JSC270 WRITTEN REPORT

Mohsin Reza, You, Peng, Wei Yu

Work Breakdown

- Mohsin – Completed modelling portion of question 2, written report
- You – Completed all the code for question 1, word clouds in question 2
- Wei – Preprocessing Steps in question 2

Section 1

A. Outputted in the code in google collab
B. In google collab
C. In google collab
D. One scenario where we might want to keep punctuation is if we want to distinguish between words that a user has used in the body of a tweet versus words used in hashtags. In this situation, we would want to keep the hashtags rather than remove them to be able to distinguish the two groups of words. In another scenario, we might want to keep things like all capitals and exclamation marks, as these are part of the way in which the user is communicating.
E. In google collab
F. In google collab
G. The length of our vocabulary is 1000, which we got from the code in section G of the google collab notebook
H. After running the Naïve Bayes Model on our data, we report a training accuracy of 68% and a testing accuracy of 42%.
I. No, we decided that it would not be appropriate to plot an ROC curve in this situation since there are multiple classes involved, and hence a false positive or false negative rate wouldn't really make sense.
J. This time, the training accuracy was 66.6% and the resting accuracy was once again 43%, so using TF-IDF vectors instead of count vectors didn't make a significant difference in our results.
K. The accuracy with lemmatization was almost the same as the accuracy with stemming, at 66.6% for training and 43% for testing, respectively.
L. Naïve Bayes is a discriminative model since it focuses on predicting labels of the data rather than generating new data

Section 2

I. The main question that our group was interested in answering was if you could use the content of a tweet i.e. what words and phrases are used to predict the number of followers a tweeter has. We found that this question was difficult to solve as there isn't an obvious answer as to what types of words a person with a high number of followers would use compared to a person with a low number of followers. We did not find any literature related to the question that we wanted to solve.
II. We extracted 2000 tweets from the API with 2 features – the content of the tweet and the number of followers the user had. The only thing we specified in the search parameters is that we wanted

to exclude retweets. This is because including retweets would mean that we had tweets from a user with vocabulary that they didn’t use, which would defeat the purpose of our analysis. Once again, we didn’t find other words of research like ours.

III. Some of the properties of our data is that we had a mean of 580 followers, and in addition, we decided to split the number of followers into 3 groups – low (0 to 1000), medium (1001 to 10000), and high (> 10000), and we created a new categorical variable to represent this. Our data had 1567 medium number of followers, 81 people with a high number, and 452 with a low number. We then performed some preprocessing steps, which can be viewed in the google collab.

IV. We decided to use the Naïve Bayes Model to answer our research question. This is a supervised model, and it works by splitting the classifying each label based on what the probability is that it is the true label. Some of the strengths of this model are that it is fast and easy to use, however, a weakness, which we also fell into, is that it can perform poorly if the training dataset is not evenly split up into the different labels. We decided to evaluate our model based on its accuracy.

V. When we ran our model, we had a testing accuracy of 75.5%, however, unfortunately our model always predicted a medium number of followers as the label as this was the majority label in the dataset. This means that our results are not significant and we can’t really draw any meaningful conclusions, since our accuracy only represents the proportion of the majority label, and not the accuracy of the model. If we had more time, we would spend significantly more time in refining our data before we decide to put it through a model. We would do this by firstly attempting to collect a larger number of observations, as well as to split the data into the training set in a way that is representative of the overall dataset.