**B00060572 David Kelly**

**Lab Sheet#3**

**Exercise 1**: Both Risen and Rift have a count of 1 in K4.txt. Risen does not appear in any another documents, Rift appears in two other documents. Based on you knowledge of IDF, explain which term will get the higher weighting, and why?

a) The IDF value is calculated by how many documents the word appears in. Although Rift and Risen both appeared once in K4.txt, Rift also occurred in 2 other documents. The uniqueness of the word “risen” occurrence document is greater.

**Exercise 2**: What has happened to the count of '1' for both risen and rift? Explain how this figure is derived.

a) A weighted value is calculated using the number of documents and the number of documents the word occur(s) in.log (N/n) Then after a idf is found it is normalized to a value relative to the number of terms in the bag of words. This is done with the calculation for

TF = .5 + ((idf\*.5)/idf value for most frequent term)

**Exercise 3**: How accurate is a classification model on 15 terms compared to model accuracy before pruning the terms?

a) The accuracy for me does not change depending on pruning in my case. 90% before pruning and 90% after pruning with 10 terms remaining.

**Exercise 4**: Do you need to keep the filters which are embedded in the process document from files operator? Explain your answer.

a) No you don’t need to keep the filters as the Bag of words must be the same in the test data and in the model data. This means that ultimately the bag of words will be the filter.

**Exercise 5**: How well did your classifier do? Can you explain the misclassifications?

a) With k-NN my model had an accuracy of 90% but when applied to test data the model correctly assigned 75% of the documents. After examining the misclassified document I could see that not terms existed in my bag of words and therefore the document was assigned to the closest class which was Kenya. With Decision tree it would have been decided with the less than branch.

**Exercise 6**: Can you improve accuracy by using a different learner? Try one other learner, and report on the results. You can do this as follows:

a) I replaced the k-NN operator with a decision tree in the process. I received a lower accuracy than with the k-NN. Correct classification occurred 50% with decision tree 75% with k-NN

**Exercise 7**: Where there are differences between the two processes, explain if that will affect the results of testing the model.

If the Dictionarystemmer was used in the test process rather than the training process, would that be OK?

a) prune below must be integer value with minimum 0 in the test data it is set to -1.

The stemmer must be the same because the stem of the word is not the same with both lovins and porter. Model uses PortersStemmer so the test should use PortersStemmer but its uses lovins