CS410: Final Project Proposal

Fall 2021

Topic [Theme5: Free Topics]

Gmail Assistant: An intelligent bot to help users navigate a deluge of emails in their inbox.

Team

Name	NetID
Ameet Deulgaonkar (captain)	ameetd2
Praveen Purohit	purohit4

Problem

For an individual, emails continue to be an important mode of “text” based communication in spite of the arrival of several modern communication platforms like social media (twitter, facebook), slack or WhatsApp especially when it comes to point to point communication (B2C or C2C). The size of an individual’s inbox is growing more than ever. It is likely that we are receiving 10 times more emails than we were 5 years ago, but most email services do not provide any features for users to cope with this deluge of email. It is really hard for users to separate useful information (needle) from noise (haystack) using the basic key-word based search capabilities that most email services provide.

Solution

Our proposal is to use probabilistic topic modelling (either LDA or PLSA) and mine the topics in an user's inbox (e.g. finance, travel, promotions, newsletters, invitations etc.) and topic distributions in the emails and organize the emails using this meta-data. Using this information we intend to solve the following key use cases.

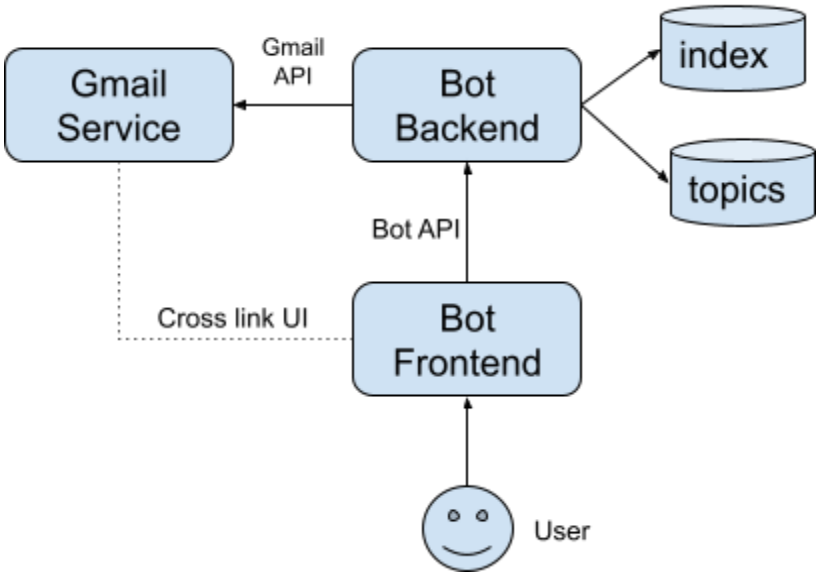
1. Automatically organize emails into categories such that the user can quickly locate emails of interest instead of scrolling the emails monotonically in chronological order or using naive keyword based search.

2. [STRETCH GOAL] A more intelligent search that uses a “combination” of state of the art ranking functions like BM25 and the mined topic information. We intend to first apply BM25 to obtain ranked emails and then improve the results by matching the topic(s) of the query (when possible) to the topic(s) of the email.

Implementation

We intend to implement the bot as a service (possibly hosted in the cloud) which downloads the user's emails and performs the text mining as mentioned in the solution section. The bot then stores the mined medata (i.e. topics and topic coverage) in a database. The bot will offer a simple WebUI for the users to register their gmail account credentials and get a “more” “organized” view of their inbox. An ideal way to build this front end would have been to have it either as a gmail plugin or a chrome extension, but we chose this approach due to paucity of time.

Here is the high level architecture of our bot :



Components

1. Backend will be built in python using metapy (search and index) and gensim (for topic modelling).
2. Frontend will be either a WebApp (e.g. react.js app) or any visualization tool like PowerBI or Tableau to view the topic modelling results.

Work Estimate

Task	Estimate
Backend: data pre-processing and storing	16 hours
Backend: topic modelling	16 hours
Backend: search and index [STRETCH GOAL]	8 hours
Backend: APIs	8 hours
Frontend: Organized email view	16 hours

Frontend: Basic email list view	8 hours