Assignment 4 Simplification (due Nov 22)

"Teachable moment": It has been reported to us that for the original assignment, the sensitivity of the SVM classifier is not only low – but is actually 0! How is that possible? Here is what happens.

- As there is a limit on our computational power, we restrict the number of terms to be approximately 100, instead of, say, 1000 terms.
- We learned in class when we discussed the SVM kernel trick that a non-linearly seperable problem becomes linearly separable at a higher dimensional space. But because we restrict ourselves to 100 terms, we have gone the other direction – that is, we have made our problem more non-linearly separable.
- We also learned in class that for non-linearly separable problems, SVM optimizes by minimizing total error of the training samples.
- Here comes the other part of the "killer" combination. Because the negative class is 4-5 times bigger than the positive class, and is well over 10,000 documents, the optimization of this SVM module is reduced to minimizing the total error of the negative documents. In other words, SVM has completely given up on the positive class – hence, sensitivity is 0.
- In summary, a "killer" combination: (i) low dimensionality; (ii) non-linearly separable; (iii) imbalance data set; and (iv)a high number of negative training samples, causes the "collapse" of the SVM module.

Simplification: Instead of forming the negative class using the remaining 15 non-*comp* folders, just form the negative class using the *talk** folders (i.e., the 3 talk.politics folders and the talk.religion folder). In other words, the binary classification problem is now *comp** vs *talk**. With this change, we simultaneously remove the imbalance and significantly reduce the number of negative training samples. A corresponding change is made to the test set as well.

- One more tip: sparse terms should be removed simultaneously across both the training and the test sets. If you have a computer that has more computational power, you can try to see if you can use more than 100 terms now with the smaller data set. This would raise the dimensionality of the problem and make the problem less non-linearly separable.