Building a Search Engine for Academia

CS-328 Project Proposal

Amey Kulkarni	Chris Francis	Nishikant Parmar
18110016	18110041	18110108

March 28, 2021

Problem Description

Build a crowd-sourced search engine that enables users to search and rank professors based on various criteria. For example, it can be used to look for professors in Machine Learning (or from a specific university, research topic etc.) ranked according to their h-index, citations, publications, etc. that can help in applying for higher studies or internships.

Tentatively we plan to use Python and some its modules such as Beautiful Soup, Requests, Pandas, nltk and csv. Other than that, MySQL or file systems for storage, Flask for backend and HTML, CSS, JavaScript for frontend.

We will have the following modules:

1. **Scraping and Converting to Text:** We will scrape Google Scholar Pages and Home Pages (if possible) of professors listed on the CSrankings repository. Information like name, affiliation, email, homepage url, research topics, citations, h-index, i10-index, year-wise citations, and details of top 100 publications will be scraped. We will divide the scraping into multiple parts and run on different IPs with random delays between requests to avoid getting blocked.
2. **Building Search Index:** We will create an Inverted Index for each of the filtering criterion after performing lemmatization and stemming on the scraped text. We will store scraped and index data separately in a file system/database for persistence.
3. **Ranking Results:** Ranking and filtering of search results can be based on any of the criteria available such as h-index, research topics, publications etc. We will give as many options as possible to users, including criteria for the last 5 years.
4. **Creating Backend and UI:** We will deploy the search engine as a public website. Clicking on a search result will display the homepage or a summary of the homepage of the professor.

Datasets

We will start by using the list of professors given at CSrankings and scrape their Google Scholar pages.