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 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. Here we have to choose a threshold for the difference in frequencies. In the current version we chose a threshold of 25%, i.e. we pick words that are at least three times more frequent in one set than in the other. For each such word we determine exactly how much more frequent it is and based on that calculate prevalence1 and prevalence2, that sum up to 1.

F3 – unique unigrams feature. This feature looks at words that can only be 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 twice.

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

**Table 2**

|  |  |  |  |
| --- | --- | --- | --- |
| expt id | feature combination | Training accuracy | Test accuracy |
| expt1 | F1 | 0.9996894409937889 | 0.9354838709677419 |
| expt2 | F2 | 0.9993788819875776 | 0.93424317617866 |
| expt3 | F3 | 0.9996894409937889 | 0.04094292803970223 |
| expt4 | F4 | 0.9996894409937889 | 0.0024813895781637717 |
| expt5 | F1+F2 | 0.9996894409937889 | 0.9330024813895782 |