CS410 Project Proposal

Team: Mission Lucy

Captain: Sam Song(samsong2)
Azim Keshwani(azimmk2)
Kavithaa Suresh Kumar(ks64)
Uttam Roy(uroy)

For our course project we have chosen to build a search engine that can support video segment search. Going with the theme of Intelligent Learning Platforms, we aim to be able to allow the user to query a classes' video lectures and see a ranked list of clips from video lectures. The problem that we aim to solve is that lecture videos may have multi-topics in it and the same topics might be in multiple lectures. Through video segment search, we are allowing users to more easily access shorter clips of relevant information that may be scattered across multiple lectures

Our project relates largely to the theme of Intelligent Learning Platform by adding more searchability to a learning platform. This project also relates to what we have learned in the first half of the class in regards to search engines, and the text retrieval problem. Although the information returned to the user is in the form of video clips, the same algorithms can largely still be applied.

To create our video segment search engine we plan to have a web crawler scrape Coursera's website for the video clips and video transcripts. These videos and transcripts will be our dataset from which we will provide the video clips. Utilizing the video transcripts we can then rank video clips from most relevant to least relevant, using BM25 as the ranking algorithm. To implement this algorithm we are planning to use Python. We are going to create a web application to demonstrate the video segmentation search where the user can enter the query and the output would be a ranked list of video segments that is relevant to the query. The query results will be evaluated by having team members manually determine relevance. That way we can evaluate the search engine using metrics like precision, recall, and F1.

Tasks:	Hours required
Web crawling to retrieve related information	10
Topic segmentation (identifying transition points)	40
Query matching	10
Evaluation	10