Requirements

If you choose this theme, please answer the following questions in your proposal:

1. **What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.**
   a. Team member 1: Urvi Awasthi
      i. Netid: urvia2
   b. Team member 2: Anupam Ojha
      i. Netid: anupamo2
2. **What topic have you chosen? Why is it a problem? How does it relate to the theme and to the class?**
   a. We have chosen to improve on an intelligent learning platform for our course project, specifically Smartmoocs. The avenue of improvement that we have chosen for this project is to explore how to better segment lectures based on topic transitions. We plan on segmenting the lecture based on topic transitions that we will manually detect. The intent of this project is to allow the user to directly jump to the topic at hand in the video without coursing through the rest of the lecture. This relates to the overarching theme, intelligent learning platforms, because we are addressing a main area of improvement for the existing platform, Smartmoocs, by leveraging what we have learned in this class, such as topic mining and analysis.
3. **Briefly describe any datasets, algorithms or techniques you plan to use**
   a. We plan on using the existing CS 410 lecture transcripts as the datasets that we will be performing topic mining on. In order to do this, we will ask our TAs how we can access the code for the existing platform, Smartmoocs, which we expect already has these datasets available to it. Once we have gotten any APIs / open source code exposing the current topic mining done for the CS 410 lectures on Smartmoocs, we plan to use Latent Dirichlet Allocation for topic extraction. Specifically, we will be preprocessing the raw text to remove stopwords, perform tokenization, lematizing, and stemming words. We plan to use existing libraries such as NLTK and gensim libraries for this preprocessing. Then, we can develop a bag of words model that we can utilize the Latent Dirichlet Allocation algorithm on, to finally extract the most relevant topics for the lecture.
4. **How will you demonstrate that your approach will work as expected? Which programming language do you plan to use**
   a. We will be building our application on top of the existing Smartmoocs platform, so demonstrating that our application can segment videos based on topic detection will be enough to show that our approach works as expected.
   b. We plan to use Python for the duration of this course project.
5. **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.**
   a. Environment setup: retrieve code for the existing platform, Smartmoocs. Setting up a github repository for the final project. (10 hours)
   b. Programming the algorithm for topic detection (10 hours)
   c. Segmenting video based on topics and time stamps detected (5 hours)
   d. Updating UI to show segmented time stamps (5 hours)
   e. Testing (10 hours)
   f. Creating the project demo (3 hours)