Natural Language Processing Project  
Comp237

Wing Yan Lau

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 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 comments data’s vocabulary contains 529 words.