Final Project Proposal

E-mail Importance Ranker and Retrospective Summarizer

Text Information Systems - CS 410

Daniel Zurawski - dzuraws2@illinois.edu (Coordinator)
Vedprakash Mishra - vrm2@illinois.edu
Gassan Soukaev - soukaev2@illinois.edu

What is the function of the tool?
We plan to create a suite of utilities that will help users handle many e-mails. The utilities will be able to rank the importance of e-mails with respect to the content and sender. The utilities will also be able to summarize a corpus of e-mails received during a specified time period using a set of keywords.

Who will benefit from such a tool?
A working professional often receive too many e-mails to reasonably address within a day. Our tool will help these users by identifying the most urgent e-mails. People also occasionally forget the specific discussion areas of e-mails received during special periods of time. Our tool provides keywords to help users recall what was being accomplished during a time period. This summary can be an effective retrospective tool for better planning in the weeks to come.

Does this kind of tools already exist? If similar tools exist, how is your tool different from them? Would people care about the difference?
There exist machine learning programs that can detect if e-mails are spam. Our tool is different in that it ranks the e-mails based on importance, so it remains up to the user whether or not an e-mail is important. Also, our tool takes into account the sender of the e-mail, whereas similar existing tools only account for the content of the e-mail. This factor may indicate that our methods will yield greater accuracy in ranking e-mail importance and, thus, users will be more satisfied with our tool.

What existing resources can you use?
Currently, there are many available modules for text mining e-mails, including MeTapy, Sklearn, SpaCy and NLTK. These modules will be useful for identifying and ranking the e-mails. Also, the Enron Corpus of e-mails will be a useful resource for testing and training our utilities.

What techniques/algorithms will you use to develop the tool? (It's fine if you just mention some vague idea.)
We plan to use variants of the BM25 for ranking the importance of e-mails. Our tool will infer query terms by using e-mails that the user has reported as an example of an important e-mail. We also plan to use graph theory algorithms to identify the importance of an e-mail based on frequent sender/receiver relationships.

How will you demonstrate the usefulness of your tool.
We plan to offer our tool to multiple users, who will evaluate the effectiveness of the tools.

A very rough timeline to show when you expect to finish what. (The timeline doesn't have to be accurate.)

1. Week 8 - Project Proposal
2. Week 10 - Get e-mail ranker working
3. Week 12 - Get summarizer working
4. Week 13 - Progress Report - Get people to evaluate it
5. Week 14 - Fine-tune ranker and summarizer
6. Week 15 - Get people to evaluate it - Finish fine-tuning
7. Week 16 - Software Code Due - Software Usage Tutorial