# HW1 – Chandler Fang

# Model performance

Base case -> remove special characters and stopwords -> apply simple stemming and lemmatization

Bag of words: 64% -> 65.6% -> 65.6%

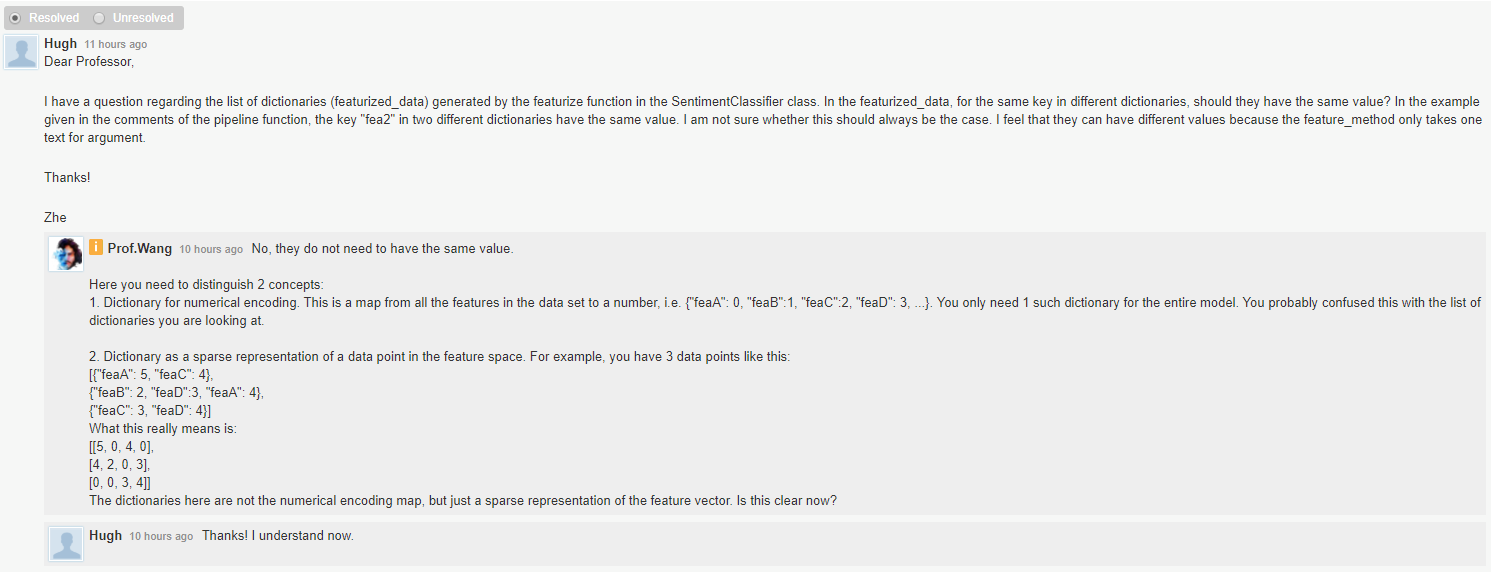
2-Gram method: 64% -> 67.3% -> 64.5%

TF-IDF method： 60% -> 63% -> 62.1%

# Challenge and Lesson I learnt

Firstly, I have zero experience in NLP. So this time I learn how to follow the process of tokenization-> Normalization -> Encoding and apply a simple NLP algo. The biggest challenge is how to start and where to start. Some lessons I have learnt are:

Lesson1: Featurizer and pipeline are different and for different purpose. As professor mentioned in Piazza answering the question. I made the same mistake before.



Lesson2: Preprocess the text to improve the prediction accuracy. Although not required, I remove some stop words and it does improve the prediction accuracy. However lemmatization doesn’t help much but maybe I didn’t implement it properly. I think I may learn more in the following sessions.

# Optional Question

Sorry I don’t get extra time to think about optional question. But if I get some time, I will first try Poisson Regression, maybe it is more suitable for count base number. I may come back and check next time.

# Note

I didn’t use the full train.txt data for training and prediction, given extremely long running time. I truncated around 30% of the data and put it into train\_sample.txt file.