CS410 Course Project Proposal

Team Name: SearchExperts (Fall 2021)

1. What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.
   a. Qin Geng (qingeng2@illinois.edu) (Capitain)
   b. Aditya Vyas (vyas7@illinois.edu)
   c. Junwei Chen (junweic4@illinois.edu)
   d. Peter Zhang (aijun2@illinois.edu)

2. What system have you chosen? Which subtopic(s) under the system?
   a. Theme: System Extension
   b. Specific Topic: ExpertSearch System

3. Briefly describe any datasets, algorithms or techniques you plan to use
   a. **datasets:** we have a huge resource of directory page URLs from MP2.1 and faculty webpage URLs from MP2.3. We will also crawl some other pages to get URLs (e.g. other URLs on the university websites, product websites, news sites, etc.)
   b. **algorithms and techniques:** we're going to build a regex-based model and a binary classifier model to help us identify specific webpages. As to extracting other information, we will use Natural Language Processing techniques to recognize names/profiles in web links and extract those. Another idea is using topic mining from MP3 to help extract top keywords in each topic which should be the common search areas.

4. If you are adding a function, how will you demonstrate that it works as expected? If you are improving a function, how will you show your implementation actually works better?
   a. **As to automatically crawling faculty webpages:** we will add the function of automatically crawling random webpages, along with the function (a regex-based function and a binary classifier) of identifying faculty directory links and faculty webpage links. Our expectation is that it can work automatically, and give a satisfying accuracy on identifying, let's set 70%.
   b. **As to extracting other information:** we will add an extracting function based on Natural Language Processing techniques to recognize and extract names/profiles in web links. Also a topic mining function in extracting common search areas. As long as these two functions can work and extract extra information, it's an improvement to the original MP2.

5. How will your code communicate with or utilize the system? It is also fine to build your own systems, just please state your plan clearly.

   Our whole project will be based on MP2.1 and MP2.3, but we're going to develop new functions to improve MP2. Our external functions will eventually be integrated into MP2.

6. Which programming language do you plan to use?

   Python and its packages

7. Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.

   There are mainly five tasks right now, and with the estimated workload (**4 team members in total**) as shown below:

Task No.	Task Description	Estimated Workload hr
Task 1	Writing a function of automatically crawling random web pages, and collecting them for the next step use.	10hr
Task 2	Building a regex-based model to identify faculty directory webpages. Improving this model to achieve an accuracy of 70%.	15hr
Task 3	Building a binary classifier model to identify faculty directory webpages. Improving this model to achieve an accuracy of 70%.	15hr
Task 4	Writing an extracting function based on Natural Language Processing techniques to recognize and extract names/profiles in web links.	15hr
Task 5	Writing a topic mining function in extracting common search areas.	15hr
Task 6	Integrate all functions and models to MP2 and test the whole project.	20hr
Total		**90hr**

We may find more tasks that can be done later as we dive into this project more in the future. We may modify some of the tasks if we find out that they are not working later.