CS294-1 Programming Assignment 3

Richard Hwang, David Huang

April 21, 2013

1 Introduction

A snapshot of Wikipedia is processed using Hadoop MapReduce. The goal is to train a classifier on the text content of articles to predict whether articles fit in a high level category.

2 Data Processing

We needed several different MapReduce jobs to process the data completely.

First, we stepped through the wikipedia file in its entirety and produced word counts. Our Mapper uses the WikipediaTokenizer and emits <token, 1>for each token. The Reducer then, simply sums the counts.

Second, we sorted the output from step one by count. In this MapReduce job, the Mapper simply inverts the pair, emitting <count, token>, and the Reducer is just the identity. This outputs a non-decreasing list of <count, token >pairs. We decided to use only the top 10,000 most frequent terms, because the data follows power law. We assigned each of these terms a unique id and put the mapping into a serialized HashMap that we could later access.

Third, we constructed the bag-of-words sparse matrix. Once again, we used the WikipediaTokenizer. For each token, we checked if it was one of the most commong words, and then added it to our sparse feature vector. The Mapper, then, emits <category, feature_vector>. The Reducer simply writes each line. Though we considered using BIDMat's saveAs function, we did not get a chance to try it.

3 Logistic Regression

Unfortunately, we were unable to train our model successfully. We were planning on training a logistic regression model, using stochastic gradient descent. As the assignment suggested, we were going to train a model local to each Mapper and have the Reducer average the model coefficients.
We did not reach this part of the assignment because of our lack of familiarity with MapReduce. Unfortunately, it took a while to get data processing jobs correctly running. Further, confusion on how to use Lucene's WikipediaTokenizer was a blocker for a long time; we have yet to find any thorough/reliable documentation or examples. Thus, though we had planned for it, we were not able to implement a logistic regression model using MapReduce.

4 Results

As stated in the previous section, we were unable to successfully train and evaluate our model.

5 Conclusion

We were able to process the Wikipedia XML file to construct a bag-of-words sparse matrix representation, but we ran out of time to construct our model and evaluate.