Topic Specific Questions:

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.**

   I will be doing this project by myself so I will be the captain. My name is Shyam Shah and my NetID is shyam3.

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

   The system I chose is ExpertSearch and the subtopic I will work on is "Automatically crawling faculty webpages" -> "Identifying webpage faculty URLs"

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

   I will be working primarily with the faculty webpage dataset (from MP2.3) for positive examples. I intend on using the links on the sign-up sheet from MP2.1 (https://docs.google.com/spreadsheets/d/198HqeztqhCHbCbcLeuOmoynnA3Z68cVxixU5vvMuUaM/edit#gid=0) to also scrape some negative examples. My plan is to approach this as a classification problem, as suggested. I intend on using different classification algorithms/techniques (such as logistic regression, k-nearest neighbors, and neural networks) and want to treat part of this as a research opportunity to teach myself some different kinds of classification algorithms that are effective with text.

4. **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?**

   I am going to treat this as adding a function to classify URLs on a given webpage. I can demonstrate that this works by splitting my dataset into a training and validation dataset and measuring precision/recall on the validation dataset.

5. **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**

   I'll start by forking the current project on Github and adding the functionality to it while developing it. There are instructions in the repo to run the application locally so I would not need to build anything else on my own. Ultimately, I would like to have it merged with the current project but that may be something I look into outside of this course.

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

   I will be using Python

7. **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.**

3h	Research algorithms
10h	Implement different algorithms
5h	Testing and tuning parameters
2h	Preparing the tutorial/documentation

Other General Questions:

1. **What is the function of the project?**

   The function is to add the ability for ExpertSearch to automatically identify links within a page as relevant without requiring human intervention.

2. **Who will benefit from this project?**

   The users and developers of ExpertSearch

3. **What existing resources can you use?**

   I can use existing Python libraries like tensorflow and sklearn to implement the models, most of the work will probably be identifying good features to use for them

4. **A very rough timeline to show when you expect to finish what. (The timeline doesn't have to be accurate.)**

November 1	Gather information on all of the models I want to implement
November 22	Finish implementing first version of models
November 29	Finish tuning and testing models
December 7	Finish creating documentation and demonstration