Project Proposal

Team NLP

1. Name and NetID: Mingwei Huang (mingwei6)

2. Topic: Intelligent Learning Platform, the Smartmoocs problem, where I want to improve the topic labeling of summarizing the transcripts. This course also introduces text clustering and text categorization which is related to this problem.

3. Data, Algorithms/Techniques: The datasets are those transcripts and manually summarized related concepts of each lecture. I prefer some end-to-end model but it should be almost impossible to build one since I do not have much labeled data. So, instead, I will try some pipeline approach: segregate the texts first and then summarize them and it is at least doable in an unsupervised way.

4. See if the performance makes sense and is useful. The result is a little bit hard to be evaluated since there is no such correct answer nor a range of acceptable answers. But it is still possible to use the manually summarized related concepts of each lecture to get some rough precision and recall in this case, but the majority of evaluation would be done manually. And I will firstly start with a baseline approach using bag-of-words/n-grams to see if the result is at least working or not, and then develop/use some models to compare with the baseline performance.

5. I think the data collection & preprocessing, implementation of the baseline approach along with the evaluation will take around 8-10 hours. The remaining time will be spent to develop/explore models that can achieve some SOTA result, or at least obviously improve the performance.