COMP9313 - Assignment1

Han YANG – z5140181

WordCount

In my solution, there are two main classes to solve the wordcount problem, TextMapper and TextSumReducer. In TextMapper, function map transforms the input into key-value pairs. In TextSumReducer, function reduce aggregates the values for each key.

There are 4 program arguments, N(ngram), minimum count, input file directory and output file directory.

Given a set of text files in input, the mapper firstly splits text into individual words, then finds all ngram as keys, and the value for each key is its count (in this step the count is 1) and the filename this ngram appears, then writes these key-value pairs into HDFS. Next, in shuffle and sort phase, the framework fetches the output of all the mappers and groups it by keys. Then each reducer computes the number of times of each ngram(key) found across all files and the list of filenames of files where this ngram appears, and writes the result that its count is equal or greater than minimum count into HDFS.

The whole program process can be seen as below:

(arguments: N(ngram) = 2, minimum count = 1)