COMP9313 Assignment-1

Amritesh Singh
z5211987

Documentation

The MapReduce program has 3 components, each belonging to a different class:

- **Main:** The main function acts as the driver for the program, creating a Configuration object and a Job object. It sets the job's jar file by finding the provided class location, provides the mapper and reducer class names, sets the Configuration object with the data type of output key and value for map and reduce, and specifies the input and output directories to be fetched from the command line. It then submits the job to the cluster and waits till it is finished. The values of **N** and **mincount** are set in the configuration data.
- **Mapper:** The mapper class reads each file in the input directory, splits it into words using a string tokenizer, and stores all the tokens in a list. Subsequently, the ngrams are derived according to the specified value of **N** (ngram length), and the intermediate key-value pair **<ngram, file_name>** is formed using context. The file name for each ngram is found using input split on context object, and then getting the file path and subsequently the file name from it.
- **Reducer:** The reducer class iterates over the list of values corresponding to each key, and stores the length of the list and the non-duplicate values in the list (found using hash set) into a string, which is then used as the **result** value. The key-value pair **<ngram, result>** is then formed using context. Only the pairs with value list length greater than or equal to the specified **mincount** are included.

MapReduce program process (N = 3, min count = 1)