10-605 Assignment 1

Namit Katariya (andrew id: nkatariy)

January 27, 2013

QUESTION 1

- Output of training and testing on the "Very Small" dataset:

```
[C24,CCAT,M14,MCAT]                          GCAT     -8837.889580619143
[E51,E512,ECAT,GCAT,GDIP]                    CCAT     -3321.9899605306355
[C15,C152,C18,C181,CCAT]                     CCAT     -842.0977201801019
[GCAT]                                       CCAT     -1393.8777561662928
[C13,CCAT,GCAT,GHEA]                         CCAT     -604.5813059774287
[C13,CCAT,M11,MCAT]                          CCAT     -1261.9526299152496
[C11,C13,CCAT,E12,ECAT,M13,M132,MCAT]        CCAT     -1867.783474781556
[C31,CCAT]                                   CCAT     -1909.4938087854025
```

-

Dataset	Percent Correct
Very Small	62.5
Small	80.44
Full	84.49

QUESTION 2A

One could discard frequently occurring words like “the”, “a”, “for” etc. A better way would be to do what we discussed in class : instead of actually incrementing variables, we output statements corresponding to these increments and write it to a file. We can retrieve the counts by sorting this file and collecting the terms occurring.

QUESTION 2B

One could keep a threshold value between 0 and 1 and predict the document to have all the labels which have a probability above this threshold. The threshold could be decided differently for different documents. For example, if the probabilities turn out to be 0.3, 0.2, 0.19, 0.15 and say more than 10 other labels each with probability around 0.02, then 0.1 would be a good threshold value.
Another method could be to have a fixed k, sort the labels according to their probabilities and label the document as having the top k labels.