I (Anna Otte) am working alone for the CS 410 Fall 2023 project, so I will be the "captain" and handle all administrative and technical tasks. My UIUC NetID is akotte2. I have chosen Theme 3: System Extension in the CS 410 Project Topics guide. For my project, I will be augmenting the ExpertSearch System by improving the methodology for converting the unstructured text in faculty webpages into more structured text to enhance the utility of the system. I plan to extract additional fields for each faculty member as well as attempt to improve the extraction currently being done on emails and faculty names since this project's description states that the current extraction techniques - regex-based and Named Entity Recognition (NER) - do not always work well.

I will use the datasets already provided by the ExpertSearch system, mainly the FacultyDataset. I plan to test many algorithms and methods to find one that improves the current system's functionalities. I will first examine if additional data cleaning and language model manipulation can improve the current implementation of ExpertSearch's extraction methodology. Next, there are many existing NER algorithms, so I will test some not currently implemented to see if I can improve the extraction functionality in that manner. Then, I will research alternatives to NER and attempt implementation of those options. Once I find a method that works better than the existing one, I will research additional fields and pieces of information for extraction and implement those extractions using the method I identified as working best. One such piece of information could be keywords from the faculty member's webpages, which could be explored using methods such as BERT embeddings.

I will demonstrate that my implementation of the faculty name and email extraction works better than the existing method by comparing my extraction outputs (e.g., my file with all names and my file with all emails) to the currently existing email and name files. Comparisons on the dataset quality will be made in OpenRefine using related tools, including regular expressions (i.e., checking that each email address is properly formatted). I will similarly demonstrate my extraction of new fields and information by examining dataset quality in OpenRefine, particularly checking the data against a set of pre-defined rules related to that specific extracted field.

My extraction code will be built separately from the current ExpertSearch system, however, I will be using their dataset as described above. Deliverables will include my code, the pre-existing dataset, and my quality and functionality checks from OpenRefine (e.g., OpenRefine history and dataset statistics).

I plan to use Python as my main programming language; I am highly familiar with this language since I use it frequently in both my school and work environments. Python also appears to be the main programming language that ExpertSearch is built in, so using Python for my project make sense in terms of integration.

Since I am a one-person team and will be completing the entire project myself, I am certain that I will readily be able to satisfy the 20*N (N = total number of students) expectation. I expect

that this project will take me greater than 20 hours to finish. The main tasks I plan to complete are as follows:

1. Background research and code review for current ExpertSearch implementation (4 hours)
2. Improve extraction methodology for faculty names and emails (10 hours)
    a. Includes iterative evaluation of current extraction methodology and new extraction methodologies I derive.
3. Research what additional information (e.g., primary faculty research interests, recent publication topics, etc.) about faculty would be useful to students (1-2 hours)
4. Add additional fields for extraction from unstructured faculty webpages to structured text (4-5 hours)
5. Write project report and create demo/presentation (3-4 hours)