1. Our names are Daniel Abdoue and Karthik Talluri, and our NetIDs are dabdoue2 and talluri4, respectively. The captain of the team is Daniel.
2. The system we will be improving is ExpertSearch, and the subtopic is automatically crawling faculty webpages.
3. We plan on accessing all the faculty webpages available by using a strategy similar to that in MP2, except on a much larger scale. We also plan on using various APIs to create the automation necessary for the project idea to be successful. In addition, we will introduce various ranking functions to ensure the project will come out to its best possible form.
4. We will demonstrate that our functionality works as expected by providing test cases of what would be found manually as well as what our code is capable of doing, and comparing the results. Given a page that we have not yet tested, we will get all the necessary URLs using our code, then do the same using a manual method, showing that not only are the results the same, but the time required to get the URLs will be significantly less.
5. The code will take in an input, say, the URL to the desired university to be scraped, and output a file with the faculty directory page, as well as all the faculty webpages that were found. This could also in theory be used to auto-populate the google doc that was used to hold the URLs we had to manually find.
6. We will be using python as our coding language as it provides an easy way to both analyze text, as well as crawl webpages, and automate these processes.
7. The main reason this project will take at least 40 hours is because of the wide variety of webpages our code has to work for. Being able to automatically crawl webpages is a time consuming task on its own, but having to account for all the various ways a website could be created and set up will require a lot of testing as well as different implementations. There are 2 main steps that need to be done it both sections of our task. The first is to try and find URLs or links on a webpage that lead to the page we are looking for, and the next step is to identity whether where we ended up is indeed what we wanted. These two steps have to be done for both faculty directory pages and for faculty webpages. The total time to complete the first step, including all testing to ensure what we are doing works on a variety of webpages, will be around 20 hours, including directory webpages and faculty webpages. Another 20 hours will be for step 2, as this step is quite similar to step 1, however, having solved step 1 does not necessarily make step 2 easier.