CS 410 Project Proposal

1. General Info:
    a. Names: Heet Parikh, Vrush Patel, Anshul Goswami, Matt Straczek, Adnan Noorullah
    b. Net IDs: hparik5@illinois.edu, vrush2@illinois.edu, anshulg3@illinois.edu, mstrac4@illinois.edu, adnann2@illinois.edu
    c. Captain: Matt Straczek (mstrac4@illinois.edu)
2. Free Topic Details:
    a. Description: Creating a query feature for YouTube where users can search for all occurrences of a word or phase that appeared in a specified Channel (during a specified date range), as well as the location of these phrases. This is a feature that can be extremely useful when attempting to quickly and efficiently find specific information from a Youtube Channel, and has not yet been done before.
    b. Planned approach: Using the HTML data from Youtube, a video's transcript can be scraped and converted into a useful data format for searching through. A bag-of-words representation would then be used for querying the data, and a BM25 ranker would rank the results.The formatting of the results would include the associated video name and timestamp of when the phrase occurred, all of which is included in the HTML data. All of this functionality would be accessible through a React web app interface.
    c. Expected outcome: The expected outcome is an easy-to-use interface that can return its users relevant videos and associated timestamps for their queries on a specific Youtube channel, providing accurate and quick information lookup.
    d. Evaluation: The accuracy of our Youtube query feature would be evaluated manually, using CTRL+F to find all occurrences of words in a video's transcript and cross-reference them with our search engine's results.
3. Programming Languages:
    a. Selenium (python) for web scraping
    b. React for interface, node.js (python) for backend logic, such as web scraping and the search engine (MetaPY)
4. Main tasks and Workload:
    1. 20 hours. Interface for web scraping (channel details, date range for videos, and a query).
    2. 30 hours. Web scraping on a select youtube channel and date range (error checking for videos with missing transcripts).
        - Traverse through each video in date range
        - Retrieve transcript
        - Formatting transcripts to useable documents (with included video name and timestamps for each word/phrase in document)
    3. 30 hours. Search engine, using queries to search through all channel's videos (documents), returns a result that includes all found timestamps and the respective video names, as well as a couple of words that appeared around the found queries.
    4. 20 hours. Display the result in the interface, below the search/config tool.