I, Alexander Zurawski (azuraws2), am doing the project solo, and will be the captain.

My chosen topic is Intelligent Learning Platform, specifically Video Segment Search. I have found this to be a problem often myself. I will see a topic in my notes and have to search an entire week’s lecture videos to find the relevant 5-minute segment. The designed system will be a search engine that will find relevant text data based on a query, which is a large topic of discussion in the first half of the course.

The datasets I will be using are the three sets of lecture videos I have access to: CS410: Text Information Systems, CS598: Foundations of Data Curation, and CS: Practical Statistical Learning. I will be using a crawler to find webpages and I will use a ranking function such as BM25 to evaluate results.

I will plan on using free Coursera courses to test the efficacy of the built system. To be more specific, I will make dummy accounts, enroll in 3-4 free courses ranging from very similar, 4 computer science courses, to not similar, 4 courses of 4 different subjects, and run the program. Since I won’t be able to get a full understanding of each free course I enroll in, I will have to make ad hoc judgements. I plan on working in Python.

The tasks of the projects are as follows:

1. Build a crawler that starts at a user’s Coursera home page and can find every web page tied to a lecture
2. Use the crawler to extract the power-point / slides for each lecture
3. Convert the power-point / slide of each lecture into text data
4. Use a ranking algorithm to retrieve the ‘best’ lecture videos

Stretch goals if the above is not 20 hours:

1. Recommend the time stamp that the lecture slide appears
2. Train, test, and tune on larger data sets to see the effects of scalability