Nicholas Ketz
Assignment 3 – Feature Engineering
Kaggle user name = NickCU
(sorry, I created the kaggle account without carefully reading assign.md so my username doesn't match my CU id)

I started by analyzing errors by looking at the confusion matrix across all labels. It seemed the most fruitful place to increase performance was targeting the 'Social Science' class which was often being mislabeled as 'Social Studies', as shown in Figure 1.

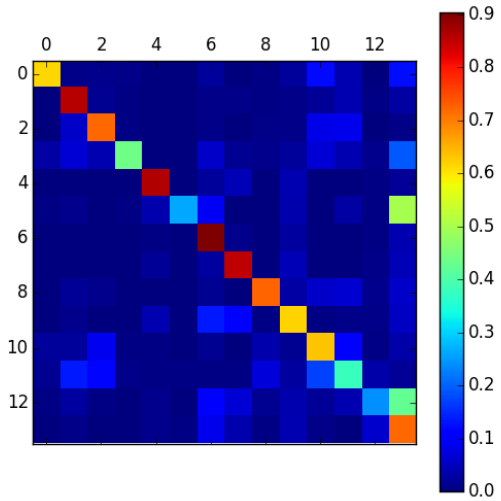

Figure 1: confusion matrix for categories normized by the number of samples in each category.

To try and increase performance I created several features as custom sklearn feature classes, that way they would be easily included through sklearn's FeatureUnion. They were z-scored counts of sentences (SentCount), capital letters (CapsCount), alpha characters (AlphaCount), and numeric characters (NonAlphaCount). I also changed the count vectorizer defaults to include bigrams and exclude standard stop words.

Using 10-fold cross-validation to evaluate test performance, I tried several different combinations of these features, including tri-grams, and non z-scored features. I found the submitted version to have the best performance.