Homework 5 for LING 572

Haotian He

February 10, 2014

Collaborators: Haotian He and Yaohua Zhuo

1 Q2

The command for filtering out unrelated features from the vector files is as below,

```
./chi_feature_selection.sh $1 $2 $3 $4 $5 $6
$1 = <original train vector file>
$2 = <original test vector file>
$3 = <chi square value>
The threshold of chi square, features below this threshold will
not be selected
$4 = <feat_list>
Each feature and its chi square value
$5 = <filtered train vector file>
Train vectors file after feature selection
$6 = <filtered test vector file>
Test vectors file after feature selection
```

Table 1: (Q2) Test accuracy talbe

p_0	χ_0^2 score	Number of related features	Test accuracy
baseline	0.0	32846	0.72
0.001	13.816	2401	0.746666666666667
0.01	9.210	3895	0.746666666666667
0.025	7.378	5223	0.753333333333333
0.05	5.991	7484	0.736666666666667
0.1	4.605	8499	0.75

2 Q3

Table 2: (Q3) Test accuracy using binary features

p_0	Test accuracy
baseline	0.82
0.001	0.856666666666667
0.01	0.866666666666667
0.025	0.86
0.05	0.826666666666667
0.1	0.86

3 Q5

(1) We can see a improvement trend from both the two tables, and both of the performances generally goes up as the p_0 value goes up.
(2) Moreover, the performance of binary features is even better than that of real-value features.
(3) One minor thing may make us note is that when p_0 equals 0.05 in both Q2 and Q3, the test accuracy does not exactly follow the general trend, and a little lower.