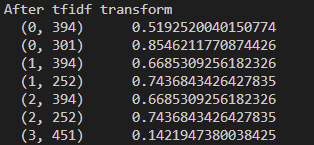
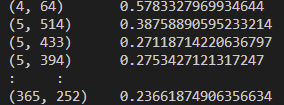
Natural Language Processing Project  
Comp237

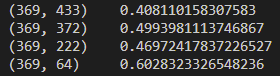
Wing Yan Lau 301229696

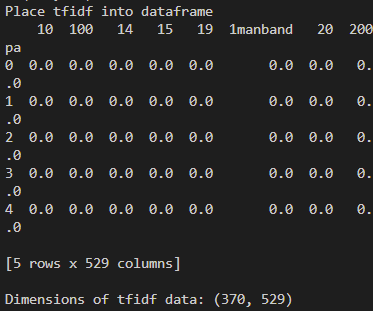
Sze Man Tang 301221595

Liangyu Wang 980025288

1. We Loaded “Youtube05-Shakira.csv” file into a Pandas data frame.
2. For this project we use only the “CONTENT” and “CLASS” columns.
3. For pre-processing, we strip all the punctuation marks from the comments in CONTENT column.
4. We then use count\_vectorizer.fit\_transform() to tokenize and vectorize all the comments using the bag of words model. We use min\_df=2 to ignore words that are in less than 20% of comments, and use max\_df=0.75 to ignore words that are in more than 75% of comments. We use CountVectorizer’s default list of stop words. The dimension of our vectorized comments data is (370, 529), meaning our bag of word model contains 529 features.
5. The data is downscaled using TfidfTransformer and placed into dataframe for further processing in the next step.

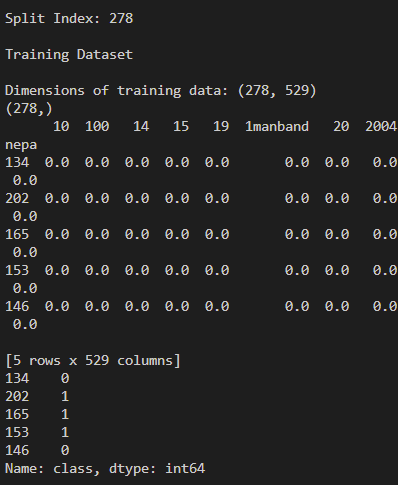
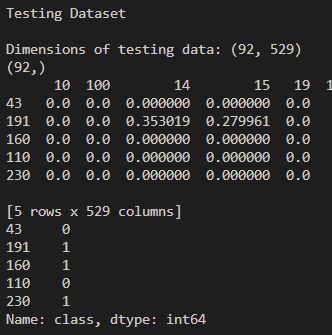




1. The data is shuffled using pandas.DataFrame.sample.
2. The shuffled data set is split into training set (75% of data) and testing set (25% of data).  
   Total # of row of data = 370

# of training data = 370\*0.75 = 278

# of testing data = 370-278 = 92

1. Naïve Bayes classifier is created and is fit with 75% of training data
2. Cross validate the model on the training data using 5-fold and print the mean results of model accuracy.  
   
3. Test the model on the test data, print the confusion matrix and the accuracy of the model.  
   Text

   Description automatically generated
4. As a group come up with 6 new comments (4 comments should be non spam and 2 comment spam) and pass them to the classifier and check the results.

* 4 non spam comments:
  + I love you Shakira!!!!!!!
  + Good song <3
  + I am still listening to it in 2023
  + Love it
* 2 spam comments:
  + Please subscribe YouTube Chanel of Centennial College
  + My home has gone and I don't have a job now. Pease click the fund raise link below and support me and my family!!!! Thanks!!!!!!

Text

Description automatically generated

Conclusion

The low number of false positive (6) and low number of false negative (3) in confusion matrix and high accuracy of the model which is close to 1 (0.9021739130434783) showed that the model is good and highly accurate.

However, there are still some limitations. Comments contain words like “link”, “subscribe” and “click”, etc., will be classified as spam. Firstly, if the comment only contains symbol, it will be trimmed and still consider as non spam. Secondly, if the video content is related to words classified as spam like “link”, “subscribe” and “click”, etc., the accuracy of this model will be lowered and new set of data should be used to train the model again in order to update the accuracy. Thirdly, accuracy of short comments is lower than that of long comments.