Project Progress Report
Personalized Email Search Engine (Intelligent Browsing)

Team Member: Ashley Yeah (ayeah2@illinois.edu)

For my personalized email search engine project, I have currently completed almost half of the tasks from the proposal. I have been able to download my entire gmail inbox up to a week ago through Google's backup functionality, which saves all emails in an mbox file. Then I was able to extract the sender, subject, and body of each email. Since the search functionality will occur on the body of the emails and I am using BM25 as my search function, I then cleaned, normalized, lemmatized, and tokenized all email bodies so that they represented just a list of words since BM25 is a bag of words retrieval function.

Now I just need to place the tokenized corpus into the BM25 ranker function, which can be done with the rank_bm25 library in python. Then I will try to tune the parameters of the function to maximize the search results. I will also need to work on the frontend implementation of how this search engine will face the user. Initially, I wanted to do a chrome extension that would connect the searched email with the actual email on the gmail site, but that is looking to be hard because there is no reference to the link of the email once all the emails have been downloaded in batch, so I may pivot and just instead make a traditional minimal web page for the search engine.

Some challenges I have faced so far have mostly been related to cleaning the texts of the email. I never realized how many different formats emails can come in from plain text to css styled, and different encodings, so having to extract the body of the email from the download was actually much more challenging than I expected. As mentioned already about my plans with the frontend, it does seem like my original plan will be a little bit harder to implement as I have no links to each email in my downloads, so I will have to try to find a way to work around that.