Requirements -

- Python
- Libraries (Numpy, Pandas)
- Modules - ast, re, pickle
- NLP - nltk
- ML - sklearn
- Linear Algebra
- Streamlit - Webapp

In [216]:
```
import re
import ast
import pickle

import numpy as np
import pandas as pd
import streamlit as st

import nltk
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer

import sklearn
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity

import warnings
warnings.filterwarnings('ignore')
```

In [ ]:
```
```

In [3]:
```
movies = pd.read_csv('tmdb_5000_movies.csv')
credits = pd.read_csv('tmdb_5000_credits.csv')
```

In [4]:
```
movies.head(3)
```

Out[4]:

	budget	genres	homepage	id	keywords	original_language	original_title	overview
0	237000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.avatarmovie.com/	19995	[{"id": 1463, "name": "culture clash"}, {"id":...	en	Avatar	In the 22nd century, a paraplegic Marine is di...
1	300000000	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	http://disney.go.com/disneypictures/pirates/	285	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	en	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...
2	245000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.sonypictures.com/movies/spectre/	206647	[{"id": 470, "name": "spy"}, {"id": 818, "name...	en	Spectre	A cryptic message from Bond’s past sends him o...

In [ ]:
```

```

In [5]:
```
credits.head()
```

Out[5]:

	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 [ ]:
```
```

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

Out[6]:
```
((4803, 20), (4803, 4))
```

In [ ]:
```
```

Merge Operation

In [7]:
```
movies.columns
```

Out[7]:
```
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'],
      dtype='object')
```

In [8]:
```
credits.columns
```

Out[8]:
```
Index(['movie_id', 'title', 'cast', 'crew'], dtype='object')
```

In [ ]:
```

```

In [9]:
```
df = movies.merge(credits, on='title')
df.head(2)
```

Out[9]:

	budget	genres	homepage	id	keywords	original_language	original_title	overview	p
0	237000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.avatarmovie.com/	19995	[{"id": 1463, "name": "culture clash"}, {"id":...	en	Avatar	In the 22nd century, a paraplegic Marine is di...	15
1	300000000	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	http://disney.go.com/disneypictures/pirates/	285	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	en	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	13

2 rows × 23 columns

In [10]:
```
df.shape
```

Out[10]:
```
(4809, 23)
```

In [ ]:
```

```

Choose best features for a movie recommendation

- movie_id
- title
- overview

- genres
- keywords
- cast
- crew

In [11]:
```
df[:2]
```

Out[11]:

	budget	genres	homepage	id	keywords	original_language	original_title	overview	p
0	237000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.avatarmovie.com/	19995	[{"id": 1463, "name": "culture clash"}, {"id":...	en	Avatar	In the 22nd century, a paraplegic Marine is di...	15
1	300000000	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	http://disney.go.com/disneypictures/pirates/	285	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	en	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	13

2 rows × 23 columns

In [ ]:
```

```

In [12]:
```
df = df[['movie_id', 'title', 'overview', 'genres', 'keywords', 'cast', 'crew']]
df.head()
```

Out[12]:

	movie_id	title	overview	genres	keywords	cast	crew
0	19995	Avatar	In the 22nd century, a paraplegic Marine is di...	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	[{"id": 1463, "name": "culture clash"}, {"id":...	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
1	285	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	[{"cast_id": 4, "character": "Captain Jack Spa...	[{"credit_id": "52fe4232c3a36847f800b579", "de...
2	206647	Spectre	A cryptic message from Bond's past sends him o...	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	[{"id": 470, "name": "spy"}, {"id": 818, "name...	[{"cast_id": 1, "character": "James Bond", "cr...	[{"credit_id": "54805967c3a36829b5002c41", "de...
3	49026	The Dark Knight Rises	Following the death of District Attorney Harve...	[{"id": 28, "name": "Action"}, {"id": 80, "nam...	[{"id": 849, "name": "dc comics"}, {"id": 853,...	[{"cast_id": 2, "character": "Bruce Wayne / Ba...	[{"credit_id": "52fe4781c3a36847f81398c3", "de...
4	49529	John Carter	John Carter is a war-weary, former military ca...	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	[{"id": 818, "name": "based on novel"}, {"id":...	[{"cast_id": 5, "character": "John Carter", "c...	[{"credit_id": "52fe479ac3a36847f813eaa3", "de...

In [ ]:
```
```

Target - 'movie_id' + 'title' + 'tags'

In [13]:
```
df['overview'][0]
```

Out[13]:
```
'In the 22nd century, a paraplegic Marine is dispatched to the moon Pandora on a unique mission, but bec
omes torn between following orders and protecting an alien civilization.'
```

In [14]:
```
df['genres'][0]
```

Out[14]:
```
'[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 8
78, "name": "Science Fiction"}]'
```

In [15]:
```
df['keywords'][0]
```

Out[15]:
```
'[{"id": 1463, "name": "culture clash"}, {"id": 2964, "name": "future"}, {"id": 3386, "name": "space wa
r"}, {"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": "spa
ce"}, {"id": 9951, "name": "alien"}, {"id": 10148, "name": "tribe"}, {"id": 10158, "name": "alien plane
t"}, {"id": 10987, "name": "cgi"}, {"id": 11399, "name": "marine"}, {"id": 13065, "name": "soldier"},
{"id": 14643, "name": "battle"}, {"id": 14720, "name": "love affair"}, {"id": 165431, "name": "anti wa
r"}, {"id": 193554, "name": "power relations"}, {"id": 206690, "name": "mind and soul"}, {"id": 209714,
"name": "3d"}]'
```

In [16]:
```
df['cast'][0]
```

Out[16]:
```
'[{"cast_id": 242, "character": "Jake Sully", "credit_id": "5602a8a7c3a3685532001c9a", "gender": 2, "i
d": 65731, "name": "Sam Worthington", "order": 0}, {"cast_id": 3, "character": "Neytiri", "credit_id":
"52fe48009251416c750ac9cb", "gender": 1, "id": 8691, "name": "Zoe Saldana", "order": 1}, {"cast_id": 2
5, "character": "Dr. Grace Augustine", "credit_id": "52fe48009251416c750aca39", "gender": 1, "id": 102
05, "name": "Sigourney Weaver", "order": 2}, {"cast_id": 4, "character": "Col. Quaritch", "credit_id":
"52fe48009251416c750ac9cf", "gender": 2, "id": 32747, "name": "Stephen Lang", "order": 3}, {"cast_id":
5, "character": "Trudy Chacon", "credit_id": "52fe48009251416c750ac9d3", "gender": 1, "id": 17647, "na
me": "Michelle Rodriguez", "order": 4}, {"cast_id": 8, "character": "Selfridge", "credit_id": "52fe480
09251416c750ac9e1", "gender": 2, "id": 1771, "name": "Giovanni Ribisi", "order": 5}, {"cast_id": 7, "c
haracter": "Norm Spellman", "credit_id": "52fe48009251416c750ac9dd", "gender": 2, "id": 59231, "name":
"Joel David Moore", "order": 6}, {"cast_id": 9, "character": "Moat", "credit_id": "52fe48009251416c750
ac9e5", "gender": 1, "id": 30485, "name": "CCH Pounder", "order": 7}, {"cast_id": 11, "character": "Ey
tukan", "credit_id": "52fe48009251416c750ac9ed", "gender": 2, "id": 15853, "name": "Wes Studi", "orde
r": 8}, {"cast_id": 10, "character": "Tsu\'Tey", "credit_id": "52fe48009251416c750ac9e9", "gender": 2,
"id": 10964, "name": "Laz Alonso", "order": 9}, {"cast_id": 12, "character": "Dr. Max Patel", "credit_
id": "52fe48009251416c750ac9f1", "gender": 2, "id": 95697, "name": "Dileep Rao", "order": 10}, {"cast_
id": 13, "character": "Lyle Wainfleet", "credit_id": "52fe48009251416c750ac9f5", "gender": 2, "id": 98
215, "name": "Matt Gerald", "order": 11}, {"cast_id": 32, "character": "Private Fike", "credit_id": "5
2fe48009251416c750aca5b", "gender": 2, "id": 154153, "name": "Sean Anthony Moran", "order": 12}, {"cas
```

In [17]:
```
df['crew'][0]
```

Out[17]:
'[{"credit_id": "52fe48009251416c750aca23", "department": "Editing", "gender": 0, "id": 1721, "job": "Editor", "name": "Stephen E. Rivkin"}, {"credit_id": "539c47ecc3a36810e3001f87", "department": "Art", "gender": 2, "id": 496, "job": "Production Design", "name": "Rick Carter"}, {"credit_id": "54491c89c3a3680fb4001cf7", "department": "Sound", "gender": 0, "id": 900, "job": "Sound Designer", "name": "Christopher Boyes"}, {"credit_id": "54491cb70e0a267480001bd0", "department": "Sound", "gender": 0, "id": 900, "job": "Supervising Sound Editor", "name": "Christopher Boyes"}, {"credit_id": "539c4a4cc3a36810c9002101", "department": "Production", "gender": 1, "id": 1262, "job": "Casting", "name": "Mali Finn"}, {"credit_id": "5544ee3b925141499f0008fc", "department": "Sound", "gender": 2, "id": 1729, "job": "Original Music Composer", "name": "James Horner"}, {"credit_id": "52fe48009251416c750ac9c3", "department": "Directing", "gender": 2, "id": 2710, "job": "Director", "name": "James Cameron"}, {"credit_id": "52fe48009251416c750ac9d9", "department": "Writing", "gender": 2, "id": 2710, "job": "Writer", "name": "James Cameron"}, {"credit_id": "52fe48009251416c750aca17", "department": "Editing", "gender": 2, "id": 2710, "job": "Editor", "name": "James Cameron"}, {"credit_id": "52fe48009251416c750aca29", "department": "Production", "gender": 2, "id": 2710, "job": "Producer", "name": "James Cameron"}, {"credit_id": "52fe48009251416c750aca3f", "department": "Writing", "gender": 2, "id": 2710, "job": "Screenplay", "name": "James Cameron"}, {"credit_id": "539c4987c3a36810ba0021a4", "department": "Art", "gender": 2, "id": 7236, "job": "Art Direction", "name": "Andrew Menzies"}, {"credit_id": "549598c3c3a3686ae9004383", "department": "Visual Effects", "gender": 0, "id": 6690, "job": "Visual Effects Producer", "name": "Jill Brooks"}, {"credit_id": "52fe48009251416c750aca4b", "department": "Production", "gender": 1, "id": 6347,

In [ ]:
```

```

In [ ]:
```

```

Let's start with Genres Column

In [18]:
```
df['genres']
```

Out[18]:
```
0       [{"id": 28, "name": "Action"}, {"id": 12, "nam...
1       [{"id": 12, "name": "Adventure"}, {"id": 14, "...
2       [{"id": 28, "name": "Action"}, {"id": 12, "nam...
3       [{"id": 28, "name": "Action"}, {"id": 80, "nam...
4       [{"id": 28, "name": "Action"}, {"id": 12, "nam...
                              ...
4804    [{"id": 28, "name": "Action"}, {"id": 80, "nam...
4805    [{"id": 35, "name": "Comedy"}, {"id": 10749, "...
4806    [{"id": 35, "name": "Comedy"}, {"id": 18, "nam...
4807                                                   []
4808                    [{"id": 99, "name": "Documentary"}]
Name: genres, Length: 4809, dtype: object
```

In [19]:
```
df['genres'][0]
```

Out[19]:
```
'[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 8
78, "name": "Science Fiction"}]'
```

In [20]:
```
ast.literal_eval(df['genres'][0])
```

Out[20]:
```
[{'id': 28, 'name': 'Action'},
 {'id': 12, 'name': 'Adventure'},
 {'id': 14, 'name': 'Fantasy'},
 {'id': 878, 'name': 'Science Fiction'}]
```

In [21]:
```
for i in ast.literal_eval(df['genres'][0]):
    print(i['name'])
```

```
Action
Adventure
Fantasy
Science Fiction
```

In [ ]:
```

```

In [22]:
```
def fetch_genres(data):

    l = []

    for i in ast.literal_eval(data):
        l.append(i['name'])

    return l
```

In [23]:
```
fetch_genres(df['genres'][0])
```

Out[23]:
```
['Action', 'Adventure', 'Fantasy', 'Science Fiction']
```

In [24]:
```
df['genres'].apply(fetch_genres)
```

Out[24]:
```
0       [Action, Adventure, Fantasy, Science Fiction]
1                         [Adventure, Fantasy, Action]
2                           [Action, Adventure, Crime]
3                    [Action, Crime, Drama, Thriller]
4                 [Action, Adventure, Science Fiction]
                              ...
4804                        [Action, Crime, Thriller]
4805                                [Comedy, Romance]
4806               [Comedy, Drama, Romance, TV Movie]
4807                                                []
4808                                     [Documentary]
Name: genres, Length: 4809, dtype: object
```

In [ ]:
```

```

In [25]:
```
df[:3]
```

Out[25]:

	movie_id	title	overview	genres	keywords	cast	crew
0	19995	Avatar	In the 22nd century, a paraplegic Marine is di...	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	[{"id": 1463, "name": "culture clash"}, {"id":...	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
1	285	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	[{"cast_id": 4, "character": "Captain Jack Spa...	[{"credit_id": "52fe4232c3a36847f800b579", "de...
2	206647	Spectre	A cryptic message from Bond’s past sends him o...	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	[{"id": 470, "name": "spy"}, {"id": 818, "name...	[{"cast_id": 1, "character": "James Bond", "cr...	[{"credit_id": "54805967c3a36829b5002c41", "de...

In [26]:
```
df['genres'] = df['genres'].apply(fetch_genres)
```

In [27]:
```
df[:3]
```

Out[27]:

	movie_id	title	overview	genres	keywords	cast	crew
0	19995	Avatar	In the 22nd century, a paraplegic Marine is di...	[Action, Adventure, Fantasy, Science Fiction]	[{"id": 1463, "name": "culture clash"}, {"id":...	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
1	285	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	[Adventure, Fantasy, Action]	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	[{"cast_id": 4, "character": "Captain Jack Spa...	[{"credit_id": "52fe4232c3a36847f800b579", "de...
2	206647	Spectre	A cryptic message from Bond’s past sends him o...	[Action, Adventure, Crime]	[{"id": 470, "name": "spy"}, {"id": 818, "name...	[{"cast_id": 1, "character": "James Bond", "cr...	[{"credit_id": "54805967c3a36829b5002c41", "de...

In [ ]:
```

```

lets go with Keywords

In [28]:
```
df['keywords']
```

Out[28]:
```
0       [{"id": 1463, "name": "culture clash"}, {"id":...
1       [{"id": 270, "name": "ocean"}, {"id": 726, "na...
2       [{"id": 470, "name": "spy"}, {"id": 818, "name...
3       [{"id": 849, "name": "dc comics"}, {"id": 853,...
4       [{"id": 818, "name": "based on novel"}, {"id":...
                              ...
4804    [{"id": 5616, "name": "united states\u2013mexi...
4805                                                   []
4806    [{"id": 248, "name": "date"}, {"id": 699, "nam...
4807                                                   []
4808    [{"id": 1523, "name": "obsession"}, {"id": 224...
Name: keywords, Length: 4809, dtype: object
```

In [29]:
```
df['keywords'][0]
```

Out[29]:
```
'[{"id": 1463, "name": "culture clash"}, {"id": 2964, "name": "future"}, {"id": 3386, "name": "space wa
r"}, {"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": "spa
ce"}, {"id": 9951, "name": "alien"}, {"id": 10148, "name": "tribe"}, {"id": 10158, "name": "alien plane
t"}, {"id": 10987, "name": "cgi"}, {"id": 11399, "name": "marine"}, {"id": 13065, "name": "soldier"},
{"id": 14643, "name": "battle"}, {"id": 14720, "name": "love affair"}, {"id": 165431, "name": "anti wa
r"}, {"id": 193554, "name": "power relations"}, {"id": 206690, "name": "mind and soul"}, {"id": 209714,
"name": "3d"}]'
```

In [30]:
```
ast.literal_eval(df['keywords'][0])
```

Out[30]:
```
[{'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, 'name': 'alien planet'},
 {'id': 10987, 'name': 'cgi'},
 {'id': 11399, 'name': 'marine'},
 {'id': 13065, 'name': 'soldier'},
 {'id': 14643, 'name': 'battle'},
 {'id': 14720, 'name': 'love affair'},
 {'id': 165431, 'name': 'anti war'},
 {'id': 193554, 'name': 'power relations'},
 {'id': 206690, 'name': 'mind and soul'},
 {'id': 209714, 'name': '3d'}]
```

In [ ]:
```
```

In [31]:
```
def fetch_keywords(data):

    l = []

    for i in ast.literal_eval(data):
        l.append(i['name'])

    return l
```

In [32]:
```
fetch_keywords(df['keywords'][0])
```

Out[32]:
```
['culture clash',
 'future',
 'space war',
 'space colony',
 'society',
 'space travel',
 'futuristic',
 'romance',
 'space',
 'alien',
 'tribe',
 'alien planet',
 'cgi',
 'marine',
 'soldier',
 'battle',
 'love affair',
 'anti war',
 'power relations',
 'mind and soul',
 '3d']
```

In [33]:
```
df['keywords'].apply(fetch_keywords)
```

Out[33]:
```
0       [culture clash, future, space war, space colon...
1       [ocean, drug abuse, exotic island, east india ...
2       [spy, based on novel, secret agent, sequel, mi...
3       [dc comics, crime fighter, terrorist, secret i...
4       [based on novel, mars, medallion, space travel...
                              ...
4804    [united states-mexico barrier, legs, arms, pap...
4805                                                   []
4806    [date, love at first sight, narration, investi...
4807                                                   []
4808              [obsession, camcorder, crush, dream girl]
Name: keywords, Length: 4809, dtype: object
```

In [34]:
```
df['keywords'] = df['keywords'].apply(fetch_keywords)
```

In [35]:
```
df[:3]
```

Out[35]:

	movie_id	title	overview	genres	keywords	cast	crew
0	19995	Avatar	In the 22nd century, a paraplegic Marine is di...	[Action, Adventure, Fantasy, Science Fiction]	[culture clash, future, space war, space colon...	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
1	285	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	[Adventure, Fantasy, Action]	[ocean, drug abuse, exotic island, east india ...	[{"cast_id": 4, "character": "Captain Jack Spa...	[{"credit_id": "52fe4232c3a36847f800b579", "de...
2	206647	Spectre	A cryptic message from Bond’s past sends him o...	[Action, Adventure, Crime]	[spy, based on novel, secret agent, sequel, mi...	[{"cast_id": 1, "character": "James Bond", "cr...	[{"credit_id": "54805967c3a36829b5002c41", "de...

In [ ]:
```
```

Cast

In [36]:
```
df['cast']
```

Out[36]:
```
0       [{"cast_id": 242, "character": "Jake Sully", "...
1       [{"cast_id": 4, "character": "Captain Jack Spa...
2       [{"cast_id": 1, "character": "James Bond", "cr...
3       [{"cast_id": 2, "character": "Bruce Wayne / Ba...
4       [{"cast_id": 5, "character": "John Carter", "c...
                              ...
4804    [{"cast_id": 1, "character": "El Mariachi", "c...
4805    [{"cast_id": 1, "character": "Buzzy", "credit_...
4806    [{"cast_id": 8, "character": "Oliver O\u2019To...
4807    [{"cast_id": 3, "character": "Sam", "credit_id...
4808    [{"cast_id": 3, "character": "Herself", "credi...
Name: cast, Length: 4809, dtype: object
```

In [37]:
```
df['cast'][0]
```

Out[37]:
'[{"cast_id": 242, "character": "Jake Sully", "credit_id": "5602a8a7c3a3685532001c9a", "gender": 2, "id": 65731, "name": "Sam Worthington", "order": 0}, {"cast_id": 3, "character": "Neytiri", "credit_id": "52fe48009251416c750ac9cb", "gender": 1, "id": 8691, "name": "Zoe Saldana", "order": 1}, {"cast_id": 25, "character": "Dr. Grace Augustine", "credit_id": "52fe48009251416c750aca39", "gender": 1, "id": 10205, "name": "Sigourney Weaver", "order": 2}, {"cast_id": 4, "character": "Col. Quaritch", "credit_id": "52fe48009251416c750ac9cf", "gender": 2, "id": 32747, "name": "Stephen Lang", "order": 3}, {"cast_id": 5, "character": "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 Spellman", "credit_id": "52fe48009251416c750ac9dd", "gender": 2, "id": 59231, "name": "Joel David Moore", "order": 6}, {"cast_id": 9, "character": "Moat", "credit_id": "52fe48009251416c750ac9e5", "gender": 1, "id": 30485, "name": "CCH Pounder", "order": 7}, {"cast_id": 11, "character": "Eytukan", "credit_id": "52fe48009251416c750ac9ed", "gender": 2, "id": 15853, "name": "Wes Studi", "order": 8}, {"cast_id": 10, "character": "Tsu\'Tey", "credit_id": "52fe48009251416c750ac9e9", "gender": 2, "id": 10964, "name": "Laz Alonso", "order": 9}, {"cast_id": 12, "character": "Dr. Max Patel", "credit_id": "52fe48009251416c750ac9f1", "gender": 2, "id": 95697, "name": "Dileep Rao", "order": 10}, {"cast_id": 13, "character": "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": 154153, "name": "Sean Anthony Moran", "order": 12}, {"cas

In [38]:
```
ast.literal_eval(df['cast'][0])
```

Out[38]:
```
[{'cast_id': 242,
  'character': 'Jake Sully',
  'credit_id': '5602a8a7c3a3685532001c9a',
  'gender': 2,
  'id': 65731,
  'name': 'Sam Worthington',
  'order': 0},
 {'cast_id': 3,
  'character': 'Neytiri',
  'credit_id': '52fe48009251416c750ac9cb',
  'gender': 1,
  'id': 8691,
  'name': 'Zoe Saldana',
  'order': 1},
 {'cast_id': 25,
  'character': 'Dr. Grace Augustine',
  'credit_id': '52fe48009251416c750aca39',
  'gender': 1,
  'id': 10205,
```

In [ ]:
```
```

In [39]:
```
def fetch_cast(data):

    l = []
    counter = 0

    for i in ast.literal_eval(data):
        if counter != 3:
            l.append(i['name'])
            counter += 1
        else:
            break

    return l
```

In [40]:
```
fetch_cast(df['cast'][0])
```

Out[40]:
```
['Sam Worthington', 'Zoe Saldana', 'Sigourney Weaver']
```

In [41]:
```
df['cast'].apply(fetch_cast)
```

Out[41]:
```
0        [Sam Worthington, Zoe Saldana, Sigourney Weaver]
1           [Johnny Depp, Orlando Bloom, Keira Knightley]
2            [Daniel Craig, Christoph Waltz, Léa Seydoux]
3            [Christian Bale, Michael Caine, Gary Oldman]
4          [Taylor Kitsch, Lynn Collins, Samantha Morton]
                              ...
4804    [Carlos Gallardo, Jaime de Hoyos, Peter Marqua...
4805         [Edward Burns, Kerry Bishé, Marsha Dietlein]
4806           [Eric Mabius, Kristin Booth, Crystal Lowe]
4807            [Daniel Henney, Eliza Coupe, Bill Paxton]
4808    [Drew Barrymore, Brian Herzlinger, Corey Feldman]
Name: cast, Length: 4809, dtype: object
```

In [42]:
```
df['cast'] = df['cast'].apply(fetch_cast)
```

In [43]:
```
df[:3]
```

Out[43]:

	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]	[{"credit_id": "52fe48009251416c750aca23", "de...
1	285	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	[Adventure, Fantasy, Action]	[ocean, drug abuse, exotic island, east india ...	[Johnny Depp, Orlando Bloom, Keira Knightley]	[{"credit_id": "52fe4232c3a36847f800b579", "de...
2	206647	Spectre	A cryptic message from Bond’s past sends him o...	[Action, Adventure, Crime]	[spy, based on novel, secret agent, sequel, mi...	[Daniel Craig, Christoph Waltz, Léa Seydoux]	[{"credit_id": "54805967c3a36829b5002c41", "de...

In [ ]:

Crew

In [44]:
```
df['crew']
```

Out[44]:
```
0       [{"credit_id": "52fe48009251416c750aca23", "de...
1       [{"credit_id": "52fe4232c3a36847f800b579", "de...
2       [{"credit_id": "54805967c3a36829b5002c41", "de...
3       [{"credit_id": "52fe4781c3a36847f81398c3", "de...
4       [{"credit_id": "52fe479ac3a36847f813eaa3", "de...
                              ...
4804    [{"credit_id": "52fe44eec3a36847f80b280b", "de...
4805    [{"credit_id": "52fe487dc3a368484e0fb013", "de...
4806    [{"credit_id": "52fe4df3c3a36847f8275ecf", "de...
4807    [{"credit_id": "52fe4ad9c3a368484e16a36b", "de...
4808    [{"credit_id": "58ce021b9251415a390165d9", "de...
Name: crew, Length: 4809, dtype: object
```

In [45]:
```
df['crew'][0]
```

Out[45]:
'[{"credit_id": "52fe48009251416c750aca23", "department": "Editing", "gender": 0, "id": 1721, "job": "Editor", "name": "Stephen E. Rivkin"}, {"credit_id": "539c47ecc3a36810e3001f87", "department": "Art", "gender": 2, "id": 496, "job": "Production Design", "name": "Rick Carter"}, {"credit_id": "54491c89c3a3680fb4001cf7", "department": "Sound", "gender": 0, "id": 900, "job": "Sound Designer", "name": "Christopher Boyes"}, {"credit_id": "54491cb70e0a267480001bd0", "department": "Sound", "gender": 0, "id": 900, "job": "Supervising Sound Editor", "name": "Christopher Boyes"}, {"credit_id": "539c4a4cc3a36810c9002101", "department": "Production", "gender": 1, "id": 1262, "job": "Casting", "name": "Mali Finn"}, {"credit_id": "5544ee3b925141499f0008fc", "department": "Sound", "gender": 2, "id": 1729, "job": "Original Music Composer", "name": "James Horner"}, {"credit_id": "52fe48009251416c750ac9c3", "department": "Directing", "gender": 2, "id": 2710, "job": "Director", "name": "James Cameron"}, {"credit_id": "52fe48009251416c750ac9d9", "department": "Writing", "gender": 2, "id": 2710, "job": "Writer", "name": "James Cameron"}, {"credit_id": "52fe48009251416c750aca17", "department": "Editing", "gender": 2, "id": 2710, "job": "Editor", "name": "James Cameron"}, {"credit_id": "52fe48009251416c750aca29", "department": "Production", "gender": 2, "id": 2710, "job": "Producer", "name": "James Cameron"}, {"credit_id": "52fe48009251416c750aca3f", "department": "Writing", "gender": 2, "id": 2710, "job": "Screenplay", "name": "James Cameron"}, {"credit_id": "539c4987c3a36810ba0021a4", "department": "Art", "gender": 2, "id": 7236, "job": "Art Direction", "name": "Andrew Menzies"}, {"credit_id": "549598c3c3a3686ae9004383", "department": "Visual Effects", "gender": 0, "id": 6690, "job": "Visual Effects Producer", "name": "Jill Brooks"}, {"credit_id": "52fe48009251416c750aca4b", "department": "Production", "gender": 1, "id": 6347,

In [ ]:
```
```

In [46]:
```
def fetch_director(data):

    l = []

    for i in ast.literal_eval(data):
        if i['job'] == 'Director':
            l.append(i['name'])

    return l
```

In [47]:
```
fetch_director(df['crew'][0])
```

Out[47]:
```
['James Cameron']
```

In [ ]:

In [48]:
```
df['crew'].apply(fetch_director)
```

Out[48]:
```
0                               [James Cameron]
1                             [Gore Verbinski]
2                                 [Sam Mendes]
3                            [Christopher Nolan]
4                               [Andrew Stanton]
                              ...
4804                          [Robert Rodriguez]
4805                              [Edward Burns]
4806                               [Scott Smith]
4807                               [Daniel Hsia]
4808    [Brian Herzlinger, Jon Gunn, Brett Winn]
Name: crew, Length: 4809, dtype: object
```

In [ ]:

In [49]:
```
df['crew'] = df['crew'].apply(fetch_director)
```

In [50]:
```
df.head()
```

Out[50]:

	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 [ ]:
```

```

Overview

In [51]:
```
df['overview']
```

Out[51]:
```
0       In the 22nd century, a paraplegic Marine is di...
1       Captain Barbossa, long believed to be dead, ha...
2       A cryptic message from Bond’s past sends him o...
3       Following the death of District Attorney Harve...
4       John Carter is a war-weary, former military ca...
                              ...
4804    El Mariachi just wants to play his guitar and ...
4805    A newlywed couple's honeymoon is upended by th...
4806    "Signed, Sealed, Delivered" introduces a dedic...
4807    When ambitious New York attorney Sam is sent t...
4808    Ever since the second grade when he first saw ...
Name: overview, Length: 4809, dtype: object
```

In [52]:
```
df['overview'][0]
```

Out[52]:
```
'In the 22nd century, a paraplegic Marine is dispatched to the moon Pandora on a unique mission, but bec
omes torn between following orders and protecting an alien civilization.'
```

In [54]:
```
'i am a boy'.split()
```

Out[54]:
```
['i', 'am', 'a', 'boy']
```

In [55]:
```
print(df['overview'][0].split())
```

```
['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 [57]:
```
df.isna().sum()
```

Out[57]:
```
movie_id    0
title       0
overview    3
genres      0
keywords    0
cast        0
crew        0
dtype: int64
```

In [58]:
```
df.dropna(inplace=True)
```

In [59]:
```
df.isna().sum()
```

Out[59]:
```
movie_id    0
title       0
overview    0
genres      0
keywords    0
cast        0
crew        0
dtype: int64
```

In [60]:
```
df['overview'].apply(lambda x: x.split())
```

Out[60]:
```
0       [In, the, 22nd, century,, a, paraplegic, Marin...
1       [Captain, Barbossa,, long, believed, to, be, d...
2       [A, cryptic, message, from, Bond’s, past, send...
3       [Following, the, death, of, District, Attorney...
4       [John, Carter, is, a, war-weary,, former, mili...
                              ...
4804    [El, Mariachi, just, wants, to, play, his, gui...
4805    [A, newlywed, couple's, honeymoon, is, upended...
4806    ["Signed,, Sealed,, Delivered", introduces, a,...
4807    [When, ambitious, New, York, attorney, Sam, is...
4808    [Ever, since, the, second, grade, when, he, fi...
Name: overview, Length: 4806, dtype: object
```

In [61]:
```
df['overview'] = df['overview'].apply(lambda x: x.split())
```

In [63]:
```
df.head(3)
```

Out[63]:

	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]

In [64]:
```
'i am a boy'
```

Out[64]:
```
'i am a boy'
```

In [65]:
```
'i am a boy'.replace(' ', '')
```

Out[65]:
```
'iamaboy'
```

In [71]:
```
df['cast'].apply(lambda x: [i.replace(' ', '') for i in x])
```

Out[71]:
```
0        [SamWorthington, ZoeSaldana, SigourneyWeaver]
1           [JohnnyDepp, OrlandoBloom, KeiraKnightley]
2            [DanielCraig, ChristophWaltz, LéaSeydoux]
3            [ChristianBale, MichaelCaine, GaryOldman]
4          [TaylorKitsch, LynnCollins, SamanthaMorton]
                              ...
4804    [CarlosGallardo, JaimedeHoyos, PeterMarquardt]
4805         [EdwardBurns, KerryBishé, MarshaDietlein]
4806           [EricMabius, KristinBooth, CrystalLowe]
4807            [DanielHenney, ElizaCoupe, BillPaxton]
4808    [DrewBarrymore, BrianHerzlinger, CoreyFeldman]
Name: cast, Length: 4806, dtype: object
```

In [ ]:
```

```

In [72]:
```
df['overview'] = df['overview'].apply(lambda x: [i.replace(' ', '') for i in x])
df['genres'] = df['genres'].apply(lambda x: [i.replace(' ', '') for i in x])
df['keywords'] = df['keywords'].apply(lambda x: [i.replace(' ', '') for i in x])
df['cast'] = df['cast'].apply(lambda x: [i.replace(' ', '') for i in x])
df['crew'] = df['crew'].apply(lambda x: [i.replace(' ', '') for i in x])
```

In [ ]:
```

```

In [73]:
```
df[:3]
```

Out[73]:

	movie_id	title	overview	genres	keywords	cast	crew
0	19995	Avatar	[In, the, 22nd, century,, a, paraplegic, Marin...	[Action, Adventure, Fantasy, ScienceFiction]	[cultureclash, future, spacewar, spacecolony, ...	[SamWorthington, ZoeSaldana, SigourneyWeaver]	[JamesCameron]
1	285	Pirates of the Caribbean: At World's End	[Captain, Barbossa,, long, believed, to, be, d...	[Adventure, Fantasy, Action]	[ocean, drugabuse, exoticisland, eastindiatrad...	[JohnnyDepp, OrlandoBloom, KeiraKnightley]	[GoreVerbinski]
2	206647	Spectre	[A, cryptic, message, from, Bond’s, past, send...	[Action, Adventure, Crime]	[spy, basedonnovel, secretagent, sequel, mi6, ...	[DanielCraig, ChristophWaltz, LéaSeydoux]	[SamMendes]

In [ ]:
```

```

In [74]:
```
[4] + [5]
```

Out[74]:
```
[4, 5]
```

In [75]:
```
['a'] + ['b']
```

Out[75]:
```
['a', 'b']
```

In [76]:
```
['a'] + [5]
```

Out[76]:
```
['a', 5]
```

In [77]:
```
df.columns
```

Out[77]:
```
Index(['movie_id', 'title', 'overview', 'genres', 'keywords', 'cast', 'crew'], dtype='object')
```

In [79]:
```
df['tags'] = df['overview'] + df['genres'] + df['keywords'] + df['cast'] + df['crew']
```

In [80]:
```
df[:3]
```

Out[80]:

	movie_id	title	overview	genres	keywords	cast	crew	tags
0	19995	Avatar	[In, the, 22nd, century,, a, paraplegic, Marin...	[Action, Adventure, Fantasy, ScienceFiction]	[cultureclash, future, spacewar, spacecolony, ...	[SamWorthington, ZoeSaldana, SigourneyWeaver]	[JamesCameron]	[In, the, 22nd, century,, a, paraplegic, Marin...
1	285	Pirates of the Caribbean: At World's End	[Captain, Barbossa,, long, believed, to, be, d...	[Adventure, Fantasy, Action]	[ocean, drugabuse, exoticisland, eastindiatrad...	[JohnnyDepp, OrlandoBloom, KeiraKnightley]	[GoreVerbinski]	[Captain, Barbossa,, long, believed, to, be, d...
2	206647	Spectre	[A, cryptic, message, from, Bond’s, past, send...	[Action, Adventure, Crime]	[spy, basedonnovel, secretagent, sequel, mi6, ...	[DanielCraig, ChristophWaltz, LéaSeydoux]	[SamMendes]	[A, cryptic, message, from, Bond’s, past, send...

In [83]:
```
print(df['tags'][0])
```

```
['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',
'cultureclash', 'future', 'spacewar', 'spacecolony', 'society', 'spacetravel', 'futuristic', 'romance',
'space', 'alien', 'tribe', 'alienplanet', 'cgi', 'marine', 'soldier', 'battle', 'loveaffair', 'antiwar',
'powerrelations', 'mindandsoul', '3d', 'SamWorthington', 'ZoeSaldana', 'SigourneyWeaver', 'JamesCamero
n']
```

In [ ]:
```

```

Updated Dataframe

In [85]:
```
data = df[['movie_id', 'title', 'tags']]
data
```

Out[85]:

	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...
...	...	...	...
4804	9367	El Mariachi	[El, Mariachi, just, wants, to, play, his, gui...
4805	72766	Newlyweds	[A, newlywed, couple's, honeymoon, is, upended...
4806	231617	Signed, Sealed, Delivered	["Signed,, Sealed,, Delivered", introduces, a,...
4807	126186	Shanghai Calling	[When, ambitious, New, York, attorney, Sam, is...
4808	25975	My Date with Drew	[Ever, since, the, second, grade, when, he, fi...

4806 rows × 3 columns

In [ ]:
```

```

In [90]:
```
' '.join(df['tags'][0])
```

Out[90]:
```
'In the 22nd century, a paraplegic Marine is dispatched to the moon Pandora on a unique mission, but bec
omes torn between following orders and protecting an alien civilization. Action Adventure Fantasy Scienc
eFiction cultureclash future spacewar spacecolony society spacetravel futuristic romance space alien tri
be alienplanet cgi marine soldier battle loveaffair antiwar powerrelations mindandsoul 3d SamWorthington
ZoeSaldana SigourneyWeaver JamesCameron'
```

In [91]:
```
data['tags'].apply(lambda x: ' '.join(x))
```

Out[91]:
```
0       In the 22nd century, a paraplegic Marine is di...
1       Captain Barbossa, long believed to be dead, ha...
2       A cryptic message from Bond’s past sends him o...
3       Following the death of District Attorney Harve...
4       John Carter is a war-weary, former military ca...
                              ...
4804    El Mariachi just wants to play his guitar and ...
4805    A newlywed couple's honeymoon is upended by th...
4806    "Signed, Sealed, Delivered" introduces a dedic...
4807    When ambitious New York attorney Sam is sent t...
4808    Ever since the second grade when he first saw ...
Name: tags, Length: 4806, dtype: object
```

In [92]:
```
data['tags'] = data['tags'].apply(lambda x: ' '.join(x))
```

In [93]:
```
data
```

Out[93]:

	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...
...	...	...	...
4804	9367	El Mariachi	El Mariachi just wants to play his guitar and ...
4805	72766	Newlyweds	A newlywed couple's honeymoon is upended by th...
4806	231617	Signed, Sealed, Delivered	"Signed, Sealed, Delivered" introduces a dedic...
4807	126186	Shanghai Calling	When ambitious New York attorney Sam is sent t...
4808	25975	My Date with Drew	Ever since the second grade when he first saw ...

4806 rows × 3 columns

In [ ]:

Now will apply NLP concepts to preprocess textual data

In [ ]:
```
# i, to, the, is , am, a .... (stopwords)
```

In [ ]:
```
i go to school
```

In [ ]:
```
go to school
```

In [ ]:
```
go school
```

In [ ]:
```
# Lower Case
# Tokenization
# Stopwords Removal
# Stemming
# Word Embedding - (Text to Vector or Number)
```

In [94]:
```
data.head(4)
```

Out[94]:

	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...

In [100]:
```
data['tags'][0]
```

Out[100]:
```
'In the 22nd century, a paraplegic Marine is dispatched to the moon Pandora on a unique mission, but bec
omes torn between following orders and protecting an alien civilization. Action Adventure Fantasy Scienc
eFiction cultureclash future spacewar spacecolony society spacetravel futuristic romance space alien tri
be alienplanet cgi marine soldier battle loveaffair antiwar powerrelations mindandsoul 3d SamWorthington
ZoeSaldana SigourneyWeaver JamesCameron'
```

In [96]:
```
import nltk
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
```

In [109]:
```
ps = PorterStemmer()

def preprocess_data(text):

    stemmed = []

    for i in text.split():
        lower = i.lower()
        stemmed.append(ps.stem(lower))

    return ' '.join(stemmed)
```

In [113]:
```
preprocess_data(data['tags'][0])
```

Out[113]:
```
'in the 22nd century, a parapleg marin is dispatch to the moon pandora on a uniqu mission, but becom tor
n between follow order and protect an alien civilization. action adventur fantasi sciencefict culturecla
sh futur spacewar spacecoloni societi spacetravel futurist romanc space alien tribe alienplanet cgi mari
n soldier battl loveaffair antiwar powerrel mindandsoul 3d samworthington zoesaldana sigourneyweav james
cameron'
```

In [115]:
```
data['tags'].apply(preprocess_data)
```

Out[115]:
```
0       in the 22nd century, a parapleg marin is dispa...
1       captain barbossa, long believ to be dead, ha c...
2       a cryptic messag from bond’ past send him on a...
3       follow the death of district attorney harvey d...
4       john carter is a war-weary, former militari ca...
                              ...
4804    el mariachi just want to play hi guitar and ca...
4805    a newlyw couple' honeymoon is upend by the arr...
4806    "signed, sealed, delivered" introduc a dedic q...
4807    when ambiti new york attorney sam is sent to s...
4808    ever sinc the second grade when he first saw h...
Name: tags, Length: 4806, dtype: object
```

In [116]:
```
data['tags'] = data['tags'].apply(preprocess_data)
```

In [117]:
```
data.head()
```

Out[117]:

	movie_id	title	tags
0	19995	Avatar	in the 22nd century, a parapleg marin is dispa...
1	285	Pirates of the Caribbean: At World's End	captain barbossa, long believ to be dead, ha c...
2	206647	Spectre	a cryptic messag from bond’ past send him on a...
3	49026	The Dark Knight Rises	follow the death of district attorney harvey d...
4	49529	John Carter	john carter is a war-weary, former militari ca...

In [218]:
```
data.to_dict()
```

Out[218]:
```
{'movie_id': {0: 19995,
  1: 285,
  2: 206647,
  3: 49026,
  4: 49529,
  5: 559,
  6: 38757,
  7: 99861,
  8: 767,
  9: 209112,
  10: 1452,
  11: 10764,
  12: 58,
  13: 57201,
  14: 49521,
  15: 2454,
  16: 24428,
  17: 1865,
  18: 41154,
```

In [ ]:

In [ ]:

Use Bag of Words (BOW) to encode your data into vector form

In [123]:
```
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(max_features=5000, stop_words='english')
```

In [122]:
```
data['tags']
```

Out[122]:
```
0       in the 22nd century, a parapleg marin is dispa...
1       captain barbossa, long believ to be dead, ha c...
2       a cryptic messag from bond’ past send him on a...
3       follow the death of district attorney harvey d...
4       john carter is a war-weary, former militari ca...
                              ...
4804    el mariachi just want to play hi guitar and ca...
4805    a newlyw couple' honeymoon is upend by the arr...
4806    "signed, sealed, delivered" introduc a dedic q...
4807    when ambiti new york attorney sam is sent to s...
4808    ever sinc the second grade when he first saw h...
Name: tags, Length: 4806, dtype: object
```

In [125]:
```
vectors = cv.fit_transform(data['tags']).toarray()
vectors
```

Out[125]:
```
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 [126]:
```
len(vectors)
```

Out[126]:
```
4806
```

In [127]:
```
vectors.shape
```

Out[127]:
```
(4806, 5000)
```

In [130]:
```
# for i in vectors[0]:
#     print(i)
```

In [ ]:
```

```

In [131]:
```
cv.get_feature_names_out()
```

Out[131]:
```
array(['000', '007', '10', ..., 'zone', 'zoo', 'zooeydeschanel'],
      dtype=object)
```

In [132]:
```
for i in cv.get_feature_names_out():
    print(i)
```

```
000
007
10
100
11
12
13
14
15
16
17
17th
18
18th
18thcenturi
19
1910
1920
1930
```

In [ ]:

In [ ]:

Calculate Cosine Similarity between vectors

In [135]:
```
from sklearn.metrics.pairwise import cosine_similarity
```

In [136]:
```
vectors
```

Out[136]:
```
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 [137]:
```
len(vectors)
```

Out[137]:
```
4806
```

In [138]:
```
similarity = cosine_similarity(vectors)
similarity
```

Out[138]:
```
array([[1.        , 0.08346223, 0.0860309 , ..., 0.04499213, 0.        ,
        0.        ],
       [0.08346223, 1.        , 0.06063391, ..., 0.02378257, 0.        ,
        0.02615329],
       [0.0860309 , 0.06063391, 1.        , ..., 0.02451452, 0.        ,
        0.        ],
       ...,
       [0.04499213, 0.02378257, 0.02451452, ..., 1.        , 0.03962144,
        0.04229549],
       [0.        , 0.        , 0.        , ..., 0.03962144, 1.        ,
        0.08714204],
       [0.        , 0.02615329, 0.        , ..., 0.04229549, 0.08714204,
        1.        ]])
```

In [142]:
```
similarity[0]
```

Out[142]:
```
array([1.        , 0.08346223, 0.0860309 , ..., 0.04499213, 0.        ,
       0.        ])
```

In [140]:
```
data[:3]
```

Out[140]:

	movie_id	title	tags
0	19995	Avatar	in the 22nd century, a parapleg marin is dispa...
1	285	Pirates of the Caribbean: At World's End	captain barbossa, long believ to be dead, ha c...
2	206647	Spectre	a cryptic messag from bond’ past send him on a...

In [143]:
```
len(similarity)
```

Out[143]:
```
4806
```

In [144]:
```
similarity.shape
```

Out[144]:
```
(4806, 4806)
```

In [ ]:

Distance between vectors

In [145]:
```
distance = similarity[0]
distance
```

Out[145]:
```
array([1.        , 0.08346223, 0.0860309 , ..., 0.04499213, 0.        ,
       0.        ])
```

In [146]:
```
for i in distance:
    print(i)
```

```
1.0000000000000002
0.08346223261119858
0.08603090020146065
0.0734718358370645
0.1892994097121204
0.10838874619051501
0.04024218182927669
0.14673479641335554
0.05923488777590923
0.0967301666813349
0.10259783520851541
0.09464970485606021
0.09037128496931669
0.04499212706658476
0.12824729401064427
0.06282808624375433
0.07894736842105264
0.13977653617040256
0.09493290614465533
```

In [ ]:

In [151]:
```
sorted(similarity[0])
```

Out[151]:
```
[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,
 0.0,
 0.0,
```

In [152]:
```
sorted(similarity[0])[-10:]
```

Out[152]:
```
[0.23174488732966075,
 0.23179316248638276,
 0.24455799402225922,
 0.24511108480187255,
 0.25038669783359574,
 0.255608593705383,
 0.2605130246476754,
 0.26901379342448517,
 0.28676966733820225,
 1.0000000000000002]
```

In [154]:
```
sorted(similarity[0], reverse=True)
```

Out[154]:
```
[1.0000000000000002,
 0.28676966733820225,
 0.26901379342448517,
 0.2605130246476754,
 0.255608593705383,
 0.25038669783359574,
 0.24511108480187255,
 0.24455799402225922,
 0.23179316248638276,
 0.23174488732966075,
 0.2278389747471728,
 0.2252817784447915,
 0.22269966704152225,
 0.21853668936906193,
 0.21239769762143662,
 0.2108663315950723,
 0.2105263157894737,
 0.20443988269091456,
 0.20437977982832192,
```

In [ ]:

In [157]:
```
list(enumerate(similarity[0]))
```

Out[157]:
```
[(0, 1.0000000000000002),
 (1, 0.08346223261119858),
 (2, 0.08603090020146065),
 (3, 0.0734718358370645),
 (4, 0.1892994097121204),
 (5, 0.10838874619051501),
 (6, 0.04024218182927669),
 (7, 0.14673479641335554),
 (8, 0.05923488777590923),
 (9, 0.0967301666813349),
 (10, 0.10259783520851541),
 (11, 0.09464970485606021),
 (12, 0.09037128496931669),
 (13, 0.04499212706658476),
 (14, 0.12824729401064427),
 (15, 0.06282808624375433),
 (16, 0.07894736842105264),
 (17, 0.13977653617040256),
 (18, 0.09493290614465533),
```

In [159]:
```
sorted(list(enumerate(similarity[0])), reverse=True)
```

Out[159]:
```
[(4805, 0.0),
 (4804, 0.0),
 (4803, 0.04499212706658476),
 (4802, 0.046829290579084706),
 (4801, 0.019252140716412975),
 (4800, 0.0),
 (4799, 0.052631578947368425),
 (4798, 0.04223886030955117),
 (4797, 0.0),
 (4796, 0.0),
 (4795, 0.0),
 (4794, 0.0),
 (4793, 0.05407380704358751),
 (4792, 0.0),
 (4791, 0.0),
 (4790, 0.0582716546748065),
 (4789, 0.060833032924035954),
 (4788, 0.0),
 (4787, 0.019117977822546817),
```

In [ ]:

In [161]:
```
sorted(list(enumerate(similarity[0])), reverse=True, key = lambda x: x[1])
```

Out[161]:
```
[(0, 1.0000000000000002),
 (1216, 0.28676966733820225),
 (2409, 0.26901379342448517),
 (3730, 0.2605130246476754),
 (507, 0.255608593705383),
 (539, 0.25038669783359574),
 (582, 0.24511108480187255),
 (1204, 0.24455799402225922),
 (1194, 0.23179316248638276),
 (778, 0.23174488732966075),
 (4048, 0.2278389747471728),
 (1920, 0.2252817784447915),
 (61, 0.22269966704152225),
 (2786, 0.21853668936906193),
 (172, 0.21239769762143662),
 (972, 0.2108663315950723),
 (322, 0.2105263157894737),
 (2333, 0.20443988269091456),
 (3608, 0.20437977982832192),
```

In [ ]:
```

```

In [163]:
```
sorted(list(enumerate(similarity[0])), reverse=True, key = lambda x: x[1])[1:6]
```

Out[163]:
```
[(1216, 0.28676966733820225),
 (2409, 0.26901379342448517),
 (3730, 0.2605130246476754),
 (507, 0.255608593705383),
 (539, 0.25038669783359574)]
```

In [ ]:
```

```

In [167]:
```
data.iloc[1216]
```

Out[167]:
```
movie_id                                                  440
title                                 Aliens vs Predator: Requiem
tags        a sequel to 2004' alien vs. predator, the icon...
Name: 1216, dtype: object
```

In [168]:
```
data.iloc[2409]
```

Out[168]:
```
movie_id                                                  679
title                                                  Aliens
tags        when ripley' lifepod is found by a salvag crew...
Name: 2409, dtype: object
```

In [ ]:
```

```

Final Recommendation Function

In [205]:
```
def recommend(movie):

    movie_index = data[data['title'] == movie].index[0]
    distance = similarity[movie_index]
    movie_list = sorted(list(enumerate(distance)), reverse=True, key = lambda x: x[1])[1:6]

    for i in movie_list:
#         print(i[0])
        print(data.iloc[i[0]].title)
```

In [206]:
```
recommend('Iron Man')
```

```
Iron Man 3
Iron Man 2
Avengers: Age of Ultron
The Avengers
Captain America: Civil War
```

In [207]:
```
recommend('Thor')
```

```
Thor: The Dark World
Clash of the Titans
After Earth
Iron Man 2
Ant-Man
```

In [208]:
```
recommend('Superman')
```

```
Superman Returns
Superman II
Iron Man 2
Superman III
Superman IV: The Quest for Peace
```

In [209]:
```
recommend('Spider-Man')
```

```
Spider-Man 3
Spider-Man 2
The Amazing Spider-Man 2
Arachnophobia
Kick-Ass
```

In [211]:
```
recommend('Casino Royale')
```

```
Grandma's Boy
The Adventures of Elmo in Grouchland
Imagine That
The Diving Bell and the Butterfly
True Lies
```

In [214]:
```
recommend('Batman')
```

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

In [215]:
```
recommend('X-Men')
```

```
X2
X-Men: The Last Stand
X-Men: Apocalypse
Iron Man 3
X-Men: First Class
```

In [ ]:

In [ ]:

In [ ]:

In [197]:
```
data.iloc[31]
```

Out[197]:
```
movie_id                                                68721
title                                              Iron Man 3
tags        when toni stark' world is torn apart by a form...
Name: 31, dtype: object
```

In [198]:
```
data.iloc[31].title
```

Out[198]:
```
'Iron Man 3'
```

In [ ]:

In [ ]:

In [173]:
```
data[data['title'] == 'Avatar']
```

Out[173]:

	movie_id	title	tags
0	19995	Avatar	in the 22nd century, a parapleg marin is dispa...

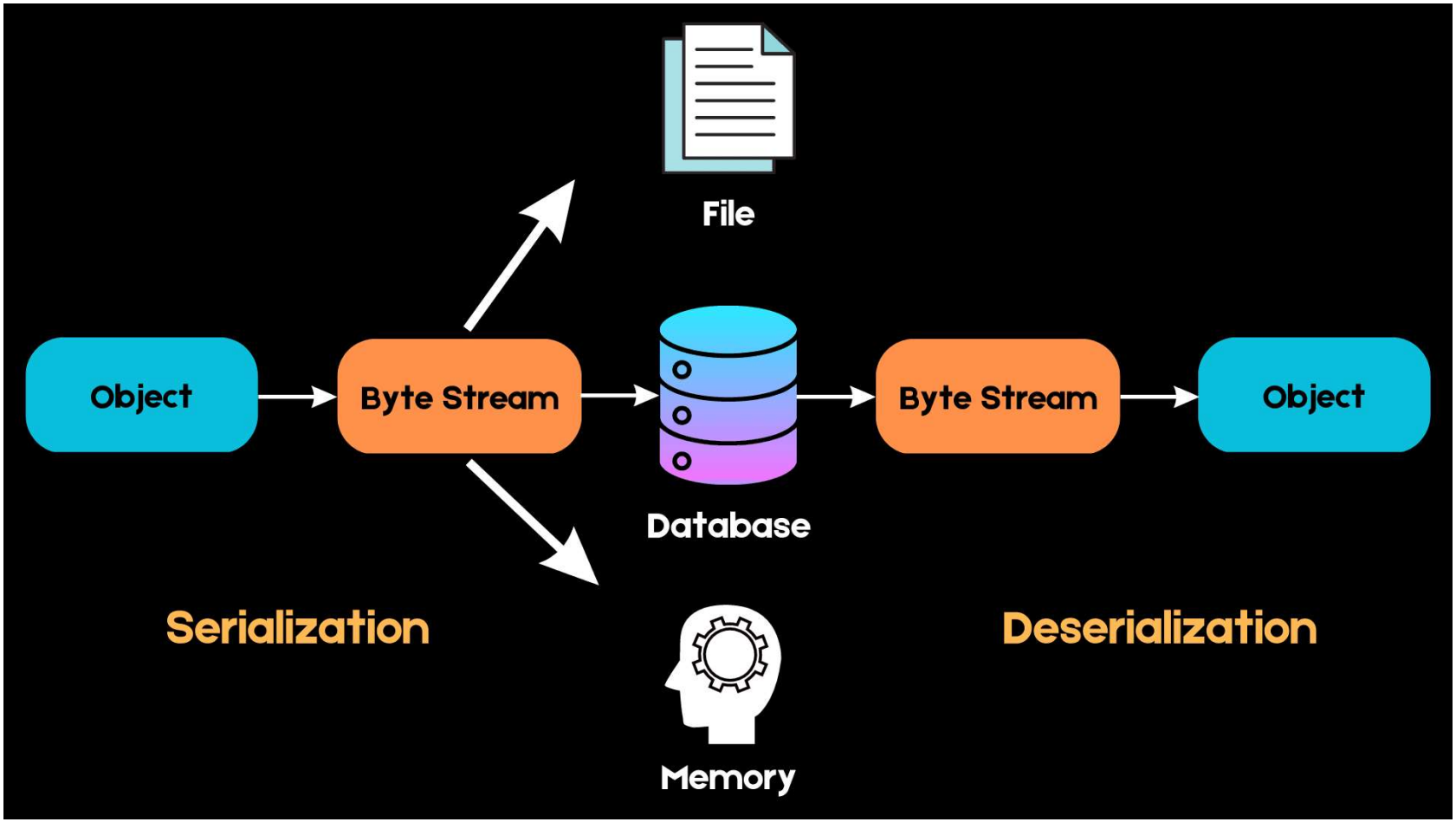
In [175]:
```
data[data['title'] == 'Avatar'].index[0]
```

Out[175]: 0

In [ ]:
```

```

In [219]:
```
import pickle
```



In [222]:
```
data.to_dict()
```

Out[222]:
```
{'movie_id': {0: 19995,
  1: 285,
  2: 206647,
  3: 49026,
  4: 49529,
  5: 559,
  6: 38757,
  7: 99861,
  8: 767,
  9: 209112,
  10: 1452,
  11: 10764,
  12: 58,
  13: 57201,
  14: 49521,
  15: 2454,
  16: 24428,
  17: 1865,
  18: 41154,
```

In [221]:
```
similarity
```

Out[221]:
```
array([[1.        , 0.08346223, 0.0860309 , ..., 0.04499213, 0.        ,
        0.        ],
       [0.08346223, 1.        , 0.06063391, ..., 0.02378257, 0.        ,
        0.02615329],
       [0.0860309 , 0.06063391, 1.        , ..., 0.02451452, 0.        ,
        0.        ],
       ...,
       [0.04499213, 0.02378257, 0.02451452, ..., 1.        , 0.03962144,
        0.04229549],
       [0.        , 0.        , 0.        , ..., 0.03962144, 1.        ,
        0.08714204],
       [0.        , 0.02615329, 0.        , ..., 0.04229549, 0.08714204,
        1.        ]])
```

In [ ]:

In [ ]:

In [226]:
```
pickle.dump(data.to_dict(), open('movie_dict.pkl', mode='wb'))
```

In [225]:
```
pickle.dump(similarity, open('similarity.pkl', mode='wb'))
```

In [ ]:

In [ ]:

In [228]:
```
data['title'].values
```

Out[228]:
```
array(['Avatar', "Pirates of the Caribbean: At World's End", 'Spectre',
       ..., 'Signed, Sealed, Delivered', 'Shanghai Calling',
       'My Date with Drew'], dtype=object)
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: