IST736 Text Mining

HW7

**Compare MNB and SVMs**

**for (Kaggle Sentiment Classification – Kaggle is optional and you may find, gather with an API, or create any good sentiment type data set)**

**Comments and Notes in BLUE ;)**

1. **Be sure to use the Assignment format. Remember that the dataset defines the topic/area.**

So far we have learned how to use sklearn to build MNB and SVMs models and evaluate them using various test methods and measures.

Now consult the sklearn documentation

and revise the instructor's given script (**using the instructor script is optional and you can also use my code. Ideally – you should practice writing your own code for SVM and NB).**

**Your Output/Results should include visualizations, tables, confusion matrices, etc.**

**Include precision and recall.**

to output confusion matrix, precision and recall values for the Kaggle Sentiment training data. Remember the sample script used 60% data for training and 40% for testing.

**While 60/40 is fine for testing and training, you are free to use 30/70 and/or to try both and compare.**

**As part of this assignment, try to use both CountVectorizer AND TfidfVectorizer for the data and the apply both to your Naïve Bayes (MNB) and your SVM.**

**Required: For the SVM – use three kernels and also try a few different costs for each kernel. Create a table to compare.**

**Also be sure to compare the accuracies of MNB and SVM.**

* **Try to find a good method to generate the top 10 indicative words for the most positive category and the most negative category from the MNB and SVMs models respectively. You can find code for this that I posted ☺ and in the asynch. You can also Google it.**

~~Task 2:~~

~~Consult the sklearn website to learn more about the CountVectorizer. Revise the script to build a MNB model and a SVMs model based on both unigram and bigram. For fair comparison, please keep the same 60% for training and the rest 40% for testing. Also keep your other vectorization parameters the same as in Task 1.~~

~~- Compare the confusion matrix and other evaluation measures (accuracy, precision, recall). Discuss whether adding bi-grams was helpful for sentiment classification, based on MNB and SVMs respectively.~~

~~Task 3:~~

~~Now revise the sample script to build your best SVMs model by tuning parameters and using the entire training data set (changing from 60% to 100%). Report what parameters you used to train the model, and its cross validation accuracy.~~

**Using Kaggle data and submitting to Kaggle is optional – not part of the assignment format or paper – but can be reported at the very end. Its fun and so I recommend it IF you have time.**

Then use this model to predict the Kaggle sentiment test data. Submit the prediction result to Kaggle, use screenshot to show your accuracy and ranking.

https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/submit