## CS 410 Project Progress Report

* **Modification of project scope:**

In the original project proposal, we planned to extract and show both research interests and top 2/3 keywords in publications. After careful consideration, we think the second part will take tremendous time to download and parse publications from sources (like Arxiv or Google Scholar). Therefore, we decided to hold off the publication keywords and first focus on research interests in this project. After we finish the visualization of research interests, we may either come back to the publication keywords, or switch to another field that is contained in the faculty bio data.

* **Progress report:**

**The following tasks in the proposal have been finished:**

1. Digest the existing data + code + visualization and test how they work and connect (10h)
2. Analyze the dataset for the pattern of the research of interest and publication keywords, and come up with corresponding algorithms (10h)
3. Implement functions to retrieve those fields (15~20h)
4. Train the new text retrieval/mining system and make some improvements (10h)

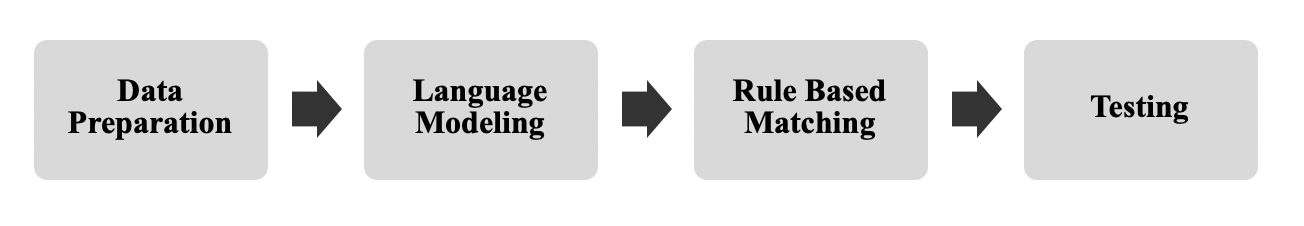
**The following tasks are pending:**

1. Adjust the web frontend to display the new fields (5~10h)
2. Test the new function as well as the whole system (5~10h)
3. Prepare the complete documentation/report (5~10h)

* **Challenges:**

1. The biggest challenge we faced was to systematically download faculty publications using faculty name/university as the query. After spending some time engineering on this, we decided to hold this off and may either come back to it later or switch to another field that is contained in the faculty bio data.
2. We also had some difficulties at the beginning running and debugging the program on Windows, as some of our teammates use Windows machine. With the support of Docker in Pycharm Professional version, we were finally able to run it on Windows.
3. We are currently struggling with html and Javascript to display the new field “research interests” on the webpage. We think this should be manageable as it is almost the last “technical” part of the project.

* **Below is a brief description of our steps to extract area of interests:**



1. Data Preparation

The dataset we use is the `faculty dataset` collected from CS 410 class. While there are 6525 numbers of documents in the dataset, we randomly sampled 25 documents for training. Then, we tokenized sentences and labeled target tokens into corresponding entities. For example, we labeled ‘database systems’ as a `research area` in the following sentence

“His primary research interests are database systems.”

('''His primary research interests are database systems, object-oriented systems and software engineering.''', { 'entities': [(35, 51, 'AREA'), (53,76, 'AREA'), (81, 101, 'AREA'), ] })

1. Language Modelling

While Spacy provides a pre-trained CNN model, we decided to build a new model from scratch with 70 numbers error-free entities.

We used 0.5 for the drop rate and default optimizer provided in Spacy.

1. Rule based Matching

To maximize the quality of outcome, we also leverage rule-based matcher using regular expressions. Since the entities we are looking for are limited to `research interest`, we first searched a sentence that includes `research` or `interest` in the validation dataset. We observed that the entities we aim to extract have a simple structure. For example, they are “My research interest is …, Research interests are.”

1. Outcome

We obtained qualitative results when performing interpolation in our initial model. We will continue testing our model on the testing dataset for future work.

Example:

Entities [('Distributed computing', 'AREA'), ('Analysis of algorithms', 'AREA'), ('Data structures', 'AREA'), ('Computational geometry', 'AREA'), ('Graph algorithms', 'AREA')] Entities [('computer networking', 'AREA'), ('computer security', 'AREA')] Entities [('CS Education', 'AREA')]