CS410 Project Proposal
October 24, 2021

Team: Topic Miner
Member/Captain: Kristina Hill (kdh5)

What topic have you chosen? Why is it a problem? How does it relate to the theme and to the class?

- Enhance the algorithms used in Smartmoocs: the project is hosted at smartmoocs.web.illinois.edu.
- As a learner I would like to be able to revisit a lecture and jump directly to the portion of the video that addresses a certain topic (for easy revising before exams, or revisiting topics that are more complicated).
- Explore better ways to segment lectures based on topic transition. At present, the lecture is segmented into uniform length 1 minute segments. A better way is to detect points of "topic change" in the lecture transcript and segment the lectures such that each segment discusses a different topic.

Briefly describe any datasets, algorithms or techniques you plan to use. What is your planned approach? What tools, systems or datasets are involved?

- Explore the Gensim toolkit in Python to develop topic clusters and segmentation points (timestamps) within each lecture on a MOOC
- Utilize the videos and transcripts from the UIUC Text Analytics Coursera course

Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team.

- [1h] Dataset acquisition and enablement
- [3h] Explore topic of video segments in the current minute-long video segments
- [2h] Examine paragraph breakpoints in video transcripts - explore topic within segments
- [6h] Implement Gensim method of developing topic clusters
- [5h] Translate results of Genesis topic segmentation into video timestamps and incorporate empirical feedback to compare timestamps with current minute-long segments.
- [2h] Compare user-friendliness of timestamps created with all three methods and provide empirical feedback.
- Possible exploration of additional algorithms for topic mining including the NLTK toolkit - compare algorithm performance using empirical feedback.