By Nitish Adhikari

Email id :nitishbuzzpro@gmail.com (mailto:nitishbuzzpro@gmail.com) (mailto:nitishbuzzpro@gmail.com (mailto:nitishbuzzpro@gmail.com)) , +91-9650740295

Linkedin : https://www.linkedin.com/in/nitish-adhikari-6b2350248 (https://www.linkedin.com/in/nitish-adhikari-6b2350248) (https://www.linkedin.com/in/nitish-adhikari-6b2350248 (https://www.linkedin.com/in/nitish-adhikari-6b2350248))

In [ ]:

Text Classification - Project

The purpose of the project is to create an NLP model to predict the reviews as positive or negetive

In [6]:
```
# Import libraries
import numpy as np
import pandas as pd
```

In [10]:
```
#read the dataframe
df = pd.read_csv('moviereviews.tsv', sep='\t')
```

In [11]:
```
df.head()
```

Out[11]:

	label	review
0	neg	how do films like mouse hunt get into theatres...
1	neg	some talented actresses are blessed with a dem...
2	pos	this has been an extraordinary year for austra...
3	pos	according to hollywood movies made in last few...
4	neg	my first press screening of 1998 and already i...

In [13]:
```
#len of the dataframe
len(df)
```

Out[13]:
```
2000
```

In [14]:
```
# check for missing values
df.isnull().sum()
```

Out[14]:
```
label      0
review    35
dtype: int64
```

In [15]:
```
#remove the missing values
df.dropna(inplace=True)
```

In [16]:
```
#recheck
df.isnull().sum()
```

Out[16]:
```
label     0
review    0
dtype: int64
```

mystring

In [20]:
```
#remove the empty strings/white space in reviews
blanks = []

#(index,label,review text)
for i,lb,rv in df.itertuples():
    if rv.isspace(): #to check if empty string
        blanks.append(i)
```

In [21]:
```
blanks
```

Out[21]:
```
[57,
 71,
 147,
 151,
 283,
 307,
 313,
 323,
 343,
 351,
 427,
 501,
 633,
 675,
 815,
 851,
 977,
 1079,
 1299,
 1455,
 1493,
 1525,
 1531,
 1763,
 1851,
 1905,
 1993]
```

In [22]:
```
#drop the empty strings at the index positiions
df.drop(blanks,inplace=True)
```

In [23]:
```
len(df)
```

Out[23]:
```
1938
```

In [28]:
```
#Split the data into train test
from sklearn.model_selection import train_test_split
```

In [29]:
```
X = df['review']
```

In [30]:
```
y = df['label']
```

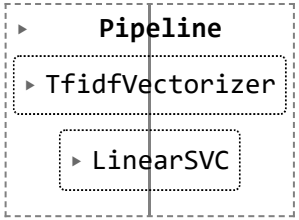
In [55]:
```
 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

In [56]:
```
#Using pipelines
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import LinearSVC
```

In [63]:
```
text_clf = Pipeline([('tfidf', TfidfVectorizer()),
                     ('clf',LinearSVC())])
```

In [58]:
```
text_clf.fit(X_train,y_train)
```

Out[58]:



In [59]:
```
X test
```

Out[59]:
```
600     eight years after its release , disney has dec...
931     it's been a long time since walt disney has de...
937     richard gere can be a commanding actor , but h...
1811    1 . he doesn't have a hard-to-decipher accent ...
1512    when i arrived in paris in june , 1992 , i was...
                            ...
615     _in brief : _ this film needs no introduction ...
1029    there are two things the american film industr...
1342    for more than a decade , anjelica huston has b...
1030    note : some may consider portions of the follo...
770     here's a word analogy : amistad is to the lost...
Name: review, Length: 582, dtype: object
```

In [60]:
```
# Making preictions
predictions = text_clf.predict(X_test)
```

In [61]:
```
# Evaluation Reports

from sklearn.metrics import confusion_matrix, classification_report, accuracy_score
```

In [64]:
```
print(confusion_matrix(y_test,predictions))
```

```
[[235  47]
 [ 41 259]]
```

In [65]:
```
print(classification_report(y_test,predictions))
```

```
              precision    recall  f1-score   support

         neg       0.85      0.83      0.84       282
         pos       0.85      0.86      0.85       300

    accuracy                           0.85       582
   macro avg       0.85      0.85      0.85       582
weighted avg       0.85      0.85      0.85       582
```

In [66]:
```
print(accuracy_score(y_test,predictions))
```

```
0.8487972508591065
```