Email Summarization Proposal

This project applies natural language processing methods to solve email text summarization using the Enron email dataset. Email overload is becoming a large time sink at work with the average office worker receiving between 40 to 90 emails a day. A system that can create concise and coherent summaries of all emails received within a timeframe can reclaim a large amount of time in the workplace. The Enron email dataset provides 150,000 emails from 150 users sent in a corporate environment. This dataset provides an environment where email summarization would be useful for the individual user and to someone looking for insights into the Enron scandal.

The Enron email data is provided for free at the following link: https://www.cs.cmu.edu/~./enron/

I would approach this as a supervised machine learning problem. While the Enron dataset has no human summaries of the email, I will use the much smaller, BC3 Corpus dataset to check the accuracy of the email summaries with the ROUGE metric. https://en.wikipedia.org/wiki/ROUGE_(metric)

The BC3 Email Corpus can be downloaded for free here:
https://www.cs.ubc.ca/cs-research/lci/research-groups/natural-language-processing/bc3.html

I will attempt an extraction-based summarization, where I find patterns in the frequency of key phrases and piece them together as a supervised machine learning problem. I will be trying to predict a concise summary of emails over a given timeframe for a given inbox checked against human summaries. The predictors would be the subject of the email, as well as the body of text itself. I use the TextRank method for NLP. https://nlpforhackers.io/textrank-text-summarization/. Which leverages word vectors generated by deep neural networks. Articles have described promising results for other deep learning methods, which I could also attempt to implement after the initial traditional approach.

The final deliverable will is a web app that can be used to explore the inboxes of the Enron executives. It can give a general summary of an inbox over the whole timeframe of the dataset. Then a tool could be used to constrain the time to a week for example, to simulate what it would look like if a user only wanted to summarize their inbox over the past week.

With the initial unsupervised learning approach, a standard workstation works well. With the deep learning approach, additional GPUs may need to be requisitioned. I personally have the following specs for this project:

1. CPU: Intel i7-6800K @ 3.4GHz.
2. 32 GB ram.
3. GPU: NVIDIA GTX 1080

Additional Resources

https://machinelearningmastery.com/gentle-introduction-text-summarization/

https://en.wikipedia.org/wiki/Automatic_summarization