CS410 Project Proposal, CollegeEvents

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.

- Chirag Rastogi: Chiragr2

2. What is your free topic? Please give a detailed description. What is the task? Why is it important or interesting? What is your planned approach? What tools, systems or datasets are involved? What is the expected outcome? How are you going to evaluate your work?

- The topic I chose is to extract events at Universities and parse all information so that it may be stored in a database.
  I will be scraping every social platform (public accounts) for images and text data and I will be extracting events from them.
  The problem can be divided as follows:
    - Unstructured text to structured text: Structured information would enhance the utility of the any system that serves event data. Primary extraction for location and time can be done using NER, however I would perform topic mining for keyword extraction to further enhance the structure.
    - Intelligent Browsing: Index events and creators and allow users to search
    - Context Creation: This is the toughest question when dealing with Events. As text data is extremely unstructured (ex: Discord), sometimes, they may have multiple dates, times, and locations. Deciphering the dates and times that occur in multiple formats is the first of the 2 tasks. This may be easier. Deciphering the Location given the text is much more complicated. For this I propose a context graph that is built using text from a given account/server and the extracted locations from a message leverage the graph to rank accordingly.

  I will be creating my own dataset by scraping social forums, and I will manually annotate a small portion of events. I assume that this would be enough for fine tuning. I will be using models from Hugging Face.
  I will evaluate my results based on the extracted features and search accuracy.

3. Which programming language do you plan to use?

- I plan on using Python

4. 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.

- Dataset extraction and annotation is time consuming. As I will be working on this alone the time taken would be about 5-7 hours per week.