LAB07 - RF IN NLP DATA
Dr. Tran Anh Tuan (University of Science, Maths & Computer Faculty)

In [ ]:
```
# Movie Review Dataset
# The Movie Review Data is a collection of movie reviews retrieved from the imdb.com website
# in the early 2000s by Bo Pang and Lillian Lee. The reviews were collected and made available
# as part of their research on natural language processing.The reviews were originally released
# in 2002, but an updated and cleaned up version was released in 2004, referred to as “v2.0“.
# The dataset is comprised of 1,000 positive and 1,000 negative movie reviews drawn from an
# archive of the rec.arts.movies.reviews newsgroup hosted at IMDB.
# The authors refer to this dataset as the “polarity dataset“.
```

In [1]:
```
import numpy as np
import re
import nltk
from sklearn.datasets import load_files
nltk.download('stopwords')
import pickle
from nltk.corpus import stopwords
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Admin\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [2]:
```
path = r"D:\\Teaching and Training\\Nhan Dang Mau\\Data\\NLP\\txt_sentoken"
movie_data = load_files(path)
X, y = movie_data.data, movie_data.target
```

In [6]:
```
print(len(X))
```

```
2000
```

In [7]:
```
display(X[0])
display(y[0])
```

```
b"arnold schwarzenegger has been an icon for action enthusiasts , since the late 80's , but lately his films h
ave been very sloppy and the one-liners are getting worse . \nit's hard seeing arnold as mr . freeze in batman
and robin , especially when he says tons of ice jokes , but hey he got 15 million , what's it matter to him ?
\nonce again arnold has signed to do another expensive blockbuster , that can't compare with the likes of the
terminator series , true lies and even eraser . \nin this so called dark thriller , the devil ( gabriel byrne
) has come upon earth , to impregnate a woman ( robin tunney ) which happens every 1000 years , and basically
destroy the world , but apparently god has chosen one man , and that one man is jericho cane ( arnold himself
) . \nwith the help of a trusty sidekick ( kevin pollack ) , they will stop at nothing to let the devil take o
ver the world ! \nparts of this are actually so absurd , that they would fit right in with dogma . \nyes , the
film is that weak , but it's better than the other blockbuster right now ( sleepy hollow ) , but it makes the
world is not enough look like a 4 star film . \nanyway , this definitely doesn't seem like an arnold movie .
\nit just wasn't the type of film you can see him doing . \nsure he gave us a few chuckles with his well known
one-liners , but he seemed confused as to where his character and the film was going . \nit's understandable ,
especially when the ending had to be changed according to some sources . \naside form that , he still walked t
hrough it , much like he has in the past few films . \ni'm sorry to say this arnold but maybe these are the en
d of your action days . \nspeaking of action , where was it in this film ? \nthere was hardly any explosions o
r fights . \nthe devil made a few places explode , but arnold wasn't kicking some devil butt . \nthe ending wa
s changed to make it more spiritual , which undoubtedly ruined the film . \ni was at least hoping for a cool e
nding if nothing else occurred , but once again i was let down . \ni also don't know why the film took so long
and cost so much . \nthere was really no super affects at all , unless you consider an invisible devil , who w
as in it for 5 minutes tops , worth the overpriced budget . \nthe budget should have gone into a better script
, where at least audiences could be somewhat entertained instead of facing boredom . \nit's pitiful to see how
scripts like these get bought and made into a movie . \ndo they even read these things anymore ? \nit sure doe
sn't seem like it . \nthankfully gabriel's performance gave some light to this poor film . \nwhen he walks dow
n the street searching for robin tunney , you can't help but feel that he looked like a devil . \nthe guy is c
reepy looking anyway ! \nwhen it's all over , you're just glad it's the end of the movie . \ndon't bother to s
ee this , if you're expecting a solid action flick , because it's neither solid nor does it have action . \ni
t's just another movie that we are suckered in to seeing , due to a strategic marketing campaign . \nsave your
money and see the world is not enough for an entertaining experience . \n"
```

```
0
```

In [8]:
```
documents = []

from nltk.stem import WordNetLemmatizer

stemmer = WordNetLemmatizer()

for sen in range(0, len(X)):
    # Remove all the special characters
    document = re.sub(r'\W', ' ', str(X[sen]))

    # remove all single characters
    document = re.sub(r'\s+[a-zA-Z]\s+', ' ', document)

    # Remove single characters from the start
    document = re.sub(r'\^[a-zA-Z]\s+', ' ', document)

    # Substituting multiple spaces with single space
    document = re.sub(r'\s+', ' ', document, flags=re.I)

    # Removing prefixed 'b'
    document = re.sub(r'^b\s+', '', document)

    # Converting to Lowercase
    document = document.lower()

    # Lemmatization
    document = document.split()

    document = [stemmer.lemmatize(word) for word in document]
    document = ' '.join(document)

    documents.append(document)
```

In [13]:
```
print(len(documents))
```

```
2000
```

In [14]:
```
display(documents[0])
```

```
'arnold schwarzenegger ha been an icon for action enthusiast since the late 80 but lately his film have been v
ery sloppy and the one liner are getting worse nit hard seeing arnold a mr freeze in batman and robin especial
ly when he say ton of ice joke but hey he got 15 million what it matter to him nonce again arnold ha signed to
do another expensive blockbuster that can compare with the like of the terminator series true lie and even era
ser nin this so called dark thriller the devil gabriel byrne ha come upon earth to impregnate woman robin tunn
ey which happens every 1000 year and basically destroy the world but apparently god ha chosen one man and that
one man is jericho cane arnold himself nwith the help of trusty sidekick kevin pollack they will stop at nothi
ng to let the devil take over the world nparts of this are actually so absurd that they would fit right in wit
h dogma nyes the film is that weak but it better than the other blockbuster right now sleepy hollow but it mak
e the world is not enough look like 4 star film nanyway this definitely doesn seem like an arnold movie nit ju
st wasn the type of film you can see him doing nsure he gave u few chuckle with his well known one liner but h
e seemed confused a to where his character and the film wa going nit understandable especially when the ending
had to be changed according to some source naside form that he still walked through it much like he ha in the
past few film ni sorry to say this arnold but maybe these are the end of your action day nspeaking of action w
here wa it in this film nthere wa hardly any explosion or fight nthe devil made few place explode but arnold w
asn kicking some devil butt nthe ending wa changed to make it more spiritual which undoubtedly ruined the film
ni wa at least hoping for cool ending if nothing else occurred but once again wa let down ni also don know why
the film took so long and cost so much nthere wa really no super affect at all unless you consider an invisibl
e devil who wa in it for 5 minute top worth the overpriced budget nthe budget should have gone into better scr
ipt where at least audience could be somewhat entertained instead of facing boredom nit pitiful to see how scr
ipt like these get bought and made into movie ndo they even read these thing anymore nit sure doesn seem like
it nthankfully gabriel performance gave some light to this poor film nwhen he walk down the street searching f
or robin tunney you can help but feel that he looked like devil nthe guy is creepy looking anyway nwhen it all
over you re just glad it the end of the movie ndon bother to see this if you re expecting solid action flick b
ecause it neither solid nor doe it have action nit just another movie that we are suckered in to seeing due to
strategic marketing campaign nsave your money and see the world is not enough for an entertaining experience'
```

In [15]:
```
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer(max_features=1500, min_df=5, max_df=0.7,
                             stop_words=stopwords.words('english'))
X = vectorizer.fit_transform(documents).toarray()
```

In [17]:
```
# The term frequency is calculated as:

# Term frequency = (Number of Occurrences of a word)/(Total words in the document)
# And the Inverse Document Frequency is calculated as:

# IDF(word) = Log((Total number of documents)/(Number of documents containing the word))
# The TFIDF value for a word in a particular document is higher if the frequency of occurrence of
# that word is higher in that specific document but lower in all the other documents.
```

In [16]:
```
from sklearn.feature_extraction.text import TfidfTransformer
tfidfconverter = TfidfTransformer()
X = tfidfconverter.fit_transform(X).toarray()
```

In [17]:
```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

In [18]:
```
from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(n_estimators=1000, random_state=0)
classifier.fit(X_train, y_train)
```

Out[18]:
```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=1000, n_jobs=None,
            oob_score=False, random_state=0, verbose=0, warm_start=False)
```

In [19]:
```
y_pred = classifier.predict(X_test)
```

In [20]:
```
y_pred
```

Out[20]:
```
array([0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1,
       1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0,
       1, 1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0,
       0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1,
       0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0,
       1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1,
       0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1,
       1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0,
       0, 0, 0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0,
       0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1,
       0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0,
       0, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0,
       0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1,
       0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 1,
       1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0,
       1, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1,
       1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1,
       0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1,
       0, 0, 1, 0])
```

In [21]:
```
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

print(confusion_matrix(y_test,y_pred))
print(classification_report(y_test,y_pred))
print(accuracy_score(y_test, y_pred))
```

```
[[180  28]
 [ 30 162]]
              precision    recall  f1-score   support

           0       0.86      0.87      0.86       208
           1       0.85      0.84      0.85       192

   micro avg       0.85      0.85      0.85       400
   macro avg       0.85      0.85      0.85       400
weighted avg       0.85      0.85      0.85       400

0.855
```

In [22]:
```
path = r"D:\\Teaching and Training\\Nhan Dang Mau\\Data\\NLP\\"
# Saving and Loading the Model
with open(path + 'text_classifier', 'wb') as picklefile:
    pickle.dump(classifier,picklefile)
```

In [23]:
```
with open(path + 'text_classifier', 'rb') as training_model:
    model = pickle.load(training_model)
```

In [24]:
```
y_pred2 = model.predict(X_test)

print(confusion_matrix(y_test, y_pred2))
print(classification_report(y_test, y_pred2))
print(accuracy_score(y_test, y_pred2))
```

```
[[180  28]
 [ 30 162]]
              precision    recall  f1-score   support

           0       0.86      0.87      0.86       208
           1       0.85      0.84      0.85       192

   micro avg       0.85      0.85      0.85       400
   macro avg       0.85      0.85      0.85       400
weighted avg       0.85      0.85      0.85       400

0.855
```

In [36]:
```
# import pandas as pd
# import numpy as np

# # load the dataset of SMS messages
# path = r"D:\\Teaching and Training\\Nhan Dang Mau\\Data\\NLP\\"
# df = pd.read_csv(path + 'SMSSPamCollection', header=None, encoding='utf-8', sep='\t')
# df.columns = ['Class', 'Raw Message']

# print(df["Class"].unique())
# display(df.head())
```

```
['ham' 'spam']
```

	Class	Raw Message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...