Requirements

If you choose this theme, please answer the following questions in your proposal:

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.
2. What topic have you chosen? Why is it a problem? How does it relate to the theme and to the class?
3. Briefly describe any datasets, algorithms or techniques you plan to use
4. How will you demonstrate that your approach will work as expected?
5. Which programming language do you plan to use?
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.

At the final stage of your project, you need to deliver the following:

- Your documented source code.
- A demo that shows your implementation actually works. If you are improving a function, compare your results to the previously available function. If your implementation works better, show it off. If not, discuss why.

Proposal Details

1. Names: Dhyey Dixit, Nabil Mohiuddin, Snehal Somalraju
   NetIDs: dhyey2, nmohiu6, snehals2
   Team Captain: Nabil Mohiuddin

2. We have chosen the intelligent browsing topic. The specific topic we have chosen involves creating an extension which optimizes a user's search for videos. This is done through parsing the transcript of the video and comparing it to the user search, leading to a more accurate search result. This is a problem as a user's search may lead to an undesirable video that does not meet the user's needs. This relates to the theme of the course as we are planning on using a web crawler to browse the transcript as a means to assist the user.

3. We plan to use a web crawler to extract the transcripts of each web crawler. We will store them as files and use Okapi BM25 to rank the videos. If needed, we will make adjustments to the algorithm. We will return the top 5 videos.

4. We will use Cranfield Evaluation Methodology to determine which documents should be marked relevant or not. Next, we will check the ranking algorithm against the documents that we determined are accurate. We will provide key metrics such as precision, recall, and mean average precision.

5. We plan on using the python programming language as a means to create this project.

6. The work should take 60 hours total to complete. The main tasks to complete are to build an extension, devise a web crawling algorithm to extract the transcripts, devise a ranking algorithm, displaying the relevant videos, and evaluate the accuracy. The estimated time for each of these tasks is difficult to estimate as each task has a varying difficulty; however, the first 3 tasks should take up the bulk of our time. The top three tasks could be split among our team members, while all of us could work together to complete the final two tasks.