Assignment 1 - Naive Bayes on Hadoop

Aditya Sharma [1]

Abstract

Here, I present and contrast two different implementations of the Naive Bayes algorithm for document classification - one on single threaded sequential environment and the other in a distributed manner using the MapReduce (Dean & Ghemawat, 2008) framework.

1. Preprocessing

I used code the following page https://www.kdnuggets.com/2018/03/text-data-preprocessing-walkthrough-python.html for preprocessing.

I removed stop words, converted to lower case and removed html code, among other things.

2. Parameters

Total number of parameters = (Number of words in the vocabulary $*$ Number of classes) $+$ Number of classes

Number of classes = 50 Number of words in vocabulary = 308552

Hence, number of parameters is 15,427,650.

3. Local vs MapReduce

It took 1 hour 07 min 36 sec for the local implementation to train, in contrast it took 2 min 10 sec for hadoop to train with 1 reducer.

Alpha, the smoothing hyperparameter was tuned to 0.05 for both implementations.

The local accuracies are higher because of more extensive preprocessing. For example, stop word removal was not done in hadoop version because of problems with nltk.

	Train	Dev	Test
Local NB	75.93	75.03	77.46
Hadoop NB	72.09	71.10	74.59

Table 1. Accuracy comparisons between local and MapReduce NB

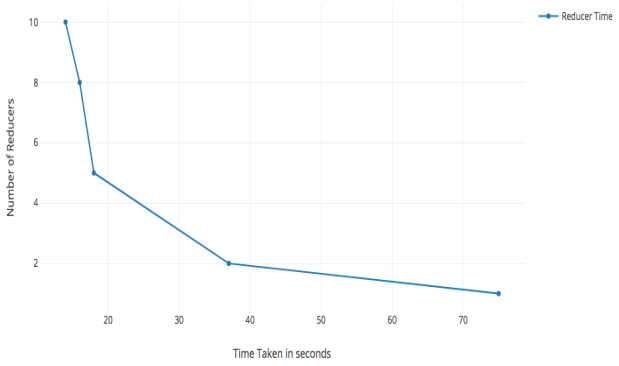

Figure 1. Training time with respect to number of reducers

The hadoop implementation was built on word count code provided on https://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/

4. Effect of Number of Reducers

As the reducers are increased the speed-up is not linear. We hypothesize that this is because the Map Reduce framework inherently has some slowdowns associated with it because of the way it handles message passing across servers.

References

Dean, J. and Ghemawat, S. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

*Equal contribution [1]Department of Computational and Data Sciences, Indian Institute of Science, Bangalore. Correspondence to: Aditya Sharma <adityasharma@iisc.ac.in>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.