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.

Team: JUEE

* 1. Snehangshu Shankar Bhattacharjee (ssb8) – Team Captain
  2. Indranil Guha (iguha4)

1. What system have you chosen? Which subtopic(s) under the system?

Under this project, we have chosen ‘System Extension’ to add new function/improve ExpertSearch functionality

1. Briefly describe any datasets, algorithms or techniques you plan to use

* Datasets: Faculty webpage URL from MP 2.3 dataset for positive examples, for negative example we can use some random URL online (TBD) for negative example.
* Algorithms: Will be exploring various topic modeling algorithm such as PLSA, LDA etc. and see what best fits.
* Techniques: Topic Mining, keyword extraction

1. 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?

* Additional features:

Topic Mining

* Will be adding a function to create a topic model using the bios data extracted from faculty page using automatic crawler. We can also try to add the text content extracted from the faculty research interest page (if available) to add to the collection to improve the topic model outcome. Saving the top keywords per topic as the common research areas. We can maintain a dictionary to map the topics to each faculty and display the top 5 research area in the faculty information on the UI.
* System improvement (optional, if time permits before final submission):
  + - UI improvement
    - Implement keyword extraction features from bios, research papers, etc.

1. 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

We will take the ***ExpertSearch baseline code*** from GitHub and add our own code on top of it for the following features:

(Source code: <https://github.com/CS410Assignments/ExpertSearch>, URL: <http://timan102.cs.illinois.edu/expertsearch//>)

* Adding a script to automatically crawling the directory pages to faculty pages and extract all the bios data.
* Adding the function for topic mining/keyword extraction
* UI improvement (optional)

1. Which programming language do you plan to use?
   * Python
   * JavaScript
2. 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.

|  |  |  |
| --- | --- | --- |
| # | Item / Task description | Efforts (hours) |
| 1 | Understanding ***ExpertSearch baseline code,*** troubleshooting (as needed) | 10+ |
| 2 | Automatic Crawling features for directory pages and associated faculty pages | 20+ |
| 3 | Adding the function for topic mining/keyword extraction | 20+ |
| 4 | UI improvement (optional, if time permits before final submission) | 10+ |
| Total | | 60+ |