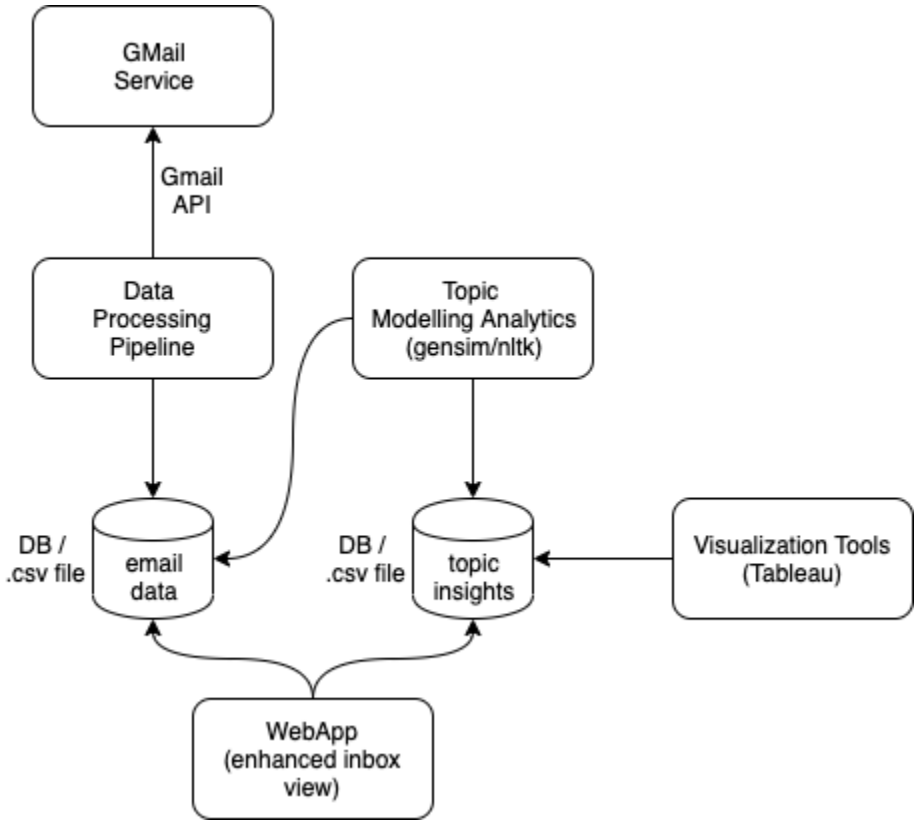
CS410: Final Project Progress Report

{ameetd2, purohit4}@illinois.edu

The goal of our project is to build an enhanced gmail client that helps users manage the deluge of emails in their inbox in a much more efficient way by employing text mining techniques learnt in this course. Specifically, our goal is to mine dominant topics in the emails in a user's inbox and automatically categorize (assign labels) them mined topics.

We have completed a detailed design of the solution, broken down the tasks among members of the team and started implementation of core modules of our solution. The following diagram captures the design of our solution.



The data processing pipeline module will fetch emails using the Gmail API, extract text from the body and perform all the data cleaning operations like stemming, removing stop words etc and store the keywords in a .csv file (or a database).

The topic modeling and analytics module will read this file and construct a vector for each email (treated as a document) and perform topic modelling using the gensim library. It will store the discovered topics for each email in a .csv file (or database).

The email client WeApp will display an enhanced view of the user's inbox by grouping emails by topics and show a summarized view that will help users read mails that are of most interest to them.

Implementation Status

The detailed task breakdown and their current status is captured in the following table.

Task	Owner	Status
Data Pipeline Implementation		
OAuth 2.0 authorization workflow for accessing GMail	ameetd2	Completed
Retrieve mail from GMail Service	ameetd2	Completed
Text extraction from mail	ameetd2	In Progress
Data cleaning and curation	ameetd2	Not Started
Storing data	ameetd2	Not Started
Text Analytics Engine		
Feature engineering	purohit4	In Progress
Topic model Implementation	purohit4	Not started
Storing results	purohit4	Not started
Front End (Visualization)		
Visualization of topic Insights in Tableau	purohit4	Not Started
WebApp for enhanced inbox view	ameetd2	Not Started

Challenges

1. We faced some challenges in getting the OAuth 2.0 authorization flow working to access a user's Gmail inbox due mostly to the multiple ways of accomplishing this (e.g. front end vs. backend based implementation) and some of the methods being deprecated by Google Identity service. After a lot of wading through the Google Identity service we were finally able to get the access token to access a user's inbox.

2. We are also facing some challenges in extracting text from emails (due to various mime types). We may restrict our solution to emails of certain mime types only if we are not able to resolve this in time for your submission.