A. Is this a multi-label classification problem? Briefly explain the implications.

Answer: This is a multi-label classification problem. Because there are two kinds of classification: binary classification and multi classification. In this case, apparently there are multiple labels, so this is a multi-label classification.

B. Show and explain any data processing/feature engineering performed.

Answer: I firstly import training data, split it into train data frame and validate data frame. Then use CountVectorizer function in sklearn to tokenize the text, then I use TfidfTransformer function to transform the word occurrences to frequencies. So I can extract features from text files. Then I use naïve bayes from sklearn to train the data, then predict the results of validate and test data. Finally restore the predicting result into the Siyu\_Xiao\_submission\_problem\_1.csv file.

C. What ML model did you use and why?

Answer: I use Naïve Bayes model. This is just a very easy question. I can use many models to do this classification. So I randomly pick this model.

D. Explain any measures you took against overfitting.

Answer: There are many methods to avoid overfitting. I use cross validating to avoid overfitting. We randomly use 2000 pieces of data as validating data, repeat this process for 10 times and use the average as validation.

E. How did you assess the efficacy of your model?

Answer: There are many methods to assess. For example, we can use the accuracy of our predicting results, we can use code running time and so on.

F. Run the model against the data in *testing\_data.csv* and provide the results in *yourname*\_submission\_problem\_1.csv

Answer: I put all of the files in this file folder and compress it into a package I sent to you. You can decompress it and directly run the code without the worry of directory.

G. What further enhancements to the model would you consider if you were given more time?

Answer: I want to use models with higher fitting ability to do this. Especially neural network. As we all know neural network can fit both linear and non-linear data, especially for the data whose features are hard to describe like sequence string data.