Course Project Proposal

Prakash Manden
Pmanden2@illinois.edu

Note: I thought did everything as per the documentation by 10/25 deadline. However, I was not aware that a separate proposal was to be uploaded until I saw some notes on Piazza today. Instruction are a bit all over the place, and I wasn't aware of the need, otherwise I would have submitted it before the deadline.

Improving a System - ExpertSystem Search

I plan to implement the following project mentioned in the 'Course Project Topic' (text copied as is from the document)

The ExpertSearch system (http://timan102.cs.illinois.edu/expertsearch//) was developed by some previous CS410 students as part of their course project! The system aims to find faculty specializing in the given research areas. The underlying data and ranker currently comes from the MP2 submissions of the previous course offering. You can read more about it here (Sections 3.6 and 4: Project are especially relevant). The code is available here. Below are some ideas to improve and expand this system. You may choose to integrate your code with the existing system, or borrow some ideas from it, or build your own systems/algorithms from scratch.

Automatically crawling faculty webpages

Recall that you developed scrapers for faculty web-pages in MP2.1, which, in general, can be a time-consuming task. So, the question is can we automate this process? Some challenges include:

- **Identifying faculty directory pages:** First, we need to identify the pages from where faculty web-pages can be mined. In MP2.1, we used faculty directory pages as the starting point to find faculty webpages. So, given a university website, can we automatically identify the directory pages? This can be posed as a classification task, i.e. classify a URL into a directory page vs. non-directory page. We have a huge resource of directory page URLs available in the sign-up sheet. These can be the "positive" examples. You can get a list of some random URLs online or crawl some other pages to get URLs(e.g. other URLs on the university websites, product websites, news sites,etc.). These would be the "negative" examples.
- Identifying faculty webpage URLs: Next, we need to extract the faculty webpages from the directory pages. This can again be posed as a classification task. Given a URL, can we identify whether it is a faculty webpage or not? We have a huge resource of faculty webpage URLs (available under MP2.3 on Coursera). These would be the "positive" examples. You can get a list of some random URLs online or crawl some other pages to get URLs (e.g. other URLs on the university websites, product websites, news sites, etc.) to get the "negative" labels.