Project Proposal

Project Name: ExpertSearch v2.0 (Auto crawler integration with faculty search improvements)

Abstract:

The UIUC ExpertSearch system has several features such as faculty search, filtering criterions, search results (with options to open faculty bio pages, emailing, location info), pagination etc. As a team, we did a deep analysis of the ExpertSearch capabilities and found several deficiencies that need to be addressed to make it a better search system. The deficiencies include lack of accuracy, lack of relevant search results and inconsistencies in the search results. These deficiencies can be addressed using the text retrieval and text mining techniques that will improve the overall search experience in the ExpertSearch system. The team will be involved in implementing features such as converting unstructured dataset to structured dataset (e.g., csv, json), identifying the key topics (e.g., areas of interest) for each of the faculties and display in the search result, introducing an admin interface that would classify faculty pages and finally improving some of the existing features in the search page for better search experience.

Project Details:

- 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.

Name	NetID	Email	Role
Sudipto Sarkar	sudipto2	sudipto2@illinois.edu	Captain
Ujjal Saha	ujjals2	ujjals2@illinois.edu	Teammate
Arnab KarSarkar	arnabk2	arnabk2@illinois.edu	Teammate

- What system have you chosen? Which subtopic(s) under the system?
  - Theme: System Extension
  - Subtopic: ExpertSearch System

- Briefly describe any datasets, algorithms or techniques you plan to use
  - We are going to use US universities faculty URLs to identify the Bios.
  - We will reference ExpertSearch System code (https://github.com/CS410Assignments/ExpertSearch). We will choose the best fit algorithm; however, we will use the text retriever and text mining algorithms which were taught in the course.

- 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?
  - We will add the admin console, where a user can search the valid URLs to crawl.
  - We are improving the search results (improved relevant searches, consistency in displaying email, location, universities etc.) and we will leverage the same ExpertSearch display user interface to validate the results by manual spot checking and testing.
  - We will test the toping mining function by random manual verification matching with the faculty page.

- 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
    - As we are building on top of an existing system, we will reuse all the major entry exit points of the system and hook up our functions or endpoints.

- Which programming language do you plan to use?
    - We will be using python as our backend language. We will also be using some of the front-end technologies such as jQuery, HTML, JavaScript etc.

- 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.
    - **Epic:** Classifier → **30 hrs.**
        - **User Story:** Crawler Implementation for a given webpage url
        - **User Story:** Adding admin interface for web page indexing
        - **User Story:** Displaying accepted/rejected web page based on url

    - **Epic:** Search Experience Enhancement → **20 hrs.**
        - **User Story:** Build a structured dataset from Unstructured datasets
        - **User Story:** Enhance the search experience with relevant search results
        - **User Story:** Better consistency in displaying links such as email, phone etc. leveraging the structured data

    - **Epic:** Topic Mining → **30 hrs.**
        - **User Story:** Using text mining techniques to extract the Areas of interest for a given faculty based
        - **User Story:** Display Areas of Interest in the faculty search result

    - **Epic:** Documentation → **20 hrs.**
        - **User Story:** Proposal Documentation
        - **User Story:** Progress Report Documentation
        - **User Story:** Final Project Report Documentation
        - **User Story:** Demo Video Presentation and Editing