1) Which tasks have been completed?
We finished programming Latent Dirichlet Allocation for topic extraction. We preprocessed the text from a test dataset (transcripts of one of the lecture videos) to remove stopwords, perform tokenization, lematizing, and stemming words, and developed a bag of words model that we performed LDA on and found keywords that would allow the user to interpret the topic. However, the results of the LDA just gave us a probabilistic bag of words. For example, for Lecture 1.4, Overview of Text Retrieval Methods, we got the following keywords:

Words: 0.042*"model" + 0.041*"document" + 0.024*"function" + 0.021*"queri" + 0.018*"word" + 0.015*"retriev" + 0.015*"rank" + 0.015*"lectur" + 0.014*"differ" + 0.013*"score"

2) Which tasks are pending?
Currently, the output of LDA does not help us to uncover key points of topic change in the lecture. We are planning on attending office hours to ask about how we can use the results of LDA to determine the placement of a topic in a lecture, or if there might be a better approach in order to determine this. We are currently looking into utilizing cosine similarity; if we know that a lecture covers X topics, we can divide the lecture into X segments based on the differences between text blocks. Then, we can run LDA on the separate segments for a clearer representation for each segment.

3) Are you facing any challenges?
Our biggest challenge right now is finding a good approach to segment the lectures based on topic transitions. However, since we have some ideas in mind, we can solve this by attending office hours and asking a TA for clarity.