BigML Assignment 4b: Naive Bayes with Hadoop API

Runyun Zhang
runyunz@andrew.cmu.edu

1. **Compare the Hadoop Streaming mode with the Hadoop API. What are the major differences? (5 points)**
   a) Hadoop Streaming supports any executable or script while, Hadoop API only support java programming.
   b) Using Hadoop API, the input/output format is pre-defined, while in Hadoop streaming the input/output are standard I/O flow.
   c) Using Hadoop API mapper/reducer class is invoked each time to deal with each record, while in Hadoop Streaming mode, the program is called once to deal with all the data.
   d) Apparently Hadoop API has better functional support, such as access HDFS.

2. **For the streaming and non-streaming modes, when would you choose one over the other? (5 points)**
   I will use streaming to test idea and draft new application. When the performance turn out to be a consideration, I will choose Hadoop API. If I need to write application with extra access to HDFS or need other utility support, I will also choose Hadoop API.

3. **In the Hadoop version of naive Bayes, how would you estimate the vocabulary size? (5 points)**
   Set a static hash map in the map class might work. If it doesn't, we might need to write a file on HDFS directly to save the hash map.

4. **How would you design the testing pipeline of large-scale naive Bayes classification using Hadoop? (5 points)**
   Separate test cases into different shards and distribute them on different worker machines. Maintain word count results on HDFS for each worker to access during the classification process. The map phase will be retrieve the count for each token in the test case, while the reduce phase will aggregate the result and perform the computation.

5. **Your answers to collaboration policy (on the first page of this handout).**
   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.