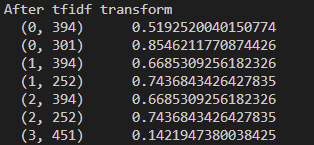
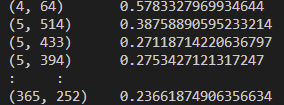
Natural Language Processing Project  
Comp237

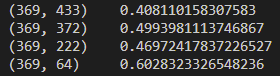
Wing Yan Lau 301229696

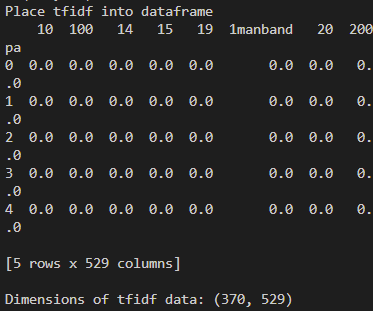
Sze Man Tang

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.

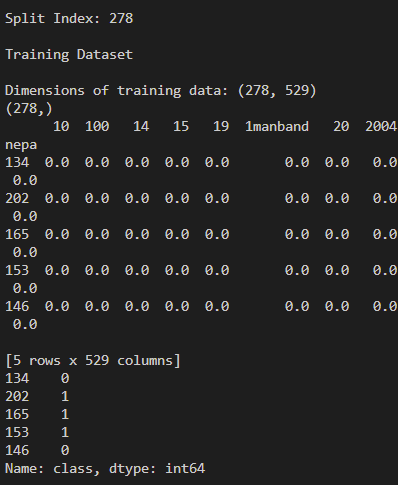
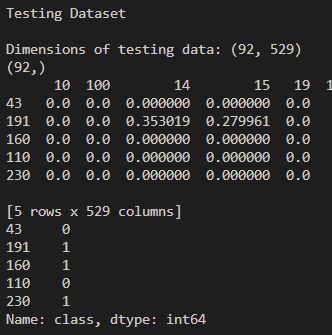




1. The data is shuffled using pandas.DataFrame.sample.
2. The shuffled data set is split into training set (75% of data) and testing set (25% of data).  
   Total # of row of data = 370

# 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