10605 BigML Assignment 4(b):

Naive Bayes with Hadoop API

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

By using the hadoop streaming, we can run map reduce job with any executable as the mapper or reducer as long as our executable reads from stdin and output to stdout.

Q2

With streaming mode, I can choose any language to implements map reduce.

With no-streaming mode, I got more functionality. For example, I can use a counter as mention in Q3.

Q3

In the map phrase, we use word as a key. So when a reduce phrase is invoked, it means there is a unique word. Then We can use output.getCounter() in the reduce function to increase the count of a global variable. So in the end, we can get the vocabulary size.

Q4

Use another map reduce job to process testing data. For a testing record, output {document id, word}. Join the output together and then do the classification.