CS3907-80/CS6444-10 Big Data and Analytics

Class project #1

R and Graph Analytics

1. Preprocess the data
2. Load text from files



In this R file, we extract text from the files. We use the gsub() and grepl() function to extract specific fields from the text. The data are collected in a data frame. Then we use na.omit() to delete the records which don’t have receivers. We also use the sqldf() function in the sqldf library to group the records by message id in case there are duplicate emails. Finally, we save the result into “email\_data.RData”.

1. Delete the less important records



In this R file, we separate the receivers and delete the less important records. Some email may have multiple receivers, therefore we split them using the strsplit() function and create rows which only have one receiver. We use count() function in the plyr library to count the frequency of senders and receivers, and then we delete the emails which are below mean. Finally, we save the result into “processed.RData”.