Mining Big Data - Map-Reduce

Assignment 1 - Hadoop

M. Vincent & A. Phansalkar
a1148120 &

School of Computer Science, The University of Adelaide

March 27, 2020

Report submitted for **COMP SCI 3306 Mining Big Data** at the School of Computer Science, University of Adelaide

Exercise 1

Exercise 2

The output from the tutorial example can be found in the folder Exercise 2/WordCount/Output. The source code, input file, and the executable JAR file used to run the Hadoop job in pseudo-distributed mode, can all be found in the folder Exercise 2/WordCount.

Exercise 3

Exercise 4

Parts 1 and 2

To answer the first two parts of this question, we developed a new MapReduce job. In the map function for this job, the length of each word is used as the key in each key-value pair, rather than the text of the word itself. The value remains 1, as in the WordCount algorithm in the tutorial. The reduce function then sums the values for each key with the same number, giving the final output.

The output from this MapReduce job can be found in folder Exercise 4/WordLength/Output. This output was used to answer the questions below.

1. There are 3102 words of length 10 in FirstInputFile.

2. There are 7019 words of length 4 in FirstInputFile.

3. The longest word in FirstInputFile is 21 characters long, and it appears once.

4. There are 306 words of length 2 in SecondInputFile.

5. There are 105 words of length 5 in SecondInputFile.

6. The most frequent length in SecondInputFile is 0 and it appears 297337 times.

Parts 3 and 4

To answer Parts 3 and 4, we used the original WordCount MapReduce job to generate a unique list of the words appearing in each of the input files, and then took this as the input for the WordLength job.

The WordCount job that was developed using the tutorial produced a list of words that included punctuation and other characters. When this output was run through the WordLength job, it showed a very large number of zero and single character words. Obviously this is not possible, so we tried modifying the WordCount job by removing all characters that were not alphabetical. This produced output that at least passed a simple sense check.

The output generated using both of the MapReduce jobs together can be found in the folder Exxercise 4/WordLength/Output. This output was used to answer the questions below.

1. There are 2263 words of length 10 in FirstInputFile.

2. There are 1911 words of length 4 in the FirstInputFile.

3. The most frequent length in FirstInputFile is 7, and it appears 4908 times.

4. There are 1819 words of length 5 in SecondInputFile.

5. There are 65 words of length 2 in SecondInputFile.

6. The second most frequent length in SecondInputFile is 8 and it appears 2800 times.

Exercise 5