Text Information Systems : CS410

Project Progress Report

"2.2 ExpertSearch System"

--

Navyaa Sanan (navyaas2)
Srivardhan Sajja (sajja3)

Progress made thus far

- Enlisted some common faculty directory URLs to give us a general sense of what they look like. This would later help us in identifying some common features.
- Determined some common features of faculty directory URLs which act as starting points for the classifier we are writing. We took the main features of URLs like word staff or directory in them and started with these obvious features. We also identified words and features that would never be part of a faculty directory page URL.
- Implemented a spider using the python scrapy package to recursively crawl through and identify all pages of a website with either 'faculty' or 'staff' in the URL, given the primary URL of the university (example: illinois.edu, berkeley.edu). Parameters such as time limit, page crawl limit, and results count limit can be set manually.
- Added some preliminary filters within the spider to not allow URLs with certain keywords such as 'mail', 'publish', 'calendar', 'research', etc., to make crawling process more efficient.
- Developed a website using the flask web framework to showcase model's results, and to amplify user experience.
- Looked at some research done in similar areas (classification of URLs) but were not able to find much that is directly relevant. The main issue we face is figuring out common features of a general faculty directory page (length of the URL, important terms to look for etc.)

Remaining tasks

- Create a testing / training data set for training our model.
- Finalizing some research papers which would be the inspiration for building our classifier.
- Base our classifier off of what results we see in the training data set.
- Fine tune our classifier to tell with greater confidence whether a website is a faculty directory URL or not.

Issues and challenges faced

- Tons of literature, scholarly articles and papers on the internet - there are tons of papers which seem relevant prima facie but turn out to be not as useful when we go into the details. This has made the task of searching for what we need for sure longer than we had anticipated.
- Too many potential URLs associated with a primary URL. A primary URL like "Illinois.edu" has several valid URLs associated with it so we have to manually stop looking at valid URLs after a certain point.
- Scraping websites with a lot of pages takes time so fine-tuning certain aspects becomes a very time-consuming task.