags'].apply(lambda x: x.lower())
```

```
C:\Users\sharv\AppData\Local\Temp\ipykernel_14220\4008985432.py:2: Setting
WithCopyWarning:
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-doc
s/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())
```

In [36]:

```
new_df.head(1)
```

Out[36]:

	movie_id	title	tags
0	19995	Avatar	in the 22nd century, a paraplegic marine is di...

In [37]:

```
new_df.head()
```

Out[37]:

	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...

In [38]:

```
import nltk
```

In [39]:

```
#removing common words like loved,l0ve, loving
from nltk.stem.porter import PorterStemmer
ps=PorterStemmer()
```

In [40]:

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

In [41]:

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

```
C:\Users\sharv\AppData\Local\Temp\ipykernel_14220\3514595201.py:1: Setting
WithCopyWarning:
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-doc
s/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(stem)
```

In [42]:

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

In [43]:

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

In [44]:

```
vectors[0]
```

Out[44]:

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

In [45]:

```
# most common 5000 words
cv.get_feature_names_out()
```

Out[45]:

```
array(['000', '007', '10', ..., 'zone', 'zoo', 'zooeydeschanel'],
      dtype=object)
```

In [46]:

```
# dis 1 movies to another movies, consine(angle) distance not euclidiead distance, of mo
from sklearn.metrics.pairwise import cosine_similarity
```

In [47]:

```
similarity = cosine_similarity(vectors)
```

In [48]:

```
similarity[0]
```

Out[48]:

```
array([1.        , 0.08346223, 0.0860309 , ..., 0.04499213, 0.        ,
       0.        ])
```

In [49]:

```
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)
```

In [50]:

```python
recommend('Batman')
```

```
Batman
Batman & Robin
Batman Begins
Batman Returns
The R.M.
```

In [51]:

```python
import pickle
```

In [52]:

```python
pickle.dump(new_df,open('movies.pkl','wb'))
```

In [53]:

```python
pickle.dump(new_df.to_dict(),open('movie_dict.pkl','wb'))
```

In [54]:

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

In [ ]:

```python

```