Things to do:

Figure out how to randomly sort the users - done

Clean the data – run the data through my cleaning program to clear out unwanted and fluff words - done

I will save the cleaned data into their own files - done

I need to sort the training set and testing set save these into their own files - done

Figure out how I want to save the data. Do I want to have a male file and female file? Do I want to just save the data? Meaning do I want to just run it all, separate the data, count each word and just save those numbers to a file? What if it takes a long time and I mess up? I should probably do each phase and save the data.

Do I really want to separate male and females into their own files according to each set? I do not know if I do. It could cause issues with data handling. I need to think it over before moving forward.

After cleaning the data, I need to sort the data by male and female and save them into a new file

Once sorted, I need to count the usage of each word for males and females and save the counts in a file

I also need to take same words from males and females to get a total of how often it appears

I then need to work on the algorithm