Text Mining

LAB 2 – B00060572

Exercise 1: Give your list of stopwords for the 15 texts. Also state how many attributes remain once you apply your stopword list.

a) Further, minister, rights, court, latest, legislation, year

Exercise 2: How many attributes are left when words of length 3 or less are removed? Do you think this filter is useful?

a) 202

Exercise 3: Include some of the phrases generated in your lab report.

a) Annan\_brokered, Archbishop\_Dublin, Mary\_Harney

Exercise 4: Give your list of stemwords/word conversion for the 15 texts. Also state how many attributes remain once you apply these conversions.

a) hospital:hospitals.\*=3, health:healthcare.\*=3, crime:criminal.\*=4, contract:contracts.\*=3, kofi:annan.\*=9, kofi:chief.\*=9

Exercise 5: Add a LovinsStemmer – evaluate the words produced. Do you think it its an improvement?

a) The Lovins Stemmer removes common endings from words, Lovins doesn’t seem concerned with the words legibility only that it is efficient at stemming a word where it appears to become common. Even changing a word from believes to belief.

Exercise 6: Replace Lovins Stemmers with Porters Stemmer. Compare the results.

a) The Porters stemmer removes common word endings but leaves more of the word than Lovin Stemmers. The work is truncated but I can still distinguish the original word more clearly.

Exercise 7: As per last week, generate a decision tree to predict each class, and then apply and evaluate the model. How accurate is the classifier?

a) 25%

Exercise 8: Does the inclusion / removal of stemmers effect the accuracy?

a) Removing Lovins Stemmer and Porters Stemmer did not affect my decision tree, however the removal of Stem Dictionary changed classifier form “Crime” to “Annan” and reduced the accuracy of the tree. Using stemmers with Naïve Bayes classifier decreased the accuracy 10% with both Porters and Lovins.

Exercise 9: Can you improve the accuracy using another classification

a) Replacing the binary tree classifier with a Naïve Bayes operator I was able to achieve 100% accuracy with the 3 classes. Naïve Bayes seems to work best when there is a lot of irrelevant data within a small data set.