**Social Media Analytics: Final Project**

**Xinyan Cai (31545891)**

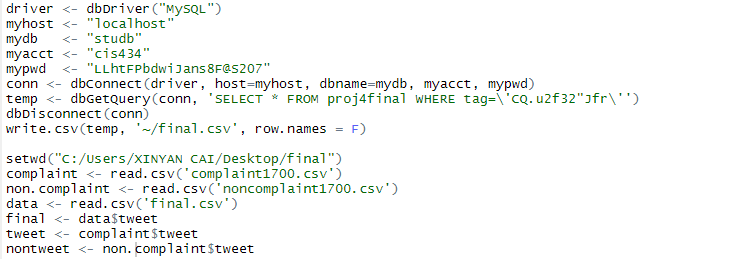
­­­­­­­­­

**Introduction**

In this report, the types of Twitter posts incorporating noncomplaint and complaint were classified using the featured negative words. The judgment of Twitter posts types could help the airplane companies to identify customers’ attitudes towards their services, therefore leading to service improvements. The whole investigation aims to find noncomplaint twitter posts.

**Data**

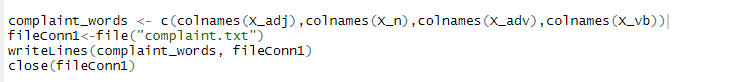
There are two big datasets include the noncomplaint twitter posts and complaint twitter posts which are utilized to build the model. After that, another data with mixed types could be classified by the model built before.

**Method**Firstly, I obtained data from the SQL server and store it as a local file called ‘final.csv’. Also, I read different files to R. 

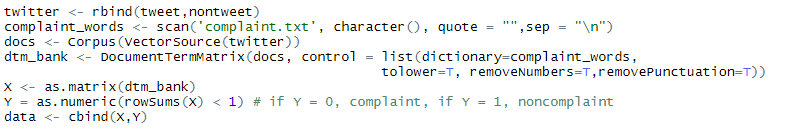
Secondly, I extracted complaint words from only the complaint dataset because the complaint words were more radical so that is could be easier to classify the complaint posts and noncomplaint posts. Here I specifically extracted adjectives, nouns, advs, and verbs from the complaint dataset and the most useful words belonged to adjectives. It is reasonable that people often use adjectives to express their emotions.



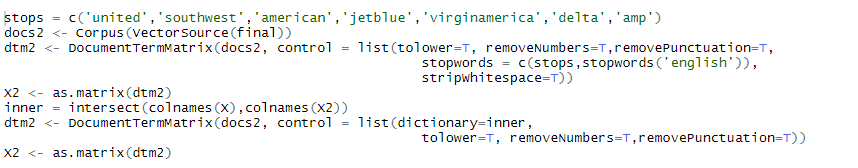
Thirdly, I found that some words extracted from complaint posts are misleading or meaningless such as ‘good’ and ‘able’. Therefore, there was an important step that I wrote the words into a local file for the further edition by hand.



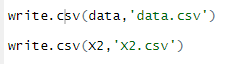
Furthermore, according to the combination of the complaint and non-complaint datasets, I derived X and Y using the complaint words I modified. It is also worth mentioning that I decided on the classification (Y) in terms of ‘rowSums’.



In order to find X2 of the dataset with mixed posts types for further prediction, I also created ‘dtm’ of this dataset.



Finally, in order to build random forest more successfully and easily, I used python to do final model building and prediction.





**Validation**

In total, I obtained 143 non-complaint twitter posts and the accuracy is around 0.7 which is not bad. Therefore, it indicates that my model is precise in general.

**Further Improvement**

Although the model can accurately classify the types of twitter posts. There are still a large number of posts that are hard to be separated. Thus, other models such as SVM, bagging, and boosting can also be tried. Meanwhile, it is also significant to extract more meaningful words among the complaint file, which is beneficial for prediction.