Michael Parrilla

John Sweeney

Denny Sabu

Assignment 1 Report

The first step in our system is to accept arguments from the command line and pass them to our map function. After the map function checks all inputs, and ensures there are no invalid inputs, it opens the input file, breaks the file up into tokens, sets each token to lower case, and removes the specified delimiters(.,:;?!-“) and stores each token in a vector. Then a two-dimensional vector is initialized to hold each token and its counter and from this two-dimensional vector a number of sub-vectors, specified by input from the command line, are produced. From here the sub-vectors are given to threads or processes, again specified by the command line, and within these blocks the tokens are counted either by the word-counter function or number-counter function which is also determined by command line arguments. As mapping is preformed, across either threads or processes, vectors of pairs are sorted and then written to shared memory. When each thread or process finishes they wait for the others to finish before moving on to the reduce phase. Once the reduce function is called each pair in shared memory is written to a vector of pairs. These pairs are then split based on the number of reduces given by the command line and are sent to either processes or threads. In their respective threads or processes the like pairs are combined, each combination increases the counter in the pair, this is done until each pair is unique, then the pairs are sorted again. Once all threads or processes have finished the separate vectors of pairs are put together and reduced again, the final vector is then written to a file whose name is specified by the command line.

//Evaluation goes here

One of the difficulties we faced for this project involved using shared memory. It was a challenge to figure out how to get data back from each individual process from the map to reduce phases. One way we solved this was to increment the address of our shared memory, this made sure each piece of data written to shared memory was some what indexed, and also made it such that the data on shared memory would not be over-written by other concurrent processes.