Ling570 Hw10

Michael Roylance (roylance@uw.edu), Olga Whelan (olgaw@uw.edu)

Q3

F1 – unigram feature

F2 – prevalence feature. This feature is available only for comparison of two directories. It assumes that particular words are more prevalent in one corpus than in the other. This feature looks at words that are present in both analyzed groups, but have a considerably different frequency count in each of them. For this feature we have to choose a threshold that defines a critical difference in frequencies. In the current version we chose a threshold of 25%, i.e. we pay attention to words that are at least three times more frequent in one set than in the other. For each such word, using its exact frequencies in each set, we determine exactly how much more frequent it is in one than in the other. Based on that ratio, we calculate prevalence1 and prevalence2, that sum up to 1.

F3 – unique unigrams feature. This feature looks at words that are only found in one of the sets. Here we have to choose a frequency threshold for the words that we want to consider. In the current version we keep words that are found in one of the sets at least three times. For each word the feature returns information about whether the feature is unique, and if yes, what set it belongs to.

F4 – unique bigrams feature. Same as F3 for bigrams.

F5 - we skipped this feature, could not get it working

F6 - average length of all words in the file

F7 - sentiment analysis of each word, we grabbed the word list from http://alexdavies.net/ - a Ph.d student at Cambridge.

Q6

Table 2

expt id	feature combination	Training accuracy	Test accuracy
expt1	F1	0.9996894409937889	0.9354838709677419
expt2	F2		

0.5055900621118012

0.5049627791563276

expt3

F3

0.9666149068322981

0.004962779156327543

expt4

F4

0.9996894409937889

0.0024813895781637717

expt5

F1+F2

0.5055900621118012

0.5049627791563276

expt6

F7

0.7611801242236025

0.7754342431761787

expt7

F6 + F7

0.7611801242236025

0.7754342431761787

expt8

F1 + F7

0.7892857142857143

0.7878411910669976

Conclusions

Unigram feature (F1) returns the highest accuracy that is difficult to match or improve.

F2 in the way it is implemented now, classifies all the texts as belonging to one of the sets, so the accuracy is defined by the percentage of the texts from that set. Not good.

F3 and F4 (unique 1 and 2-grams) do very well for training, but in testing they classify left as right and right as left; so the accuracies are close to zero.

F7 by itself did a decent job with about 76% and 78% accuracy, respectively. This can be attributed to the sentiment of the words being present.

F1 + F7 did slightly better than just F7 by itself, this is most likely do to the superior unigram accuracy.

These accuracies give us an insight that we incorrectly implemented our vectors, because the ideas behind the features seemed reasonable.