Report on project - word_co-occurrence

Subhrajyoti Pradhan

U79333962

Introduction -

In this project, we attempt to write an application that counts the number of times words appear next to each other.

Description of Driver -

In the driver, we establish the classes outputkey, outputvalue, mapper and reducer according to our needs.

Description of Mapper -

For each line that the mapper receives, it splits it into words. It then checks for first iterations of words, which are then added to a buffer. For subsequent iterations, it outputs as a pair of the words as the key and the number 1 as a value.

Description of Reducer -

In this application, the reducer is to find the sum of the values for every key.

Data Flow -

The mapper reads each lines. The keys are subsequently generated (two words joined by a comma). The reducer counts the time everything appears.

Test and Output -

We first check the file for errors in Eclipse and compile it to a jar file. We unzip the .tgz file and create a directory in hadoop. We execute everything using command -

Hadoop jar coo.jar stubs.WordCo /user/training/coo coo1

Sample Output -

eat,chickens 1

eat,do 1

eat,dried 1

eat,each 1

eat,get 1

eat,go 1

eat,grass 2

eat,him 2

eat,his 3

eat,honey 1

eat,husks 1

eat,i 3

eat,in 2

eat,iron 1