1.Progress Made

Implemented PLSA, and have found the general list of topics for the lectures.

1. LDA
   LDA is under development. Inspection of the output of PLSA reveals problems discussed in later sections.
2. Data Collection
   As we tried to scrape text data as a background model for LDA, we find text8 dataset is a better background.
   We are currently using the CS410 text data as the training dataset for our model.
3. .vtt data to text data conversion
   We use python scripts to read .vtt data and produce formatted text data.
4. Model for evaluation
   We have surveyed several ways to evaluate the result. The list of topics for CS410 is useful.

2. Remaining Tasks

1. Develop a model to predict the number of topics in a lecture.
2. Develop ways to interpret LDA word distributions.
3. Develop models to separate prerequisite and actual topic.
4. Write a python library that is easy to use
5. Comment the code.
6. Write the documentation
7. Resolve the challenges.

3.Challenges and Issues

1. Intuitively interpret the word distributions as topics.
2. Make good use of the timestamps in a .vtt file.
3. Learn complex concepts in lectures, e.g. phrases and sentences.
4. Look up new ways to interpret text data.