COMP9313 - 2019T2 Assignment #1 (MapReduce)

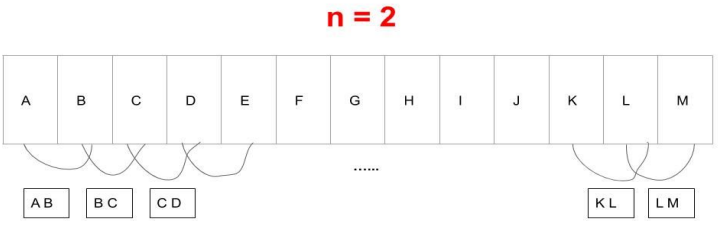
z5147810 Meng Lingxu

Environment: CSE Lab environment, Java 11, Hadoop 1.2.1
Working period: 2019.06.24 - 2019.06.30

1. Input Processing
Totally four input arguments are required:
args[0]: The value N for the ngram.
args[1]: The minimum count for an ngram to be included in the output file.
args[2]: The directory containing the files in input.
args[3]: The directory where the output file will be stored.

An error catcher is set in the main function, if the number of input arguments is not equal to 4, program will exit and throw an exception. For convenient, the program will delete the former output folder and create a new one when the program runs successfully.
Use *setInt* to store args[0] and args[1] to the configuration that set to the job. Use getConfiguration to get values from context to get the value that map/reduce functions need.

2. Map



Actually, grading all n-grams from a long sentence is sliding an array of strings. Like the figure above, get all slices from A to M, 2 is the slice length.
Firstly, the program used a stringTokenizer to read the content from input files. Then put all stringTokenizer object elements into a new ArrayList for the sliding part. The map output format would be <Text, TimeAndFileName>.

3. Reduce
Because the input folder may includes not only one single file, so only one reduce progress is not enough. The program adds a combiner to deal with the aggresion part. All K, V pairs would be summed during the combiner and the true reduce function will store all K, V pairs in a hashmap. The reduce function will traversal the hashmap and grab pairs those has a sum larger than the threshold(args[1]).

4. Producing Output
Aiming to write the exact output format as the output template, the program defined a customized Hadoop writable structure: TimeAndFileName<IntWritable, Text>, which is a combination of IntWritable and Text. The IntWritable part is the time that the phrase shows among all input files. The Text part is the filenames of the phrase sources, which is a set, which means each filename only show once, though not sorted.