**Report Homework 2**

**Daniel Ribeiro Silva**

**drsilva**

**Question 1**

Total run time for abstract: 14min49sec

Total run time for links: 1h28min13sec

**Question 2**

I’ve created a bash script that runs the code ten times for each assignment on the local copy of the RCV1.small\_train.txt file from assignment 1.

The code used is the following:

For Assignment 2

*javac \*.java*

*time {*

*cat ../data/RCV1.small\_train.txt | java -Xmx128m NBTrain | sort -k1,1 | java -Xmx128m MergeCounts > outputA.txt*

*}*

For Assignment 1

*javac \*.java*

*time {*

*cat ../RCV1/RCV1.small\_train.txt | java NBTrain > outputB.txt*

*}*

The result was:

Average time for training in Assignment 1: **2.953 s**

Average time for training in Assignment 2: **5.566 s**

**Question 2**

One way of getting the most informative features is to analyze, for each feature, its conditional probability to a given label and compare that probability with relation to other labels. For example, if we want to know if the feature F is an informative feature for any label, we will take the conditional probability P(F|Y) for each label Y. Now we compare those conditional probabilities among all labels. For example, let be the label that has the highest and let be the average probability of having F, i.e. .

We can analyze the ratio . If r is very high, it probably means that F is an informative feature of the label . This approach can vary in many ways. We can, for instance, compute the ratio r with relation to the second highest value, or with the median value. The advantage of such an approach is that is only uses values that are pre-computed (the probabilities) and has linear complexity in the size of the vocabulary and on the number of labels.

Another possible approach (though much less efficient) is to remove one feature at a time and compute the results for the testing part. If the result quality decreases, then that feature was probably important for some label. We can compare the previous prediction with the new prediction to know what label(s) that actually is.

**Question 4**

Vinay Vyas and I talked about the memory efficiency topic in this homework and how the java BufferedReader handles the memory (he explained it to me).

Vinay Vyas, Yu-Hsin Huo and I talked about question 3, discussing possible approaches to that question.

Jefferey Gee and I talked about the techniques of testing for large datasets and about how estimating the value of the training vocabulary.