Burak Araz

1818939

CEng 495 Cloud Computing Programming Assignment 2 Report

I used 2.5.2 Hadoop Version. In this assignment, we have two task. First task is finding the similar words. Before starting to map reduce operation, I established the configuration and set methods in the main. In the map function, I took the sentence id to the rowID variable and in the while loop, listOfElement was filled with the element of single line input. Then to delete replicate element in the list, I used practical approach which I came across in the web. In this approach, first I filled the HashSet with the element of List, then I filled the List with the HashSet values which eliminate duplicate values. After deletion, I combined two words in the line with the for loops and words become like (word1,word2) or (word2,word1) depends on lexicographic order. I created all combinations of tuples in the line. The last thing that I did in map function is that write these keys and values to the context like <(word1,word2) 1>.

In the reduce function, I summed the same word tuples and if the result is equal or bigger than the threshold value which was given by input as a “K”, I wrote the result to the context. And the result is like <(word1,word2) 3>.

Our second task is word similarity histogram. We need to print the number of n similar for each value in a new line. Firstly, I called again first job but this time the name of the output file was inter_output_path because I defined second job in the main and I gave this intermediate result as an input to the SecondMapper and the SecondReducer. After I got the intermediate_result which was same with the first task, in the second map function firstly I called the itr.nextToken() to get the value <(word1,word2) 3> like <3>, and wrote these values to the context.

And in the SecondReducer function I summed the same value and wrote the result to context which was output file given by the input. The result is like <1 8, 2 1, 3 1>. After these operations I deleted the intermediate file from the hadoop file system. In the second task the job order is FirstMapper -> FirstReducer -> SecondMapper -> SecondReducer. Also, before starting the second task, first task was being waited to finish.