CS 410 Text Information Systems

Final Project Proposal

“Company Similarity Recommender”

Team

Member and captain (just one): Chaaru Dingankar
NetID: chaarud2

Topic description

The goal of the project is to create a recommendation system that will suggest companies similar to a selected company (ie, a content-based filtering system). The data used will be publicly available company description data from Crunchbase, which is a site that houses information about many companies worldwide. This is an interesting use case because it’s important to be able to understand what companies are similar to others (a question that has a valuable answer whether you’re a sales person, a researcher, a job seeker, or a potential investor).

The approach I plan on taking is to set up a python search engine using either the MeTA or Whoosh libraries, and then augment it with the relevant systems to be able to support a content-based filtering system use case (that is, I would need to add the concept of a user profile that changes over time and is initialized a certain way). My expected outcome is being able to deliver companies that are relevant to a user’s profile. To evaluate my work, I will go through the recommendation process myself and see how feedback was incorporated, and whether the recommendations make intuitive sense.

Programming language

I plan on using Python primarily, for the data processing and delivery of recommendations. I may also use Scala if appropriate, for peripheral operations like data cleaning or loading.

Workload justification plan

Here’s some tasks that are part of my rough plan, and approximate time estimations for each:

1. Assemble and clean data set: 2-4 hours
2. Set up scoring/search engine with the data set: 3-5 hours
3. Augment scoring with user profile initialization: 3-5 hours
4. Augment with thresholding: 2-4 hours
5. Augment with feedback/updating user profile and threshold: 10-12 hours

All together, this represents at least 20 hours of work.