## CS 410 Project Proposal

* **Team members**

Joseph Angulo (jangulo2)

Bruno Seo (sbseo2)

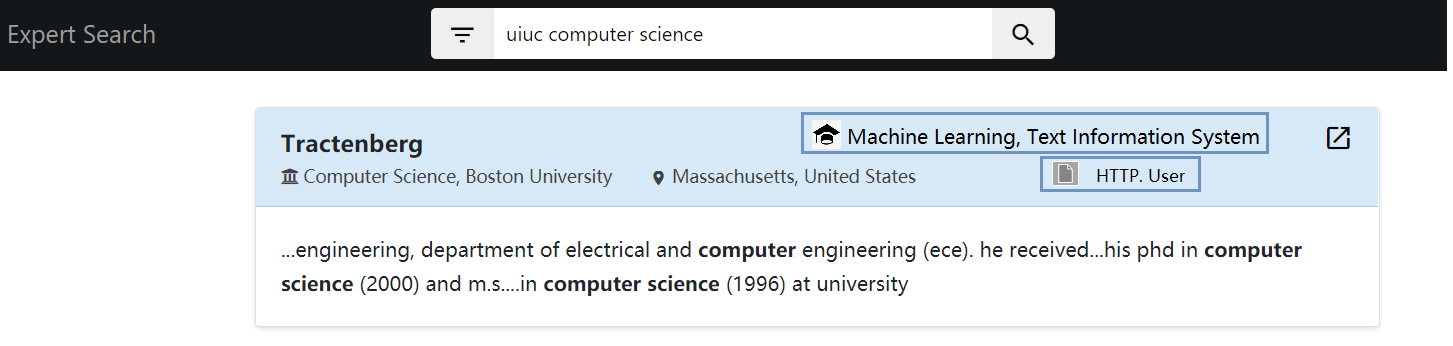
Xiaohan Liu (xliu120) – captain/coordinator

* **Project topic**

Improve the ExpertSearch System: **Extracting relevant information from faculty bios**

* **If you are adding a function, how will you demonstrate that it works as expected?**

In this project, we are going to add two functions to extract ***faculty research interests*** and ***top 2 (or 3) keywords in publications***.In the search result page, the two new fields will be displayed after the name and address items (see highlighted area below).



* **Briefly describe the datasets, algorithms or techniques you plan to use**

We are going to reuse the existing data of the search engine, which is part of the codebase in [GitHub](https://github.com/CS410Fall2020/ExpertSearch). We plan to use Named Entity Recognition (NER) algorithm along with other data mining techniques introduced in this course.

* **How will your code communicate with or utilize the system?**

We are also going to rely the existing code framework: new functions will be added to the ./extraction folder. New results will be merged together with existing search results and displayed on the webpage.

* **Which programming language do you plan to use?**

Python will be the main language for extraction while Javascript and HTML for visualization.

* **Main tasks and estimated time cost (60~80hrs)**

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