

Topic: Intelligent Learning Platform

- Build a search engine that can support video segment search for Courseera lecture videos. This would allow an user to type in a query and see a ranked list of short video segments so that the user can precisely locate which segment to watch in a lecture in order to know more about a concept. We intend to build the search engine based on the scrapping/indexing/ranking concepts learnt from this course. The end solution will provide an User Interface for the user to type in the query and to view the results as links to the video segments

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.

Matthew Kryczka - kryczka3

Selvaganapathy Thirugnanam - st26

Diana Arita - dianama2

Kartik Patel – kartikp2

Captain – Kartik Patel

2. What topic have you chosen? Why is it a problem? How does it relate to the theme and to the class?
 - Intelligent Learning Platform- Build a search engine that can support video segment search. This would allow you to type in a query and see a ranked list of short video segments so that you can precisely locate which segment to watch in a lecture in order to know more about a concept.
 - Apart from searching videos by titles, there is no easy way to search by video content. This makes harder for users who are interested in finding precise locations of the videos where a concept or a term is mentioned.

- Assuming, Each Video segment as a document, we will try to address the "Pull" mode of the text retrieval problem by ranking the these documents based on the transcripts available for each video segment and return the segments that best matches the query.

3. Briefly describe any datasets, algorithms or techniques you plan to use

- We plan to use course lectures videos (i.e videos of CS 410: Text Information Systems) as the dataset. At a very high level, we tend to perform the following :
 - Scrape the videos along with the transcripts
 - Extract documents out of scrapped videos and transcripts and build an association between them
 - Use Tokenizer(Stemming and other normalization techniques) to extract lexical units (words)
 - Use an Indexer (inverted Index) for faster response from the Search Engine.
 - Perform Ranking/Scoring based on Probabilistic retrieval functions (Eg: BM25)
 - So UI will be built for the user to submit the query. Once submitted, the query is further normalized and sent to ranking function to score against the data available in the index. The top ranked results will then be displayed in the UI. The results would provide links to the video segments.

4. How will you demonstrate that your approach will work as expected? Which programming language do you plan to use?

- Demo video for the demonstration. We will try to search for concept as keywords used in our lectures and the results should contain video segments related to the concepts.
- scrapping, indexing, ranking ---> Python
- UI ---> Javascript frameworks (either Angular or React)
- Backend ---> Python/Flask

5. Please justify that the workload of your topic is at least $20 \times N$ 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.

- Group of 4, 80 hours of total work

Scraping the video text	10 hours
Indexing	20 hours
Ranking	20 hours
UI	10 hours
Mapping the video	20 hours

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.