Assignment 1 - MLLD

Abhimanyu Vatta [1]

1. Problem Statement

Given DBPedia training dataset, it is required to train a Multinomial Naive Bayes Classifier for classification of DBPedia development and testing datasets with in-memory and streaming based Map Reduce approach.

2. Pre-Processing

Preprocessing on Dataset

1. Documents were tokenized into seperate words.
2. Words were lowercased to avoid repetation due to case sensitivity.
3. Words were stripped of any punctuation marks to avoid repetation due to occurence of a word with punctuation mark.
4. Stopwords as given in NLTK library, with some additions were removed as they in general would not help determining any context attached to a document (label in this case).
5. Smoothing parameter 'm' was tested with different values and finally set to 500.

3. Model Description

There are two pass of Map-Reduce for both train and test. Cache to be shared by all reducers is created using a third Map-Reduce.

1. **Training** - First pass produces events like like "Y=label X=word word_count". In second pass the output is such that the key is a word and values are count of word for a particular label. This output and test data are used as input for testing.
2. **Testing** - First pass produces key as "label of data" and values as "words and its counts". Second Pass uses this output and test data to predict the labels of the test data. The final output is the document and its predicted labels, number of correct predictions and total number of data.

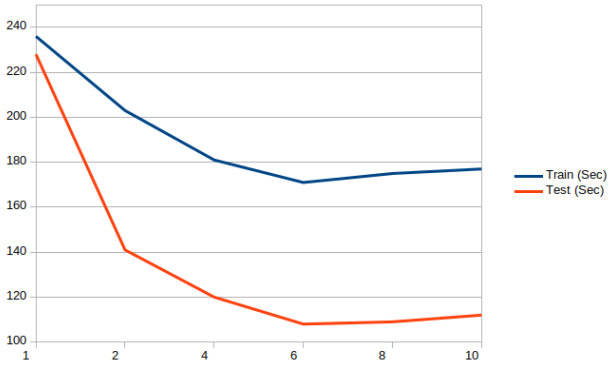

Figure 1. Run time in sec (Y-axis) for number of reducers from 2 to 10 (X-axis)

4. Accuracy

	Development(%)	Test(%)
local	97.5	97.81
hadoop	76.56	79.87

5. Run time

	Train(sec)	Test(sec)
local	33	42
hadoop(10)	201	112
hadoop(8)	165	107
hadoop(6)	171	108
hadoop(4)	181	120
hadoop(2)	203	141
hadoop(1)	236	238

Exponential decrease in time is notices when using more reducers with lowest at 6. Then it starts to flatten which can be see in Figure 1.

[1]Indian Institute of Science, Bangalore. Correspondence to: Abhimanyu Vatta <v.manyu@gmail.com>.

6. Number of parameters

1. Naive Bayes - 279548
2. Map-Reduce Version - 396845

REFERENCES

Cohen, W. Naive bayes and Map-Reduce. Technical report, 2015. https://www.semanticscholar.org/paper/Naive-Bayes-and-MapReduce-Cohen/a397eb310921897ef8a140668b623de618da7606.

Noll, M. G. Writing an hadoop mapreduce program in python, 2013.