Project Proposal - Chicken and Rice

Group Members Information

Name	NetID	Email	Coordinator
Nanyi Yang	nanyiry2	nanyiry2@illinois.edu	✔
Ivan Zhong	ninghan2	ninghan2@illinois.edu	
Cody Wang	yaohuiw2	yaohuiw2@illinois.edu	

Chosen Topic and Why

We have chosen the intelligent browsing theme and specifically will try to create a way to intelligently browse faculty pages in order to reduce the amount of browsing/time needed to find a relevant professor for interesting research topics and Master's programs.

Algorithms & Dataset

For the search algorithm, we plan to use BM25.

For the datasets we will use, the web scraper will act as a way to build our own datasets. Initially, we will try to have functionality that scarpes and searches only the current page, but if there is time, we can extend it to the scraper storign results to a database and the search engine searching that database. This way, users will be able to search various professors/programs at different institutions.

Demonstration

To demonstrate that our program will work, we can use the Chrome extension on faculty websites that have not been tested before in addition to those that we have already tested. This will demonstrate that our scraper is robust enough to deal with various ways of rendering text on websites.

For the search engine, we can also build a simple user interface to allow users to search the documents based on user-specified parameters.

Programming Language

We will be using Python, and probably the beautifulSoup and Selenium for the web scraper

Workload

3 Members - 60 Hours

- Scraper
    - Chrome extension - 10 hours
    - Robust scraper using Selenium - 5-10 hours (considering one of the MPs was something similar, so it may take less time)
    - Potentially a database that stores all the scraped data, which removes the need for repeated scrapping - 10 hours - Potentially deploy database?
- Search Engine
    - 30 hours
- Integration
    - 5-15 hours