Homework 4 – Supervised machine learning

1. As a first step work through the code for the example in this introduction to RTextTools

https://journal.r-project.org/archive/2013-1/collingwood-jurka-boydstun-etal.pdf

You will be classifying congressional bill titles for major topic (20 topics total) using part of the data to train a model, and another part to test its performance. Begin by ONLY using the SVM model, which is the fastest (and also one of the best).

[Note: you may encounter an error along the way. Read the error message and see if you can fix the data to make it go away!]

2. What proportion of the cases are being used to train the model, and what proportion to test?
3. Compute overall precision, recall and the F-score for your results.
4. Next, cross-validate your results using a 10 fold cross validation. Are the results the same as your initial precision and recall results?

Part II

5. Now start over by training and testing the performance of three algorithms (this is going to take longer)
6. Report ensemble agreement for these three algorithms. Explain the difference between coverage and accuracy.

Part III

7. In the results generated for the SVM algorithm, you will find the human-assigned labels and the predicted labels. Use them to create a confusion matrix (you'll need to explore options for doing this)
8. Where does the algorithm perform best and worst across the topics? Are the errors for the worst topic randomly distributed or more systematic? Are they more likely to be false positives or false negatives? What might you do to improve algorithm performance?