BigML Assignment 2: Small Memory Footprint Streaming Naive Bayes

Hao Gao (haog)

Did you receive any help whatsoever from anyone in solving this assignment? No
Did you give any help whatsoever to anyone in solving this assignment? No

Q1

For Accuracy, both of the datasets are almost the same. For the runtime, the links datasets is slower than the abstract dataset.
The tokens in the links dataset are much longer, they will consume more memory to build the hashtable.

Q2

Assignment1	8.05s
Assignment2	80.26s

Use a shell script to measure the time

```
#!/bin/bash
start_time=`date +%s`
for i in 1 2 3 4 5 6 7 8 9 10
do
   cat RCV1.small_train.txt | java -Xmx128m NBTrain |
                 sort -k1,1 -t ';' -T . | java -Xmx128m MergeCounts > /dev/null
done
end_time=`date +%s`
echo execution time was `(expr $end_time - $start_time)` s.
```

```
#!/bin/bash
start_time=`date +%s`
for i in 1 2 3 4 5 6 7 8 9 10
do
   cat RCV1.small_train.txt | java  NBTrain > /dev/null
done
end_time=`date +%s`
echo execution time was `(expr $end_time - $start_time)` s.
```

Q3

We can do feature selection.
For example, use correlation feature selection. The feature with higher correlation will be more informative.