COMP9313 Assignment 1 MapReduce

z5014567, Senlin Deng

Introduction

MapReduce is the data processing component of Hadoop, which transform lists of input data elements into lists of output data elements. There are two main frameworks of a MapReduce program, which are Map process and Reduce process. Moreover, there are several intermediate steps (i.e. partitioner, combiner and shuffling-sorting).

In this assignment, Mapper and Reducer are the only two main technologies which are implemented. All inputs and outputs are stored locally instead of accessing from HDFS remotely. Moreover, it will be assumed that there are only one reduce in the assignment. And all intermediate steps (i.e. combiner) will be considered as black box to simplify this project.

Main Function

The template of this function is copied from Hadoop wordcount2.0 sample code. The main features for this function are:

1. Copy the arguments `<Ngram> <MinCount> <IN> <OUT>` as a string list and pass the information to `Configuration`.
2. Create a MapReduce Job. Set basic information to the job (i.e. `map -> combiner -> reducer` and `output(key, value)`).
3. Set the input and output directory. They are local storage given by arguments `<IN> <OUT>`

Mapper

In this assignment, customized mapper *NgramMapper* is designed to solve the problem, which extends the basic *Mapper* class. The input key and value pairs are using the default format, while the output value is customized type which extends *writable.*

FilenameIntWritable is a customized Hadoop value which contains two fields:

1. `private Text filename`
2. `private IntWritable frequency`

For each pair, filenames are valued by FileSplit and frequency is initialized by one. And the key of each pair is the Ngram. When Ngram is 2, concrete example:

`<"a hadoop", {1, "file01.txt"}>`

And the pair will be sent to Reducer when it has been set by:

```
context.write(word, new FilenameIntWritable(one,filename));
```

Combiner and Reducer

Customized reducer and combiner are called NgramReducer and NgramCombiner respectively, keys are `Text` type while values are `FileNameIntWritable` while type. Aggregates the occurrence as well as the file name list. Only output the Ngram has an occurrence greater than the `<MinCount>` are printed.

(more comments showed on the code)