Project Proposal

Group Members:
Grace Chang (ycchang4) - Captain
Bohan Liu (bohan3)
Yipeng Yang (yipengy2)
Leo Yang (junjiey3)

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

Chosen Topic: Concept View.
Given a MOOC lecture, preferably its text transcript. We want to identify the topics covered in the lecture by generating words and phrases to represent the topics. Meanwhile, our desired output would focus on the content of the lecture and exclude related, but uncovered concepts. The difficulty in this problem is to clearly identify all the concepts in a lecture and exclude the uncovered topics. The choice and representation of topics is an interesting problem to solve. This project has close relation to the theme of the class. It's within the scope of text analysis.

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

We plan to use the PLSA/LDA algorithm to extract major topics from the transcript and select words with the largest probabilities in each topic to represent the concepts. We also plan to devise an algorithm to distinguish between concepts taught in this class and those mentioned as prerequisites. It may be implemented by retrieving some background knowledge model from the Internet and then comparing the concepts extracted from the video with the background model. We can add the taught concepts in each video to a set/tree-like data structure so that for future videos in the same course we can easily identify the concepts that have already been taught. Another idea is that we can extract certain features (e.g. patterns like "why do we use <concept>") to to perform a binary classification between newly taught concepts and prerequisites.
We will run our algorithm on the transcripts of the two moocs of CS410 and will use the given human annotated dataset for evaluation.

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

We will calculate the precision and recall of the top K results of our algorithm compared with the given dataset, assuming that the human annotations are reliable. We can compare the precision/recall of different techniques against the baseline, which is simply extracting the top

concepts using LDA. Also, we can go through a specific video to demonstrate whether the results actually make sense from a student's perspective. We plan to use Python for this project.

4. 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.

Main Task & Estimated Time Cost:
Implement LDA and fine tune parameters (5-10h)
Find a background language model (5-10h)
Devise algorithm to distinguish between newly taught concepts and prerequisites (25-30h)
Devise algorithm to store previous concepts (10-15h)
Fetch all transcripts and evaluate using given dataset (5-10h)
Improve algorithm (20-30h)