CS410: Text Information Systems
Course Project Proposal

1. What are the names and NetIDs of all your team members? Who is the captain?

Name: Andre Roberts
NetID: andrer2
Email: andrer2@illinois.edu
Team Captain: Andre Roberts

Name: Caleb Thomas
NetID: calebt4
Email: calebt4@illinois.edu

Name: Robin Young
NetID: robiny2
Email: robiny2@illinois.edu

Name: Derek Liu
NetID: derekjl2
Email: derekjl2@illinois.edu

2. What topic have you chosen? Why is it a problem? How does it relate to the theme and to the class?
    a. Intelligent Browsing for Cousera and Campuswire
    b. Information for each class can be spread across lecture videos, text book chapters, and Campuswire conversations which makes it hard to track down when you need it.
    c. It relates to the theme because the end result of the project is a system that will help the user more easily find relevant information according to the user’s query. To achieve the required functionality from the system several techniques learned in this class will be needed such as web scraping, building an inverted index, and retrieving relevant information based on a user’s query.

3. Briefly describe any datasets, algorithms or techniques you plan to use
    a. The two primary data sources for this project will be Coursera and Campuswire, and linked content from these sources such as textbook

chapters. The text data gathered from these two sources will have to be converted into an inverted index.

   b. A corpus will be constructed from the scraped text documents and relevant methods like inverted index and TF-IDF will be used to evaluate queries and return relevant information after the search engine is implemented as a google chrome extension.

4. How will you demonstrate that your approach will work as expected?
   a. The final product will be a search engine extension, and the approach will be evaluated by whether the search engine retrieves relevant information and provides the means to access that information.

5. Which programming language do you plan to use?
   a. Python and Javascript,

6. 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.
   a. Web scraping Campuswire and Coursera
      i. Because Campuswire and Coursera pages are rendered on the client, an understanding of the underlying API or a library like Selenium will be required to obtain the necessary text information.
   b. Making the obtained information searchable
      i. Will need to implement some kind of browser extension to help link questions with content and vice versa