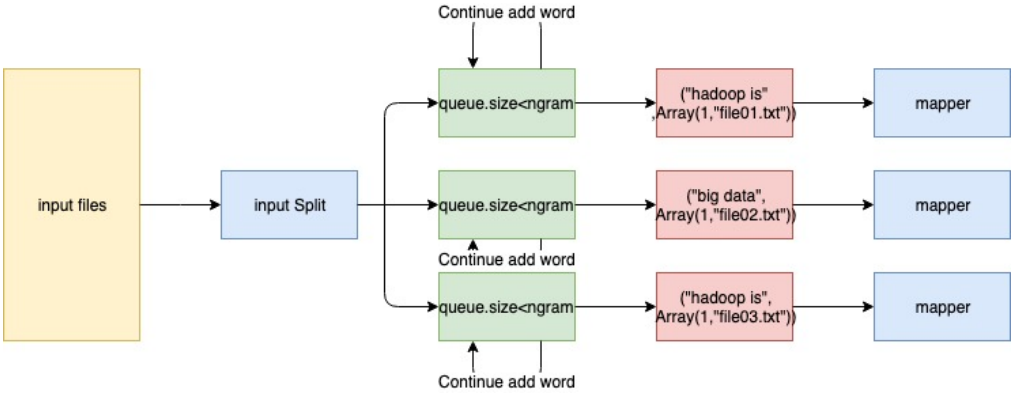
Comp9313 Ass1

Student ID: z5184142

Student Name: Jiachen Li

To this assignment, I modify the word_count procedure from the lab1. First I defined a class which is called Array, which contains a parameter amount and another file_name. This class is used as the value of each key in the mapper.

Next Defining a variable which is called ngram (Datatype: Integer) in TokenizerMapper and the value of ngram comes from args[0], and then defining a queue in the mapper class. What will happen next is that there will be a loop continue to add the context of file to the queue word by word. Once the queue is full, then the context in the queue will be the key in mapper and the value will be set as amount 1 and file_name = the file name. Then they will be recorded to the context. After that the first element will be removed from the queue, and the next one adds into the queue, this continues to happen until the loop ends.



The output of the mapper will be given to the reducer. In reducer there will be a variable which is called count, it is used to filter the sum of whose amount value does not meet the requirement. Also there will be a list to receive the value of all the file_name of each key. Once the sum of the amount value meets the requirement, the list will be converted to a set and eliminate all the repeated file_name, then converted to the list again. After that, all the processed data will be recorded to the output file.