CS 410 Text Information Systems

Final Project Progress Report

Which tasks have been completed? Which tasks are pending?

Here's the tasks from my proposal, annotated with progress:

1. **COMPLETED**: Assemble and clean data set: 2-4 hours
2. **IN PROGRESS**: Set up scoring/search engine with the data set: 3-5 hours
3. **NOT BEGUN**: Augment scoring with user profile initialization: 3-5 hours
4. **NOT BEGUN**: Augment with thresholding: 2-4 hours
5. **NOT BEGUN**: Augment with feedback/updating user profile and threshold: 10-12 hours

Are you facing any challenges?

Deciding on a framework/toolset and understanding the state of the art when it comes to systems that are relatively easy to set up is challenging, but it's not a blocker - just taking a bit longer than expected.

Specific feedback from reviews:

1. *There is no clear indication on the interface that will be provided*

   > Answer: I don't have plans to build any sort of GUI. The UI will be a command line, with text input for initial profile setup, and text output for recommendations. The main reason I'm not building a GUI is that it'd probably take a lot longer than the project target time allotment.

2. *Although since your current proposal doesn't incorporate new data pulls from Crunchbase, it does make me think it's more like a search engine than a recommender system. With recommender systems, there is typically a "dynamic info source" (Wk 6.5 slide 4) and the system has to decide whether to recommend a new item each time a new item is added. I suggest confirming with TAs as you may need to augment your project to include a web scraper that can pull fresh content from Crunchbase with some level of frequency.*

   > Answer: This is a good point. I don't think augmenting to include a web scraper is going to fit within the project target time allotment, but it's definitely something I'll plan to do if possible. If absolutely necessary, the idea of "fresh content" is something that could be approximated with a static dataset (eg by holding out a portion of the data and simulating fresh content).