COMP 9313 - 2019T2 Assignment 1
Student: z5146092

The Implementation of Mapper

At first, the content of file is split by spaces and then store into a string array. A file name is also recorded by using FileSplit function. Then a N-gram queue implementation (see Fig 1) is used to form the N-gram word. When the N-gram word is formed, it is written into the mapper context as a key, and the file name (in Text format) is written as a value.

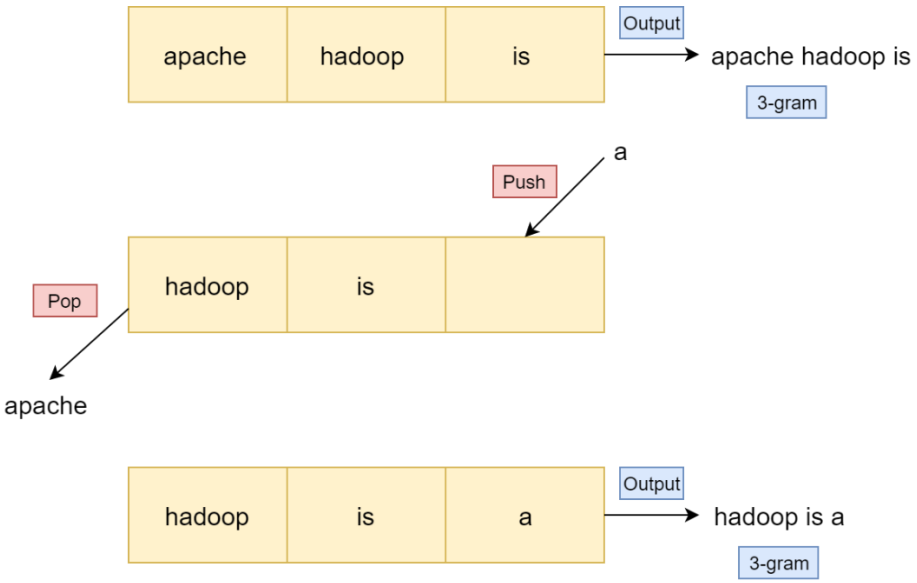

Fig 1. N-gram Queue Implementation

The Implementation of Reducer

Since the mapper function sends keys and values in Text format, the word count is a part of work in Reducer function. Once, the Reducer receives a key, it will be followed by several values (iterable Text) of file names. The whole procedure is shown on Fig 2. With the TreeSet data structure used, duplicate filenames can be eliminated, and filenames are automatically sorted. Then with the check of minimum count, the unsatisfied n-gram words are abandoned. Finally, the result is generated by concatenating strings and written into context.

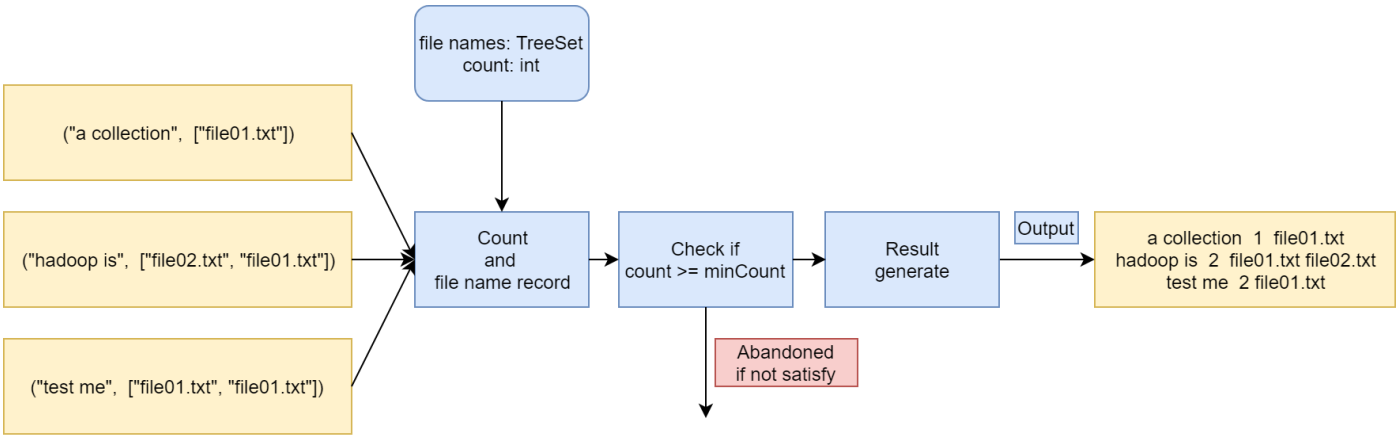

Fig 2. Reducer Implementation

Other Implementation

In main function, the MapReduce job configuration is set, and the class details are given to the MapReduce API. Also, the command-line arguments are parsed with exception handling.