Assignment 4b

Chen Sun
chens1

1. Compare the Hadoop Streaming mode with the Hadoop API. What are the major differences?

The major differences are that you have to implement the **Mapper** and **Reducer** under the Hadoop API specification. They can only executable java files. However, streaming could run under any executables as long as they can read in data and output data.

2. For the streaming and non-streaming modes, when would you choose one over the other?

When the job is relatively easy with existing nodes, such as we wish to take the input and do the word count. We can use streaming to do the work.

Also, when the input data is well formatted, and no or few exceptions would produce throughout the work, I would use streaming.

Also, when the data only rely on the same key to do some staff, it is better to do that in streaming, since the scripting language is simpler.

On the controversy cases, when the mapper and reducer need more than counting, and when the key is not the only determining factor to consider, mapper and reducer should be using hadoop API.

Also, configurations can be passed through mappers and classes to control some executions. This can't be done via streaming.

3. In the Hadoop version of naive Bayes, how would you estimate the vocabulary size?

To estimate, we can estimate by the totalRecords / 4, based on the assumption that each label has similar frequency distribution.

We can also write each mapper's voc size to configure, and print them out. And sum them up and estimate the overlapping rate, then we can estimate the result of the vocsize.

4. How would you design the testing pipeline of large-scale naive Bayes classication using Hadoop?

Read in the vocabulary table, and the test instance. Map by the label and key's initial letter to different nodes, and set up table for each nodes.

And for each doc word to test, map them to mapper according to their first letter. And calculate the prob value. And finally calculate the prob of each label and determine the label.

Also, if there are many test instances, we can also map them according to test id.

5. Answer the questions in the collaboration policy on page 1.

No.

No.