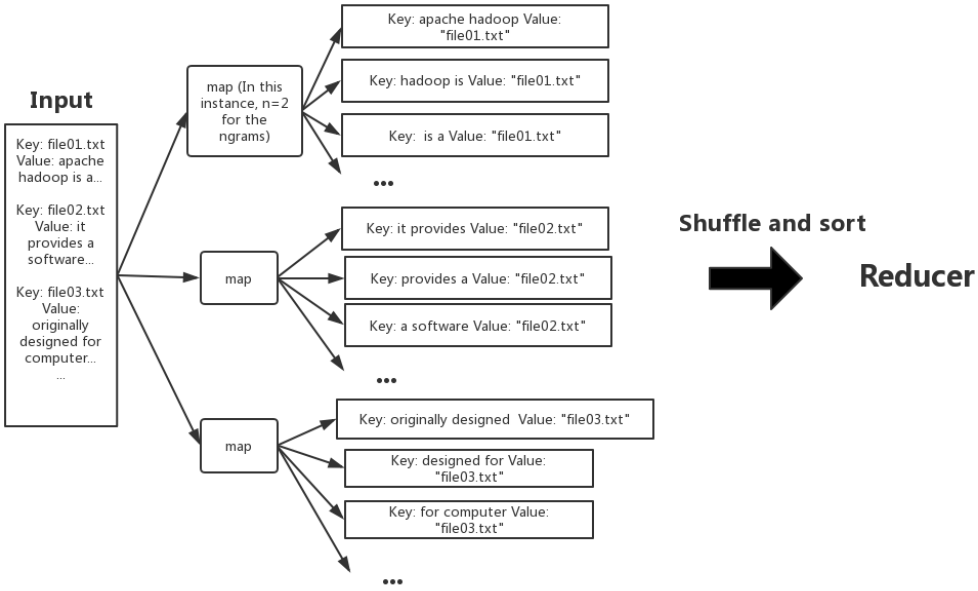
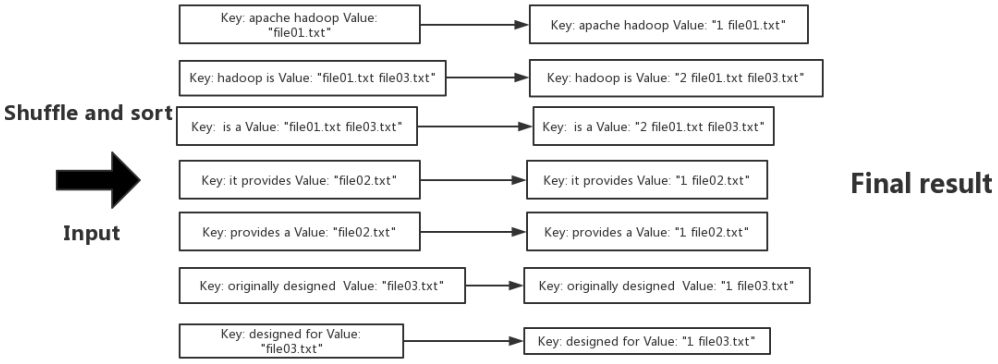
COMP9313 - 2019T2 Assignment #1 (MapReduce)

Student Number: z5195349 Student Name: Wenxun Peng

There are three main parts in my code. First, it's the **Mapper part.** In map function, using the original key, files' names to extract the contents (getting the ngrams) as value. And then, I use ngrams as key and the corresponding file name as their value to give the Reduce function. If there are same bigrams in the same file, for example, two "hadoop is" in file01.txt, it will produce two identical messages ("hadoop is" is the key and "file01.txt" is the value).



Second part is the **reduce function.** Reducer gets the input value from map and the data will be shuffled and sorted. Therefore, in reduce function, I count the bigrams and detect their file name. For example, if there are two same bigrams, such as "is a", they appear in different file, file01.txt and file02.txt. The counting will be 2 and the final result will be "is a 2 file01.txt file02.txt".



The last one is the **main function.** In addition to the various declarations and definitions introduced like in lecture, an additional part is added to check whether the input is correct. (args [0] and args [1] must be positive integers and whether args [3] and args [4] files exist).