In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
from IPython.display import display
from sklearn.preprocessing import LabelEncoder, MinMaxScaler, StandardScaler
import pickle
```

In [2]:
```
import nltk
import re
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
```

In [3]:
```
df = pd.read_csv('kaggle_movie_train.csv')
```

In [4]:
```
le = LabelEncoder()

df['genre'] = le.fit_transform(df['genre'])
```

In [5]:
```
df.head(15)
```

Out[5]:

	id	text	genre
0	0	eady dead, maybe even wishing he was. INT. 2ND...	8
1	2	t, summa cum laude and all. And I'm about to l...	2
2	3	up Come, I have a surprise.... She takes him ...	3
3	4	ded by the two detectives. INT. JEFF'S APARTME...	8
4	5	nd dismounts, just as the other children reach...	3
5	6	breadth of the bluff. Gabe pulls out his ancie...	8
6	7	uilding. A MAN in pajamas runs out into the ra...	8
7	9	ELLES AND RITA HAYWORTH Just disgustingly rich...	3
8	10	Memphis goes back into the garage, Budgy cack...	8
9	11	e reels as the world spins. Sweat pours off hi...	0
10	12	ng nasty now. Percy starts to play an old reli...	3
11	14	maps you know. Swines. Before the war we help...	3
12	15	ng ones always break your heart. Donna nods, g...	3
13	16	ilm feeling to it. Andrew WHIPS AROUND to face...	3
14	17	ousse as JACK watches in silence. PARRY molds ...	3

In [6]:
```
df['genre'].value_counts()
```

Out[6]:
```
3    8873
8    6824
2    2941
0    2392
7     613
4     456
5     270
1     147
6      63
Name: genre, dtype: int64
```

In [7]:
```
df.drop('id', axis=1, inplace=True)
```

In [8]:
```
# Cleaning the text
corpus = []
ps = PorterStemmer()

for i in range(0, df.shape[0]):

  # Cleaning special character from the dialog/script
  dialog = re.sub(pattern='[^a-zA-Z]', repl=' ', string=df['text'][i])

  # Converting the entire dialog/script into lower case
  dialog = dialog.lower()

  # Tokenizing the dialog/script by words
  words = dialog.split()

  # Removing the stop words
  dialog_words = [word for word in words if word not in set(stopwords.words('english'))]

  # Stemming the words
  words = [ps.stem(word) for word in dialog_words]

  # Joining the stemmed words
  dialog = ' '.join(words)

  # Creating a corpus
  corpus.append(dialog)
```

In [9]:
```
# Creating the Bag of Words model
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(max_features=10000, ngram_range=(1,2))
X = cv.fit_transform(corpus).toarray()
y = df['genre'].values
```

In [10]:
```
# Model Building
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_state=0)
```

In [11]:
```
# Fitting Naive Bayes to the Training set
from sklearn.naive_bayes import MultinomialNB
nb_classifier = MultinomialNB(alpha=0.1)
nb_classifier.fit(X_train, y_train)
```

Out[11]:
```
MultinomialNB(alpha=0.1, class_prior=None, fit_prior=True)
```

In [13]:
```
# Creating a pickle file for the CountVectorizer
file = open('cv-transform.pkl', 'wb')
pickle.dump(cv,file)

# Creating a pickle file for the Multinomial Naive Bayes model
filename = open('movie-genre-model.pkl', 'wb')
pickle.dump(nb_classifier,filename)
```

In [14]:
```
file.close()
filename.close()
```