1.Progress Made

1. Data Collection

We wrote a Python script to convert the given .vtt text data to a collection of documents where the transcript of each lecture is represented by a document.

2. PLSA

We tried applying the PLSA algorithm to the collection of transcripts. We selected the top m topics for each transcript and chose the top n words to represent each topic. However, the results were not very good because it's difficult to identify concepts from the topic word distributions. This might also be due to the fact that the top topics for each document did not capture the concepts very well, since concepts like "text analytics" may occur only a few times in a transcript. We later implement LDA to see if a background model will make any difference.

3. RAKE-NLTK

We also tried using the RAKE-NLTK Python library to extract the keyphrases in each lecture transcript. Some results were good compared with the given human annotated dataset ("unigram language model" appears in both our result and the human labeling for lecture 20 of text analytics course):

```
lecture 20
['might bias towards using one topic', 'continue discussing probabilistic topic models', 'mission would allow us', 'two
unigram language models', 'two unigram language models', 'maximum likelihood estimator later', 'background word distribu
tion denoted', 'unigram language models', 'unigram language models', 'component like theta sub']
```

	Lecture 19	multiplier approach	
67	Lecture 20	Probalistic Topic Models: Mixture of Unigram Language Models; probabilistic topic models; mixture model; unigram; language model; mixture of two unigram language models; component model; filtering out background; words; generative model	Looks good wi by Reviewer#2

However, it seems that besides noun phrases, RAKE-NLTK tends to capture subject-verb-object phrases like "might bias towards using one

topic” in the first image. This leads to the lack of relevant concepts in the top keyphrases returned by the algorithm.

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. How to interpret the topic word distributions in PLSA/LDA as concepts.
2. How to make good use of the timestamps in a .vtt file. Also, maybe explore the structural information in a lecture transcript (for example, words like “first”, “also”, “next” may suggest a switch in topic).