In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import os
```

In [2]:
```
movie = pd.read_csv(r"C:\Users\srivi\OneDrive\Documents\Projects\TMDB_5000Movies
credits = pd.read_csv(r"C:\Users\srivi\OneDrive\Documents\Projects\TMDB_5000Movi
```

In [3]:
```
movie.head()
```

Out[3]:

	budget	genres	homepage	id	keyword
0	237000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.avatarmovie.com/	19995	[{"id 146 "name "cultur clash" {"id":
1	300000000	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	http://disney.go.com/disneypictures/pirates/	285	[{"id 27("name "ocean" {"id": 72("na
2	245000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.sonypictures.com/movies/spectre/	206647	[{"id 47("name "spy" {"id": 81 "name
3	250000000	[{"id": 28, "name": "Action"}, {"id": 80, "nam...	http://www.thedarkknightrises.com/	49026	[{"id 84 "name "d comics" {"id 853,
4	260000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://movies.disney.com/john-carter	49529	[{"id 81 "name "based o novel" {"id":

In [4]:
```
credits.head()
```

Out[4]:

	movie_id	title	cast	crew
0	19995	Avatar	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
1	285	Pirates of the Caribbean: At World's End	[{"cast_id": 4, "character": "Captain Jack Spa...	[{"credit_id": "52fe4232c3a36847f800b579", "de...
2	206647	Spectre	[{"cast_id": 1, "character": "James Bond", "cr...	[{"credit_id": "54805967c3a36829b5002c41", "de...
3	49026	The Dark Knight Rises	[{"cast_id": 2, "character": "Bruce Wayne / Ba...	[{"credit_id": "52fe4781c3a36847f81398c3", "de...
4	49529	John Carter	[{"cast_id": 5, "character": "John Carter", "c...	[{"credit_id": "52fe479ac3a36847f813eaa3", "de...

In [5]:
```
movie.shape
```

Out[5]:
```
(4803, 20)
```

In [6]:
```
credits.shape
```

Out[6]:
```
(4803, 4)
```

In [7]:
```
#I want to merge the two datasets
#so I only have to work with one dataset and all variables together
movies= movie.merge(credits, on='title')
movies.head()
```

Out[7]:

	budget	genres	homepage	id	keyword
0	237000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.avatarmovie.com/	19995	[{"id 146 "name "cultur clash" {"id":
1	300000000	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	http://disney.go.com/disneypictures/pirates/	285	[{"id 27("name "ocean" {"id": 72("na
2	245000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.sonypictures.com/movies/spectre/	206647	[{"id 47("name "spy" {"id": 818 "name
3	250000000	[{"id": 28, "name": "Action"}, {"id": 80, "nam...	http://www.thedarkknightrises.com/	49026	[{"id 849 "name "d comics" {"id 853,
4	260000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://movies.disney.com/john-carter	49529	[{"id 818 "name "based o novel" {"id":

5 rows × 23 columns

In [8]:
```
movies.shape
```

Out[8]:
```
(4809, 23)
```

In [9]:
```
#I want to remove the columns that are insignificant in this system
#To do that I want a list of all the columns so I can pick and choose
movies.columns
```

Out[9]:
```
Index(['budget', 'genres', 'homepage', 'id', 'keywords', 'original_language',
       'original_title', 'overview', 'popularity', 'production_companies',
       'production_countries', 'release_date', 'revenue', 'runtime',
       'spoken_languages', 'status', 'tagline', 'title', 'vote_average',
       'vote_count', 'movie_id', 'cast', 'crew'],
      dtype='object')
```

In [10]:
```
movies= movies[['movie_id', 'title', 'overview', 'genres',
                'keywords', 'cast', 'crew']]
movies.head()
#I have filtered the dataset so I only have the relavant columns
```

Out[10]:

	movie_id	title	overview	genres	keywords	cast	
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", "...	"52fe480092514
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...	"52fe4232c3a36
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...	"54805967c3a36
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...	"52fe4781c3a36
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...	"52fe479ac3a36

In [11]:
```
movies.shape
```

Out[11]:
```
(4809, 7)
```

In [12]:
```
movies.isnull().sum()
#now i am checking for any missing values in the relavant columns only
```

Out[12]:
```
movie_id    0
title       0
overview    3
genres      0
keywords    0
cast        0
crew        0
dtype: int64
```

In [13]:
```
movies.dropna(inplace=True)
#i am dropping all the null values
```

In [14]:
```
movies.isnull().sum()
```

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

In [15]:
```
movies.duplicated().sum()
#checking to see if there are any duplicate values in the DF
```

Out[15]:
```
0
```

In [16]:
```
movies.iloc[0]['genres']
#the genres column is presented as a list
```

Out[16]:
```
'[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, "na
me": "Fantasy"}, {"id": 878, "name": "Science Fiction"}]'
```

In [17]:
```
movies.dtypes
```

Out[17]:
```
movie_id     int64
title       object
overview    object
genres      object
keywords    object
cast        object
crew        object
dtype: object
```

In [18]:
```
import ast
#helps convert a string to list
```

In [19]:
```
def convert(text):
    # appends the strings in the columns into a list
    l = []  # initialises an empty list

    for i in ast.literal_eval(text):  # converts the string to a list
        l.append(i['name'])  # adds the 'name' field to the list

    return l  # returns the list
```

In [20]:
```
movies['genres'] = movies['genres'].apply(convert)
```

```
# i applied the function i made to the genre column
```

In [21]:
```
movies.head()
#to see the changes that have been made
```

Out[21]:

	movie_id	title	overview	genres	keywords	cast	
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", "...	"52fe4800925141
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...	"52fe4232c3a368
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...	"54805967c3a368
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...	"52fe4781c3a368
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...	"52fe479ac3a368

In [22]:
```
movies.iloc[0]['keywords']
#checking to see how the keywords column looks
```

Out[22]:

```
'[{"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"}, {"id": 9685, "name":
"futuristic"}, {"id": 9840, "name": "romance"}, {"id": 9882, "name": "space"},
{"id": 9951, "name": "alien"}, {"id": 10148, "name": "tribe"}, {"id": 10158, "n
ame": "alien planet"}, {"id": 10987, "name": "cgi"}, {"id": 11399, "name": "mar
ine"}, {"id": 13065, "name": "soldier"}, {"id": 14643, "name": "battle"}, {"i
d": 14720, "name": "love affair"}, {"id": 165431, "name": "anti war"}, {"id": 1
93554, "name": "power relations"}, {"id": 206690, "name": "mind and soul"}, {"i
d": 209714, "name": "3d"}]'
```

In [23]:
```
movies['keywords'] = movies['keywords'].apply(convert)
#applied the function to the keywords function too
```

In [24]:
```
movies.head()
#to see the changes that have been made
```

Out[24]:

	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", "...	"52fe4800925141
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...	"52fe4232c3a368
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...	"54805967c3a368
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...	"52fe4781c3a368
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...	"52fe479ac3a368

In [25]:
```
movies.iloc[0]['cast']
#checking to see how the cast column looks
```

Out[25]:

```
'[{"cast_id": 242, "character": "Jake Sully", "credit_id": "5602a8a7c3a36855320
01c9a", "gender": 2, "id": 65731, "name": "Sam Worthington", "order": 0}, {"cas
t_id": 3, "character": "Neytiri", "credit_id": "52fe48009251416c750ac9cb", "gen
der": 1, "id": 8691, "name": "Zoe Saldana", "order": 1}, {"cast_id": 25, "chara
cter": "Dr. Grace Augustine", "credit_id": "52fe48009251416c750aca39", "gende
r": 1, "id": 10205, "name": "Sigourney Weaver", "order": 2}, {"cast_id": 4, "ch
aracter": "Col. Quaritch", "credit_id": "52fe48009251416c750ac9cf", "gender":
2, "id": 32747, "name": "Stephen Lang", "order": 3}, {"cast_id": 5, "characte
r": "Trudy Chacon", "credit_id": "52fe48009251416c750ac9d3", "gender": 1, "id":
17647, "name": "Michelle Rodriguez", "order": 4}, {"cast_id": 8, "character":
"Selfridge", "credit_id": "52fe48009251416c750ac9e1", "gender": 2, "id": 1771,
"name": "Giovanni Ribisi", "order": 5}, {"cast_id": 7, "character": "Norm Spell
man", "credit_id": "52fe48009251416c750ac9dd", "gender": 2, "id": 59231, "nam
e": "Joel David Moore", "order": 6}, {"cast_id": 9, "character": "Moat", "credi
t_id": "52fe48009251416c750ac9e5", "gender": 1, "id": 30485, "name": "CCH Pound
er", "order": 7}, {"cast_id": 11, "character": "Eytukan", "credit_id": "52fe480
09251416c750ac9ed", "gender": 2, "id": 15853, "name": "Wes Studi", "order": 8},
{"cast_id": 10, "character": "Tsu\'Tey", "credit_id": "52fe48009251416c750ac9e
9", "gender": 2, "id": 10964, "name": "Laz Alonso", "order": 9}, {"cast_id": 1
2, "character": "Dr. Max Patel", "credit_id": "52fe48009251416c750ac9f1", "gend
er": 2, "id": 95697, "name": "Dileep Rao", "order": 10}, {"cast_id": 13, "chara
cter": "Lyle Wainfleet", "credit_id": "52fe48009251416c750ac9f5", "gender": 2,
"id": 98215, "name": "Matt Gerald", "order": 11}, {"cast_id": 32, "character":
"Private Fike", "credit_id": "52fe48009251416c750aca5b", "gender": 2, "id": 154
153, "name": "Sean Anthony Moran", "order": 12}, {"cast_id": 33, "character":
"Cryo Vault Med Tech", "credit_id": "52fe48009251416c750aca5f", "gender": 2, "i
d": 397312, "name": "Jason 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, "charac
ter": "Lock Up Trooper", "credit_id": "52fe48009251416c750aca67", "gender": 2,
"id": 986734, "name": "Kelly Kilgour", "order": 15}, {"cast_id": 36, "characte
r": "Shuttle Pilot", "credit_id": "52fe48009251416c750aca6b", "gender": 0, "i
d": 1207227, "name": "James Patrick Pitt", "order": 16}, {"cast_id": 37, "chara
cter": "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": "52fe48009251416c750aca73", "ge
nder": 2, "id": 1019578, "name": "Peter Dillon", "order": 18}, {"cast_id": 39,
"character": "Tractor Operator / Troupe", "credit_id": "52fe48009251416c750aca7
7", "gender": 0, "id": 91443, "name": "Kevin Dorman", "order": 19}, {"cast_id":
40, "character": "Dragon Gunship Pilot", "credit_id": "52fe48009251416c750aca7
b", "gender": 2, "id": 173391, "name": "Kelson Henderson", "order": 20}, {"cast
_id": 41, "character": "Dragon Gunship Gunner", "credit_id": "52fe48009251416c7
50aca7f", "gender": 0, "id": 1207236, "name": "David Van Horn", "order": 21},
{"cast_id": 42, "character": "Dragon Gunship Navigator", "credit_id": "52fe4800
9251416c750aca83", "gender": 0, "id": 215913, "name": "Jacob Tomuri", "order":
22}, {"cast_id": 43, "character": "Suit #1", "credit_id": "52fe48009251416c750a
ca87", "gender": 0, "id": 143206, "name": "Michael Blain-Rozgay", "order": 23},
{"cast_id": 44, "character": "Suit #2", "credit_id": "52fe48009251416c750aca8
b", "gender": 2, "id": 169676, "name": "Jon Curry", "order": 24}, {"cast_id": 4
6, "character": "Ambient Room Tech", "credit_id": "52fe48009251416c750aca8f",
"gender": 0, "id": 1048610, "name": "Luke Hawker", "order": 25}, {"cast_id": 4
7, "character": "Ambient Room Tech / Troupe", "credit_id": "52fe48009251416c750
aca93", "gender": 0, "id": 42288, "name": "Woody Schultz", "order": 26}, {"cast
_id": 48, "character": "Horse Clan Leader", "credit_id": "52fe48009251416c750ac
a97", "gender": 2, "id": 68278, "name": "Peter Mensah", "order": 27}, {"cast_i
d": 49, "character": "Link Room Tech", "credit_id": "52fe48009251416c750aca9b",
"gender": 0, "id": 1207247, "name": "Sonia Yee", "order": 28}, {"cast_id": 50,
"character": "Basketball Avatar / Troupe", "credit_id": "52fe48009251416c750aca
9f", "gender": 1, "id": 1207248, "name": "Jahnel Curfman", "order": 29}, {"cast
_id": 51, "character": "Basketball Avatar", "credit_id": "52fe48009251416c750ac
```

aa3", "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_id": 53, "character": "Troupe", "credit_id": "52fe48009251416c750acaab", "gender": 0, "id": 1207250, "name": "Lisa Roumain", "order": 32}, {"cast_id": 54, "character": "Troupe", "credit_id": "52fe48009251416c750acaaf", "gender": 1, "id": 83105, "name": "Debra Wilson", "order": 33}, {"cast_id": 57, "character": "Troupe", "credit_id": "52fe48009251416c750acabb", "gender": 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, "character": "Troupe", "credit_id": "52fe48009251416c750acab7", "gender": 0, "id": 1207252, "name": "Jodie Landau", "order": 36}, {"cast_id": 58, "character": "Troupe", "credit_id": "52fe48009251416c750acabf", "gender": 0, "id": 1207254, "name": "Julie Lamm", "order": 37}, {"cast_id": 59, "character": "Troupe", "credit_id": "52fe48009251416c750acac3", "gender": 0, "id": 1207257, "name": "Cullen B. Madden", "order": 38}, {"cast_id": 60, "character": "Troupe", "credit_id": "52fe48009251416c750acac7", "gender": 0, "id": 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, "character": "Troupe", "credit_id": "52fe48019251416c750acad7", "gender": 0, "id": 1207263, "name": "Tamica Washington-Miller", "order": 43}, {"cast_id": 65, "character": "Op Center Staff", "credit_id": "52fe48019251416c750acadb", "gender": 1, "id": 1145098, "name": "Lucy Briant", "order": 44}, {"cast_id": 66, "character": "Op Center Staff", "credit_id": "52fe48019251416c750acadf", "gender": 2, "id": 33305, "name": "Nathan Meister", "order": 45}, {"cast_id": 67, "character": "Op Center Staff", "credit_id": "52fe48019251416c750acae3", "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 Chamberlain", "order": 47}, {"cast_id": 69, "character": "Op Center Staff", "credit_id": "52fe48019251416c750acaeb", "gender": 0, "id": 1207265, "name": "Paul Yates", "order": 48}, {"cast_id": 70, "character": "Op Center Duty Officer", "credit_id": "52fe48019251416c750acaef", "gender": 0, "id": 1207266, "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, "character": "Dancer", "credit_id": "52fe48019251416c750acaf7", "gender": 0, "id": 1207267, "name": "Melvin Leno Clark III", "order": 51}, {"cast_id": 73, "character": "Dancer", "credit_id": "52fe48019251416c750acafb", "gender": 0, "id": 1207268, "name": "Carvon Futrell", "order": 52}, {"cast_id": 74, "character": "Dancer", "credit_id": "52fe48019251416c750acaff", "gender": 0, "id": 1207269, "name": "Brandon Jelkes", "order": 53}, {"cast_id": 75, "character": "Dancer", "credit_id": "52fe48019251416c750acb03", "gender": 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": 77, "character": "Dancer", "credit_id": "52fe48019251416c750acb0b", "gender": 0, "id": 1207272, "name": "Christopher Nolen", "order": 56}, {"cast_id": 78, "character": "Dancer", "credit_id": "52fe48019251416c750acb0f", "gender": 0, "id": 1207273, "name": "Christa Oliver", "order": 57}, {"cast_id": 79, "character": "Dancer", "credit_id": "52fe48019251416c750acb13", "gender": 0, "id": 1207274, "name": "April Marie Thomas", "order": 58}, {"cast_id": 80, "character": "Dancer", "credit_id": "52fe48019251416c750acb17", "gender": 0, "id": 1207275, "name": "Bravita A. Threatt", "order": 59}, {"cast_id": 81, "character": "Mining Chief (uncredited)", "credit_id": "52fe48019251416c750acb1b", "gender": 0, "id": 1207276, "name": "Colin Bleasdale", "order": 60}, {"cast_id": 82, "character": "Veteran Miner (uncredited)", "credit_id": "52fe48019251416c750acb1f", "gender": 0, "id": 107969, "nam

e": "Mike Bodnar", "order": 61}, {"cast_id": 83, "character": "Richard (uncredi
ted)", "credit_id": "52fe48019251416c750acb23", "gender": 0, "id": 1207278, "na
me": "Matt Clayton", "order": 62}, {"cast_id": 84, "character": "Nav\'i (uncred
ited)", "credit_id": "52fe48019251416c750acb27", "gender": 1, "id": 147898, "na
me": "Nicole Dionne", "order": 63}, {"cast_id": 85, "character": "Trooper (uncr
edited)", "credit_id": "52fe48019251416c750acb2b", "gender": 0, "id": 1207280,
"name": "Jamie Harrison", "order": 64}, {"cast_id": 86, "character": "Trooper
(uncredited)", "credit_id": "52fe48019251416c750acb2f", "gender": 0, "id": 1207
281, "name": "Allan Henry", "order": 65}, {"cast_id": 87, "character": "Ground
Technician (uncredited)", "credit_id": "52fe48019251416c750acb33", "gender": 2,
"id": 1207282, "name": "Anthony Ingruber", "order": 66}, {"cast_id": 88, "chara
cter": "Flight Crew Mechanic (uncredited)", "credit_id": "52fe48019251416c750ac
b37", "gender": 0, "id": 1207283, "name": "Ashley Jeffery", "order": 67}, {"cas
t_id": 14, "character": "Samson Pilot", "credit_id": "52fe48009251416c750ac9f
9", "gender": 0, "id": 98216, "name": "Dean Knowsley", "order": 68}, {"cast_i
d": 89, "character": "Trooper (uncredited)", "credit_id": "52fe48019251416c750a
cb3b", "gender": 0, "id": 1201399, "name": "Joseph Mika-Hunt", "order": 69},
{"cast_id": 90, "character": "Banshee (uncredited)", "credit_id": "52fe48019251
416c750acb3f", "gender": 0, "id": 236696, "name": "Terry Notary", "order": 70},
{"cast_id": 91, "character": "Soldier (uncredited)", "credit_id": "52fe48019251
416c750acb43", "gender": 0, "id": 1207287, "name": "Kai Pantano", "order": 71},
{"cast_id": 92, "character": "Blast Technician (uncredited)", "credit_id": "52f
e48019251416c750acb47", "gender": 0, "id": 1207288, "name": "Logan Pithyou", "o
rder": 72}, {"cast_id": 93, "character": "Vindum Raah (uncredited)", "credit_i
d": "52fe48019251416c750acb4b", "gender": 0, "id": 1207289, "name": "Stuart Pol
lock", "order": 73}, {"cast_id": 94, "character": "Hero (uncredited)", "credit_
id": "52fe48019251416c750acb4f", "gender": 0, "id": 584868, "name": "Raja", "or
der": 74}, {"cast_id": 95, "character": "Ops Centreworker (uncredited)", "credi
t_id": "52fe48019251416c750acb53", "gender": 0, "id": 1207290, "name": "Gareth
Ruck", "order": 75}, {"cast_id": 96, "character": "Engineer (uncredited)", "cre
dit_id": "52fe48019251416c750acb57", "gender": 0, "id": 1062463, "name": "Rhian
Sheehan", "order": 76}, {"cast_id": 97, "character": "Col. Quaritch\'s Mech Sui
t (uncredited)", "credit_id": "52fe48019251416c750acb5b", "gender": 0, "id": 60
656, "name": "T. J. Storm", "order": 77}, {"cast_id": 98, "character": "Female
Marine (uncredited)", "credit_id": "52fe48019251416c750acb5f", "gender": 0, "i
d": 1207291, "name": "Jodie Taylor", "order": 78}, {"cast_id": 99, "character":
"Ikran Clan Leader (uncredited)", "credit_id": "52fe48019251416c750acb63", "gen
der": 1, "id": 1186027, "name": "Alicia Vela-Bailey", "order": 79}, {"cast_id":
100, "character": "Geologist (uncredited)", "credit_id": "52fe48019251416c750ac
b67", "gender": 0, "id": 1207292, "name": "Richard Whiteside", "order": 80},
{"cast_id": 101, "character": "Na\'vi (uncredited)", "credit_id": "52fe48019251
416c750acb6b", "gender": 0, "id": 103259, "name": "Nikie Zambo", "order": 81},
{"cast_id": 102, "character": "Ambient Room Tech / Troupe", "credit_id": "52fe4
8019251416c750acb6f", "gender": 1, "id": 42286, "name": "Julene Renee", "orde
r": 82}]'

In [26]:

```
def convert_cast(text):
    l = []

    for i in ast.literal_eval(text):
        #converts the string to list of dictionaries

        if "uncredited" not in i["character"].lower():
            # filters out uncredited roles

            l.append(i["name"])
            # adds only credited names to the list

    return l
```

In [27]:
```
# apply convert_cast function to the 'cast' column
movies['cast'] = movies['cast'].apply(convert_cast)
```

In [28]:
```
movies.head()
#to see the changes that have been made
```

Out[28]:

	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...	[Sam Worthington, Zoe Saldana, Sigourney Weave...	"52fe480092514
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, ...	"52fe4232c3a36
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, R...	"54805967c3a36
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, A...	"52fe4781c3a36
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,...	"52fe479ac3a36

In [29]:
```
len(movies.iloc[0]['cast'])
#checking to see if the length of the list has reduced to just the credited
```

Out[29]: 62

In [30]:
```
movies.iloc[0]['crew']
#checking to see how the crew column looks
```

Out[30]:

```
'[{"credit_id": "52fe48009251416c750aca23", "department": "Editing", "gender":
0, "id": 1721, "job": "Editor", "name": "Stephen E. Rivkin"}, {"credit_id": "53
9c47ecc3a36810e3001f87", "department": "Art", "gender": 2, "id": 496, "job": "P
roduction Design", "name": "Rick Carter"}, {"credit_id": "54491c89c3a3680fb4001
cf7", "department": "Sound", "gender": 0, "id": 900, "job": "Sound Designer",
"name": "Christopher Boyes"}, {"credit_id": "54491cb70e0a267480001bd0", "depart
ment": "Sound", "gender": 0, "id": 900, "job": "Supervising Sound Editor", "nam
e": "Christopher Boyes"}, {"credit_id": "539c4a4cc3a36810c9002101", "departmen
t": "Production", "gender": 1, "id": 1262, "job": "Casting", "name": "Mali Fin
n"}, {"credit_id": "5544ee3b925141499f0008fc", "department": "Sound", "gender":
2, "id": 1729, "job": "Original Music Composer", "name": "James Horner"}, {"cre
dit_id": "52fe48009251416c750ac9c3", "department": "Directing", "gender": 2, "i
d": 2710, "job": "Director", "name": "James Cameron"}, {"credit_id": "52fe48009
251416c750ac9d9", "department": "Writing", "gender": 2, "id": 2710, "job": "Wri
ter", "name": "James Cameron"}, {"credit_id": "52fe48009251416c750aca17", "depa
rtment": "Editing", "gender": 2, "id": 2710, "job": "Editor", "name": "James Ca
meron"}, {"credit_id": "52fe48009251416c750aca29", "department": "Production",
"gender": 2, "id": 2710, "job": "Producer", "name": "James Cameron"}, {"credit_
id": "52fe48009251416c750aca3f", "department": "Writing", "gender": 2, "id": 27
10, "job": "Screenplay", "name": "James Cameron"}, {"credit_id": "539c4987c3a36
810ba0021a4", "department": "Art", "gender": 2, "id": 7236, "job": "Art Directi
on", "name": "Andrew Menzies"}, {"credit_id": "549598c3c3a3686ae9004383", "depa
rtment": "Visual Effects", "gender": 0, "id": 6690, "job": "Visual Effects Prod
ucer", "name": "Jill Brooks"}, {"credit_id": "52fe48009251416c750aca4b", "depar
tment": "Production", "gender": 1, "id": 6347, "job": "Casting", "name": "Marge
ry Simkin"}, {"credit_id": "570b6f419251417da70032fe", "department": "Art", "ge
nder": 2, "id": 6878, "job": "Supervising Art Director", "name": "Kevin Ishiok
a"}, {"credit_id": "5495a0fac3a3686ae9004468", "department": "Sound", "gender":
0, "id": 6883, "job": "Music Editor", "name": "Dick Bernstein"}, {"credit_id":
"54959706c3a3686af3003e81", "department": "Sound", "gender": 0, "id": 8159, "jo
b": "Sound Effects Editor", "name": "Shannon Mills"}, {"credit_id": "54491d58c3
a3680fb1001ccb", "department": "Sound", "gender": 0, "id": 8160, "job": "Fole
y", "name": "Dennie Thorpe"}, {"credit_id": "54491d6cc3a3680fa5001b2c", "depart
ment": "Sound", "gender": 0, "id": 8163, "job": "Foley", "name": "Jana Vance"},
{"credit_id": "52fe48009251416c750aca57", "department": "Costume & Make-Up", "g
ender": 1, "id": 8527, "job": "Costume Design", "name": "Deborah Lynn Scott"},
{"credit_id": "52fe48009251416c750aca2f", "department": "Production", "gender":
2, "id": 8529, "job": "Producer", "name": "Jon Landau"}, {"credit_id": "539c493
7c3a36810ba002194", "department": "Art", "gender": 0, "id": 9618, "job": "Art D
irection", "name": "Sean Haworth"}, {"credit_id": "539c49b6c3a36810c10020e6",
"department": "Art", "gender": 1, "id": 12653, "job": "Set Decoration", "name":
"Kim Sinclair"}, {"credit_id": "570b6f2f9251413a0e00020d", "department": "Art",
"gender": 1, "id": 12653, "job": "Supervising Art Director", "name": "Kim Sincl
air"}, {"credit_id": "54491a6c0e0a26748c001b19", "department": "Art", "gender":
2, "id": 14350, "job": "Set Designer", "name": "Richard F. Mays"}, {"credit_i
d": "56928cf4c3a3684cff0025c4", "department": "Production", "gender": 1, "id":
20294, "job": "Executive Producer", "name": "Laeta Kalogridis"}, {"credit_id":
"52fe48009251416c750aca51", "department": "Costume & Make-Up", "gender": 0, "i
d": 17675, "job": "Costume Design", "name": "Mayes C. Rubeo"}, {"credit_id": "5
2fe48009251416c750aca11", "department": "Camera", "gender": 2, "id": 18265, "jo
b": "Director of Photography", "name": "Mauro Fiore"}, {"credit_id": "5449194d0
e0a26748f001b39", "department": "Art", "gender": 0, "id": 42281, "job": "Set De
signer", "name": "Scott Herbertson"}, {"credit_id": "52fe48009251416c750aca05",
"department": "Crew", "gender": 0, "id": 42288, "job": "Stunts", "name": "Woody
Schultz"}, {"credit_id": "5592aefb92514152de0010f5", "department": "Costume & M
ake-Up", "gender": 0, "id": 29067, "job": "Makeup Artist", "name": "Linda DeVet
ta"}, {"credit_id": "5592afa492514152de00112c", "department": "Costume & Make-U
p", "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", "department": "Art", "gender": 2, "id": 33303, "job": "Art Direction", "name": "Simon Bright"}, {"credit_id": "54959c069251417a81001f3a", "department": "Visual Effects", "gender": 0, "id": 113145, "job": "Visual Effects Supervisor", "name": "Richard Martin"}, {"credit_id": "54959a0dc3a3680ff5002c8d", "department": "Crew", "gender": 2, "id": 58188, "job": "Visual Effects Editor", "name": "Steve R. Moore"}, {"credit_id": "52fe48009251416c750aca1d", "department": "Editing", "gender": 2, "id": 58871, "job": "Editor", "name": "John Refoua"}, {"credit_id": "54491a4dc3a3680fc30018ca", "department": "Art", "gender": 0, "id": 92359, "job": "Set Designer", "name": "Karl J. Martin"}, {"credit_id": "52fe48009251416c750aca35", "department": "Camera", "gender": 1, "id": 72201, "job": "Director of Photography", "name": "Chiling Lin"}, {"credit_id": "52fe48009251416c750ac9ff", "department": "Crew", "gender": 0, "id": 89714, "job": "Stunts", "name": "Ilram Choi"}, {"credit_id": "54959c529251416e2b004394", "department": "Visual Effects", "gender": 2, "id": 93214, "job": "Visual Effects Supervisor", "name": "Steven Quale"}, {"credit_id": "54491edf0e0a267489001c37", "department": "Crew", "gender": 1, "id": 122607, "job": "Dialect Coach", "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": "539c4967c3a368653d001a94", "department": "Art", "gender": 0, "id": 132604, "job": "Art Direction", "name": "Andy McLaren"}, {"credit_id": "52fe48009251416c750aca45", "department": "Crew", "gender": 0, "id": 236696, "job": "Motion Capture Artist", "name": "Terry Notary"}, {"credit_id": "54959e02c3a3680fc60027d2", "department": "Crew", "gender": 2, "id": 956198, "job": "Stunt Coordinator", "name": "Garrett Warren"}, {"credit_id": "54959ca3c3a3686ae300438c", "department": "Visual Effects", "gender": 2, "id": 957874, "job": "Visual Effects Supervisor", "name": "Jonathan Rothbart"}, {"credit_id": "570b6f519251412c74001b2f", "department": "Art", "gender": 0, "id": 957889, "job": "Supervising Art Director", "name": "Stefan Dechant"}, {"credit_id": "570b6f62c3a3680b77007460", "department": "Art", "gender": 2, "id": 959555, "job": "Supervising Art Director", "name": "Todd Cherniawsky"}, {"credit_id": "539c4a3ac3a36810da0021cc", "department": "Production", "gender": 0, "id": 1016177, "job": "Casting", "name": "Miranda Rivers"}, {"credit_id": "539c482cc3a36810c1002062", "department": "Art", "gender": 0, "id": 1032536, "job": "Production Design", "name": "Robert Stromberg"}, {"credit_id": "539c4b65c3a36810c9002125", "department": "Costume & Make-Up", "gender": 2, "id": 1071680, "job": "Costume Design", "name": "John Harding"}, {"credit_id": "54959e6692514130fc002e4e", "department": "Camera", "gender": 0, "id": 1177364, "job": "Steadicam Operator", "name": "Roberto De Angelis"}, {"credit_id": "539c49f1c3a368653d001aac", "department": "Costume & Make-Up", "gender": 2, "id": 1202850, "job": "Makeup Department Head", "name": "Mike Smithson"}, {"credit_id": "5495999ec3a3686ae100460c", "department": "Visual Effects", "gender": 0, "id": 1204668, "job": "Visual Effects Producer", "name": "Alain Lalanne"}, {"credit_id": "54959cdfc3a3681153002729", "department": "Visual Effects", "gender": 0, "id": 1206410, "job": "Visual Effects Supervisor", "name": "Lucas Salton"}, {"credit_id": "549596239251417a81001eae", "department": "Crew", "gender": 0, "id": 1234266, "job": "Post Production Supervisor", "name": "Janace Tashjian"}, {"credit_id": "54959c859251416e1e003efe", "department": "Visual Effects", "gender": 0, "id": 1271932, "job": "Visual Effects Supervisor", "name": "Stephen Rosenbaum"}, {"credit_id": "5592af28c3a368775a00105f", "department": "Costume & Make-Up", "gender": 0, "id": 1310064, "job": "Makeup Artist", "name": "Frankie Karena"}, {"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 Direction", "name": "Robert Bavin"}, {"credit_id": "539c4a71c3a36810da0021e0", "department": "Costume & Make-Up", "gender": 0, "id": 1330567, "job": "Costume Supervisor", "name": "Anthony Almaraz"}, {"credit_id": "539c4a8ac3a36810ba0021e4", "departmen

t": "Costume & Make-Up", "gender": 0, "id": 1330570, "job": "Costume Superviso
r", "name": "Carolyn M. Fenton"}, {"credit_id": "539c4ab6c3a36810da0021f0", "de
partment": "Costume & Make-Up", "gender": 0, "id": 1330574, "job": "Costume Sup
ervisor", "name": "Beth Koenigsberg"}, {"credit_id": "54491ab70e0a267480001ba
2", "department": "Art", "gender": 0, "id": 1336191, "job": "Set Designer", "na
me": "Sam Page"}, {"credit_id": "544919d9c3a3680fc30018bd", "department": "Ar
t", "gender": 0, "id": 1339441, "job": "Set Designer", "name": "Tex Kadonaga"},
{"credit_id": "54491cf50e0a267483001b0c", "department": "Editing", "gender": 0,
"id": 1352422, "job": "Dialogue Editor", "name": "Kim Foscato"}, {"credit_id":
"544919f40e0a26748c001b09", "department": "Art", "gender": 0, "id": 1352962, "j
ob": "Set Designer", "name": "Tammy S. Lee"}, {"credit_id": "5495a115c3a3680ff5
002d71", "department": "Crew", "gender": 0, "id": 1357070, "job": "Transportati
on Coordinator", "name": "Denny Caira"}, {"credit_id": "5495a12f92514130fc002e9
4", "department": "Crew", "gender": 0, "id": 1357071, "job": "Transportation Co
ordinator", "name": "James Waitkus"}, {"credit_id": "5495976fc3a36811530026b0",
"department": "Sound", "gender": 0, "id": 1360103, "job": "Supervising Sound Ed
itor", "name": "Addison Teague"}, {"credit_id": "54491837c3a3680fb1001c5a", "de
partment": "Art", "gender": 2, "id": 1376887, "job": "Set Designer", "name":
"C. Scott Baker"}, {"credit_id": "54491878c3a3680fb4001c9d", "department": "Ar
t", "gender": 0, "id": 1376888, "job": "Set Designer", "name": "Luke Caska"},
{"credit_id": "544918dac3a3680fa5001ae0", "department": "Art", "gender": 0, "i
d": 1376889, "job": "Set Designer", "name": "David Chow"}, {"credit_id": "54491
9110e0a267486001b68", "department": "Art", "gender": 0, "id": 1376890, "job":
"Set Designer", "name": "Jonathan Dyer"}, {"credit_id": "54491967c3a3680faa001b
5e", "department": "Art", "gender": 0, "id": 1376891, "job": "Set Designer", "n
ame": "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 J
ohnson"}, {"credit_id": "54491b1dc3a3680faa001b8c", "department": "Art", "gende
r": 0, "id": 1376895, "job": "Assistant Art Director", "name": "Mike Stassi"},
{"credit_id": "54491b79c3a3680fbb001826", "department": "Art", "gender": 0, "i
d": 1376897, "job": "Construction Coordinator", "name": "John Villarino"}, {"cr
edit_id": "54491baec3a3680fb4001ce6", "department": "Art", "gender": 2, "id": 1
376898, "job": "Assistant Art Director", "name": "Jeffrey Wisniewski"}, {"credi
t_id": "54491d2fc3a3680fb4001d07", "department": "Editing", "gender": 0, "id":
1376899, "job": "Dialogue Editor", "name": "Cheryl Nardi"}, {"credit_id": "5449
1d86c3a3680fa5001b2f", "department": "Editing", "gender": 0, "id": 1376901, "jo
b": "Dialogue Editor", "name": "Marshall Winn"}, {"credit_id": "54491d9dc3a3680
faa001bb0", "department": "Sound", "gender": 0, "id": 1376902, "job": "Supervis
ing Sound Editor", "name": "Gwendolyn Yates Whittle"}, {"credit_id": "54491dc10
e0a267486001bce", "department": "Sound", "gender": 0, "id": 1376903, "job": "So
und Re-Recording Mixer", "name": "William Stein"}, {"credit_id": "54491f500e0a2
6747c001c07", "department": "Crew", "gender": 0, "id": 1376909, "job": "Choreog
rapher", "name": "Lula Washington"}, {"credit_id": "549599239251412c4e002a2e",
"department": "Visual Effects", "gender": 0, "id": 1391692, "job": "Visual Effe
cts Producer", "name": "Chris Del Conte"}, {"credit_id": "54959d54c3a36831b8001
d9a", "department": "Visual Effects", "gender": 2, "id": 1391695, "job": "Visua
l Effects Supervisor", "name": "R. Christopher White"}, {"credit_id": "54959bdf
9251412c4e002a66", "department": "Visual Effects", "gender": 0, "id": 1394070,
"job": "Visual Effects Supervisor", "name": "Dan Lemmon"}, {"credit_id": "54959
71d92514132ed002922", "department": "Sound", "gender": 0, "id": 1394129, "job":
"Sound Effects Editor", "name": "Tim Nielsen"}, {"credit_id": "5592b25792514152
cc0011aa", "department": "Crew", "gender": 0, "id": 1394286, "job": "CG Supervi
sor", "name": "Michael Mulholland"}, {"credit_id": "54959a329251416e2b004355",
"department": "Crew", "gender": 0, "id": 1394750, "job": "Visual Effects Edito
r", "name": "Thomas Nittmann"}, {"credit_id": "54959d6dc3a3686ae9004401", "depa
rtment": "Visual Effects", "gender": 0, "id": 1394755, "job": "Visual Effects S
upervisor", "name": "Edson Williams"}, {"credit_id": "5495a08fc3a3686ae300441
c", "department": "Editing", "gender": 0, "id": 1394953, "job": "Digital Interm

ediate", "name": "Christine Carr"}, {"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": "54959734c3a3686ae10045e0", "department": "Sound", "gender": 0, "id": 1400906, "job": "Sound Effects Editor", "name": "Christopher Scarabosio"}, {"credit_id": "549595dd92514130fc002d79", "department": "Production", "gender": 0, "id": 1401784, "job": "Production Supervisor", "name": "Jennifer Teves"}, {"credit_id": "549596009251413af70028cc", "department": "Production", "gender": 0, "id": 1401785, "job": "Production Manager", "name": "Brigitte Yorke"}, {"credit_id": "549596e892514130fc002d99", "department": "Sound", "gender": 0, "id": 1401786, "job": "Sound Effects Editor", "name": "Ken Fischer"}, {"credit_id": "549598229251412c4e002a1c", "department": "Crew", "gender": 0, "id": 1401787, "job": "Special Effects Coordinator", "name": "Iain Hutton"}, {"credit_id": "549598349251416e2b00432b", "department": "Crew", "gender": 0, "id": 1401788, "job": "Special Effects Coordinator", "name": "Steve Ingram"}, {"credit_id": "54959905c3a3686ae3004324", "department": "Visual Effects", "gender": 0, "id": 1401789, "job": "Visual Effects Producer", "name": "Joyce Cox"}, {"credit_id": "5495994b92514132ed002951", "department": "Visual Effects", "gender": 0, "id": 1401790, "job": "Visual Effects Producer", "name": "Jenny Foster"}, {"credit_id": "549599cbc3a3686ae1004613", "department": "Crew", "gender": 0, "id": 1401791, "job": "Visual Effects Editor", "name": "Christopher Marino"}, {"credit_id": "549599f2c3a3686ae100461e", "department": "Crew", "gender": 0, "id": 1401792, "job": "Visual Effects Editor", "name": "Jim Milton"}, {"credit_id": "54959a51c3a3686af3003eb5", "department": "Visual Effects", "gender": 0, "id": 1401793, "job": "Visual Effects Producer", "name": "Cyndi Ochs"}, {"credit_id": "54959a7cc3a36811530026f4", "department": "Crew", "gender": 0, "id": 1401794, "job": "Visual Effects Editor", "name": "Lucas Putnam"}, {"credit_id": "54959b91c3a3680ff5002cb4", "department": "Visual Effects", "gender": 0, "id": 1401795, "job": "Visual Effects Supervisor", "name": "Anthony \'Max\' Ivins"}, {"credit_id": "54959bb69251412c4e002a5f", "department": "Visual Effects", "gender": 0, "id": 1401796, "job": "Visual Effects Supervisor", "name": "John Knoll"}, {"credit_id": "54959cbbc3a3686ae3004391", "department": "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": "Visual Effects Supervisor", "name": "Guy Williams"}, {"credit_id": "54959de4c3a3681153002750", "department": "Crew", "gender": 0, "id": 1401804, "job": "Stunt Coordinator", "name": "Stuart Thorp"}, {"credit_id": "54959ef2c3a3680fc60027f2", "department": "Lighting", "gender": 0, "id": 1401805, "job": "Best Boy Electric", "name": "Giles Coburn"}, {"credit_id": "54959f07c3a3680fc60027f9", "department": "Camera", "gender": 2, "id": 1401806, "job": "Still Photographer", "name": "Mark Fellman"}, {"credit_id": "54959f47c3a3681153002774", "department": "Lighting", "gender": 0, "id": 1401807, "job": "Lighting Technician", "name": "Scott Sprague"}, {"credit_id": "54959f8cc3a36831b8001df2", "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": "54959fb6c3a3686af3003f54", "department": "Visual Effects", "gender": 0, "id": 1401810, "job": "Animation Director", "name": "Taisuke Tanimura"}, {"credit_id": "54959fd2c3a36831b8001e02", "department": "Costume & Make-Up", "gender": 0, "id": 1401812, "job": "Set Costumer", "name": "Lilia Mishel Acevedo"}, {"credit_id": "54959ff9c3a3686ae300440c", "department": "Costume & Make-Up", "gender": 0, "id": 1401814, "job": "Set Costumer", "name": "Alejandro M. Hernandez"}, {"credit_id": "5495a0ddc3a3686ae10046fe", "department": "Editing", "gender": 0, "id": 1401815, "job": "Digital Intermediate", "name": "Marvin Hal

l"}, {"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, "job": "Conceptual Design", "name": "Seth Engstrom"}, {"credit_id": "5525d5809251417276002b06", "department": "Crew", "gender": 0, "id": 1447362, "job": "Visual Effects Art Director", "name": "Eric Oliver"}, {"credit_id": "554427ca925141586500312a", "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", "name": "Arun Ram-Mohan"}, {"credit_id": "5592af4692514152d5001355", "department": "Costume & Make-Up", "gender": 0, "id": 1457305, "job": "Makeup Artist", "name": "Georgia Lockhart-Adams"}, {"credit_id": "5592b2eac3a36877470012a5", "department": "Crew", "gender": 0, "id": 1466035, "job": "CG Supervisor", "name": "Thrain Shadbolt"}, {"credit_id": "5592b032c3a36877450015f1", "department": "Crew", "gender": 0, "id": 1483220, "job": "CG Supervisor", "name": "Brad Alexander"}, {"credit_id": "5592b05592514152d80012f6", "department": "Crew", "gender": 0, "id": 1483221, "job": "CG Supervisor", "name": "Shadi Almassizadeh"}, {"credit_id": "5592b090c3a36877570010b5", "department": "Crew", "gender": 0, "id": 1483222, "job": "CG Supervisor", "name": "Simon Clutterbuck"}, {"credit_id": "5592b0dbc3a368774b00112c", "department": "Crew", "gender": 0, "id": 1483223, "job": "CG Supervisor", "name": "Graeme Dem mocks"}, {"credit_id": "5592b0fe92514152db0010c1", "department": "Crew", "gender": 0, "id": 1483224, "job": "CG Supervisor", "name": "Adrian Fernandes"}, {"credit_id": "5592b11f9251415df8001059", "department": "Crew", "gender": 0, "id": 1483225, "job": "CG Supervisor", "name": "Mitch Gates"}, {"credit_id": "5592b15dc3a3687745001645", "department": "Crew", "gender": 0, "id": 1483226, "job": "CG Supervisor", "name": "Jerry Kung"}, {"credit_id": "5592b18e925141645a0004ae", "department": "Crew", "gender": 0, "id": 1483227, "job": "CG Supervisor", "name": "Andy Lomas"}, {"credit_id": "5592b1bfc3a368775d0010e7", "department": "Crew", "gender": 0, "id": 1483228, "job": "CG Supervisor", "name": "Sebastian Marino"}, {"credit_id": "5592b2049251415df8001078", "department": "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": "Crew", "gender": 0, "id": 1483232, "job": "CG Supervisor", "name": "Michael Takarangi"}, {"credit_id": "5592b345c3a36877470012bb", "department": "Crew", "gender": 0, "id": 1483233, "job": "CG Supervisor", "name": "David Weitzberg"}, {"credit_id": "5592b37cc3a368775100113b", "department": "Crew", "gender": 0, "id": 1483234, "job": "CG Supervisor", "name": "Ben White"}, {"credit_id": "573c8e2f9251413f5d000094", "department": "Crew", "gender": 1, "id": 1621932, "job": "Stunts", "name": "Min Windle"}]'

In [31]:

```python
def fetch_crew(text):
    unique_names = set()
    valid_jobs = {'Director', 'Writer', 'Producer',
                  'Cinematographer', 'Editor', 'Composer'}

    for i in ast.literal_eval(text):
        if i['job'] in valid_jobs:
            unique_names.add(i['name'])  #automatically handles duplicates
```

```
    return list(unique_names)  #convert set back to list
```

In [32]:
```
movies['crew'] = movies['crew'].apply(fetch_crew)
```

In [33]:
```
movies.head()
```

Out[33]:

	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 Weave...	[Jon Landau, James Cameron, John Refoua, Steph...
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, ...	[Pat Sandston, Jerry Bruckheimer, Peter Kohn, ...
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, R...	[Michael G. Wilson, Lee Smith, Barbara Broccol...
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, A...	[Christopher Nolan, Charles Roven, Lee Smith, ...
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, Lindsey Collins, Jim Morris, ...

In [34]:
```
movies.iloc[0]['crew']
#checking to see how the cast column looks after the changes
```

Out[34]:
```
['Jon Landau', 'James Cameron', 'John Refoua', 'Stephen E. Rivkin']
```

In [35]:
```
movies.iloc[0]['overview']
#checking to see how the overview column looks after the changes
```

Out[35]:
```
'In the 22nd century, a paraplegic Marine is dispatched to the moon Pandora on a unique mission, but becomes torn between following orders and protecting an a lien civilization.'
```

In [36]:
```
movies['overview'] = movies['overview'].apply(lambda x:x.split())
movies.iloc[0]['overview']
#changing the string to list format so i can concatinate the columns
```

Out[36]:
```
['In',
 'the',
 '22nd',
 'century,',
 'a',
 'paraplegic',
 'Marine',
 'is',
 'dispatched',
 'to',
 'the',
 'moon',
 'Pandora',
 'on',
 'a',
 'unique',
 'mission,',
 'but',
 'becomes',
 'torn',
 'between',
 'following',
 'orders',
 'and',
 'protecting',
 'an',
 'alien',
 'civilization.']
```

In [37]:
```
movies.head()
```

Out[37]:

	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 Weave...	[Jon Landau, James Cameron, John Refoua, Steph...
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, ...	[Pat Sandston, Jerry Bruckheimer, Peter Kohn, ...
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, R...	[Michael G. Wilson, Lee Smith, Barbara Broccol...
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, A...	[Christopher Nolan, Charles Roven, Lee Smith, ...
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, Lindsey Collins, Jim Morris, ...

In [38]:
```
def remove_space(word):
    l=[]
    for i in word:
        l.append(i.replace(" ","")) #this replaces a space with no space
    return l
```

In [39]:
```
movies['cast'] = movies['cast'].apply(remove_space)
movies['crew'] = movies['crew'].apply(remove_space)
movies['genres'] = movies['genres'].apply(remove_space)
movies['keywords'] = movies['keywords'].apply(remove_space)
```

In [40]:
```
movies.head()
```

Out[40]:

	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, ...	
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, Ste...	Je
2	206647	Spectre	[A, cryptic, message, from, Bond’s, past, send...	[Action, Adventure, Crime]	[spy, basedonnovel, secretagent, sequel, mi6, ...	[DanielCraig, ChristophWaltz, LéaSeydoux, Ralp...	[M
3	49026	The Dark Knight Rises	[Following, the, death, of, District, Attorney...	[Action, Crime, Drama, Thriller]	[dccomics, crimefighter, terrorist, secretiden...	[ChristianBale, MichaelCaine, GaryOldman, Anne...	[Ch L
4	49529	John Carter	[John, Carter, is, a, war-weary,, former, mili...	[Action, Adventure, ScienceFiction]	[basedonnovel, mars, medallion, spacetravel, p...	[TaylorKitsch, LynnCollins, SamanthaMorton, Wi...	[

In [41]:
```
movies['tags'] = movies['overview'] + movies['genres'] +
movies['overview'] + movies['keywords'] + movies['cast'] + movies['crew']
#i am concatinating all the columns into one column named tags
```

In [42]:
```
movies.head()
#checking to see if the required changes have been made
```

Out[42]:

	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, ...	
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, Ste...	Je
2	206647	Spectre	[A, cryptic, message, from, Bond's, past, send...	[Action, Adventure, Crime]	[spy, basedonnovel, secretagent, sequel, mi6, ...	[DanielCraig, ChristophWaltz, LéaSeydoux, Ralp...	[M
3	49026	The Dark Knight Rises	[Following, the, death, of, District, Attorney...	[Action, Crime, Drama, Thriller]	[dccomics, crimefighter, terrorist, secretiden...	[ChristianBale, MichaelCaine, GaryOldman, Anne...	[Ch L
4	49529	John Carter	[John, Carter, is, a, war-weary,, former, mili...	[Action, Adventure, ScienceFiction]	[basedonnovel, mars, medallion, spacetravel, p...	[TaylorKitsch, LynnCollins, SamanthaMorton, Wi...	[

In [43]:
```
new_df = movies[['movie_id', 'title', 'tags']]
#creating a new df with only movie_id, title and tags
```

In [44]:
```
new_df.head()
#checking to see if the required changes have been made
```

Out[44]:

	movie_id	title	tags
0	19995	Avatar	[In, the, 22nd, century,, a, paraplegic, Marin...
1	285	Pirates of the Caribbean: At World's End	[Captain, Barbossa,, long, believed, to, be, d...
2	206647	Spectre	[A, cryptic, message, from, Bond's, past, send...
3	49026	The Dark Knight Rises	[Following, the, death, of, District, Attorney...
4	49529	John Carter	[John, Carter, is, a, war-weary,, former, mili...

In [45]:
```
new_df['tags'] = new_df['tags'].apply(lambda x: " ".join(x))
#takes each list of words in the 'tags' column
```

```
#turns it into a single sentence with spaces in between
```

```
C:\Users\srivi\AppData\Local\Temp\ipykernel_15892\365633490.py:1: SettingWithCopy
Warning:
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/stabl
e/user_guide/indexing.html#returning-a-view-versus-a-copy
  new_df['tags'] = new_df['tags'].apply(lambda x: " ".join(x))
```

In [46]:
```
new_df.head()
#checking to see if the required changes have been made
```

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...

In [47]:
```
new_df.iloc[0]['tags']
#seeing only the first element of the tags column
```

Out[47]:
```
'In the 22nd century, a paraplegic Marine is dispatched to the moon Pandora on
a unique mission, but becomes torn between following orders and protecting an a
lien civilization. Action Adventure Fantasy ScienceFiction In the 22nd century,
a paraplegic Marine is dispatched to the moon Pandora on a unique mission, but
becomes torn between following orders and protecting an alien civilization. cul
tureclash future spacewar spacecolony society spacetravel futuristic romance sp
ace alien tribe alienplanet cgi marine soldier battle loveaffair antiwar powerr
elations mindandsoul 3d SamWorthington ZoeSaldana SigourneyWeaver StephenLang M
ichelleRodriguez GiovanniRibisi JoelDavidMoore CCHPounder WesStudi LazAlonso Di
leepRao MattGerald SeanAnthonyMoran JasonWhyte ScottLawrence KellyKilgour James
PatrickPitt SeanPatrickMurphy PeterDillon KevinDorman KelsonHenderson DavidVanH
orn JacobTomuri MichaelBlain-Rozgay JonCurry LukeHawker WoodySchultz PeterMensa
h SoniaYee JahnelCurfman IlramChoi KylaWarren LisaRoumain DebraWilson ChrisMala
TaylorKibby JodieLandau JulieLamm CullenB.Madden JosephBradyMadden FrankieTorre
s AustinWilson SaraWilson TamicaWashington-Miller LucyBriant NathanMeister Gerr
yBlair MatthewChamberlain PaulYates WrayWilson JamesGaylyn MelvinLenoClarkIII C
arvonFutrell BrandonJelkes MicahMoch HanniyahMuhammad ChristopherNolen ChristaO
liver AprilMarieThomas BravitaA.Threatt DeanKnowsley JuleneRenee JonLandau Jame
sCameron JohnRefoua StephenE.Rivkin'
```

In [48]:
```
new_df['tags'] = new_df['tags'].apply(lambda x:x.lower())
#changing the content in the tags column to be lowercase
```

```
C:\Users\srivi\AppData\Local\Temp\ipykernel_15892\3446560648.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())
```

In [49]:
```
new_df.iloc[0]['tags']
#checking to see if the required changes have been made
```

Out[49]:
```
'in the 22nd century, a paraplegic marine is dispatched to the moon pandora on a unique mission, but becomes torn between following orders and protecting an alien civilization. action adventure fantasy sciencefiction in the 22nd century, a paraplegic marine is dispatched to the moon pandora on a unique mission, but becomes torn between following orders and protecting an alien civilization. cultureclash future spacewar spacecolony society spacetravel futuristic romance space alien tribe alienplanet cgi marine soldier battle loveaffair antiwar powerrelations mindandsoul 3d samworthington zoesaldana sigourneyweaver stephenlang michellerodriguez giovanniribisi joeldavidmoore cchpounder wesstudi lazalonso dileeprao mattgerald seananthonymoran jasonwhyte scottlawrence kellykilgour jamespatrickpitt seanpatrickmurphy peterdillon kevindorman kelsonhenderson davidvanhorn jacobtomuri michaelblain-rozgay joncurry lukehawker woodyschultz petermensah soniayee jahnelcurfman ilramchoi kylawarren lisaroumain debrawilson chrismalataylorkibby jodielandau julielamm cullenb.madden josephbradymadden frankietorres austinwilson sarawilson tamicawashington-miller lucybriant nathanmeister gerryblair matthewchamberlain paulyates wraywilson jamesgaylyn melvinlenoclarkiii carvonfutrell brandonjelkes micahmoch hanniyahmuhammad christophernolen christaoliver aprilmariethomas bravitaa.threatt deanknowsley julenerenee jonlandau jamescameron johnrefoua stephene.rivkin'
```

In [50]:
```
new_df.shape
```

Out[50]:
```
(4806, 3)
```

What we will do with the tags next:

So far, I have created a descriptive tag for each movie, combining relevant details such as cast, crew, genre, and plot summary into a single textual representation.

1. Vectorisation Using Count Vectoriser:

- Converted the textual tags into numerical data by applying Count Vectorisation, which transforms text into a matrix of token counts.

2. Global Word Frequency Analysis:

- Combined tags from all 4,806 movies into a single large text corpus.
- Identified unique words and extracted key features related to movies.
- Counted the 9,000 most frequently occurring words in this corpus.

3. Feature Representation:

- Created a dataset with 9,000 columns, each representing one of the most common words.
- Compared each movie's original tags against these words to determine their frequency.
- Represented each movie as a feature vector based on word frequency.

4. Similarity Calculation:

- Plotted the movie vectors in a multi-dimensional space.
- Used Cosine Similarity to measure the distance between vectors, where a smaller angle indicates higher similarity.

5. Recommendation Generation:

- Based on the computed similarities, identified and recommended the top 5 most similar movies for a given input movie.

In [52]:
```
import nltk
from nltk.stem import PorterStemmer
```

In [53]:
```
ps = PorterStemmer()
```

In [54]:
```
def stems(text):
    l = [] #initialises an empty list
    for i in text.split():
        #splits the input text into individual words (loops)
        l.append(ps.stem(i))
        #pplies stemming to each word and adds to list

    return " ".join(l)
    #adds stemmed words nack into single string and returns
```

In [55]:
```
new_df['tags'] = new_df['tags'].apply(stems)
```

```
C:\Users\srivi\AppData\Local\Temp\ipykernel_15892\3973021881.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(stems)
```

In [56]:
```
new_df.iloc[0]['tags']
```

Out[56]:
```
'in the 22nd century, a parapleg marin is dispatch to the moon pandora on a uni
qu mission, but becom torn between follow order and protect an alien civilizati
on. action adventur fantasi sciencefict in the 22nd century, a parapleg marin i
s dispatch to the moon pandora on a uniqu mission, but becom torn between follo
w order and protect an alien civilization. cultureclash futur spacewar spacecol
oni societi spacetravel futurist romanc space alien tribe alienplanet cgi marin
soldier battl loveaffair antiwar powerrel mindandsoul 3d samworthington zoesald
ana sigourneyweav stephenlang michellerodriguez giovanniribisi joeldavidmoor cc
hpounder wesstudi lazalonso dileeprao mattgerald seananthonymoran jasonwhyt sco
ttlawr kellykilgour jamespatrickpitt seanpatrickmurphi peterdillon kevindorman
kelsonhenderson davidvanhorn jacobtomuri michaelblain-rozgay joncurri lukehawk
woodyschultz petermensah soniaye jahnelcurfman ilramchoi kylawarren lisaroumain
debrawilson chrismala taylorkibbi jodielandau julielamm cullenb.madden josephbr
adymadden frankietorr austinwilson sarawilson tamicawashington-mil lucybri nath
anmeist gerryblair matthewchamberlain paulyat wraywilson jamesgaylyn melvinleno
clarkiii carvonfutrel brandonjelk micahmoch hanniyahmuhammad christophernolen c
hristaoliv aprilmariethoma bravitaa.threatt deanknowsley julenerene jonlandau j
amescameron johnrefoua stephene.rivkin'
```

In [57]:
```
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(max_features=9000, stop_words= 'english')
```

In [58]:
```
vector = cv.fit_transform(new_df['tags']).toarray()
```

In [59]:
```
vector
```

Out[59]:
```
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)
```

In [60]:
```
vector.shape
```

Out[60]:
```
(4806, 9000)
```

In [61]:
```
from sklearn.metrics.pairwise import cosine_similarity
```

In [62]:
```
similarity = cosine_similarity(vector)
```

In [63]:
```
similarity
```

Out[63]:
```
array([[1.        , 0.05183087, 0.03850607, ..., 0.02234951, 0.        ,
        0.        ],
       [0.05183087, 1.        , 0.02963767, ..., 0.03058161, 0.        ,
        0.03241444],
       [0.03850607, 0.02963767, 1.        , ..., 0.02839952, 0.        ,
        0.        ],
       ...,
       [0.02234951, 0.03058161, 0.02839952, ..., 1.        , 0.04131623,
        0.03494283],
       [0.        , 0.        , 0.        , ..., 0.04131623, 1.        ,
        0.08758482],
       [0.        , 0.03241444, 0.        , ..., 0.03494283, 0.08758482,
        1.        ]])
```

In [64]:
```
similarity.shape
#comparing the 4806 rows with themselves
```

Out[64]: (4806, 4806)

In [65]:
```
new_df[new_df['title'] == 'Spider-Man'].index[0]
#can calculate the index of each movie
```

Out[65]: 159

In [66]:
```
def recommend(movie):
    index = new_df[new_df['title'] == movie].index[0]
    #find he index of the given movie in dataset

    distances = sorted(list(enumerate(similarity[index])),
                       reverse=True, key= lambda x: x[1])
    #compute the similarity score with all movies
    #sort them in descending order based on simalrity score

    for i in distances[1:6]:
        print(new_df.iloc[i[0]].title)
        #print the title of the top 5 similar movies
```

In [67]:
```
recommend('Spider-Man')
#checking to see if it is working by inserting a movie
```

```
Spider-Man 3
Spider-Man 2
The Amazing Spider-Man 2
21 Jump Street
Arachnophobia
```

In [68]:
```
recommend('The Dark Knight Rises')
```

```
The Dark Knight
Batman Forever
Batman
Batman Returns
Batman
```

In [69]:
```
import pickle
import os

# create the directory if it doesn't exist
os.makedirs("artifacts", exist_ok=True)

# save the data using pickle
pickle.dump(new_df, open("artifacts/movie_list.pkl", "wb"))
pickle.dump(similarity, open("artifacts/similarity.pkl", "wb"))
```

In [70]:
```
import os

#path of the artifacts folder
artifacts_path = os.path.abspath("artifacts")
print("Artifacts folder is located at:", artifacts_path)
```

```
Artifacts folder is located at: C:\Users\srivi\Documents\artifacts
```

In [ ]:

In [ ]: