Calculating Word Co-Occurrence

Eclipse project: word\_co-occurrence

Java files:

WordCoMapper.java (Mapper)

SumReducer.java (Reducer from WordCount)

WordCo.java (Driver)

Test directory (HDFS): shakespeare

Exercise directory: ~/workspace/word\_co-occurence

JAR File: wordco.jar

**In this exercise, you will write an application that counts the number of times words appear next to each other.**

Test your application using the files in the shakespeare folder you previously copied into HDFS in the “Using HDFS” exercise.

NOTE: This implementation is a specialization of Word Co-‐Occurrence as we describe it in the notes; in this case **we are only interested in pairs of words**

**which appear directly next to each other*.***

1. Change directories to the word\_co-occurrence directory within the exercises directory.
2. Complete the Driver and Mapper fixme files; you can use the standard SumReducer from the WordCount project as your Reducer. Your Mapper’s

intermediate output should be in the form of a Text object as the key, and an IntWritable as the value; the key will be word1,word2, and the value will be 1.

**Extra Credit**

If you have extra time, please complete these additional challenges:

**Challenge 1**: Use the StringPairWritable key type from the “Implementing a Custom WritableComparable” exercise. If you completed the exercise (in the writables project) copy that code to the current project. Otherwise copy the class from the writables solution package.

**Challenge 2**: Write a second MapReduce job to sort the output from the first job so that the list of pairs of words appears in ascending frequency.

**Challenge 3**: Sort by descending frequency instead (sort that the most frequently occurring word pairs are first in the output.) Hint: you’ll need to

extend org.apache.hadoop.io.LongWritable.Comparator.

**END**