Progress Report

Team: Free Topics Team
Team Leader: Dajun Lin (dajunl2@illinois.edu)
Teammates: Tsz Yui So (tyso2@illinois.edu), Amy Zhao (yiminz3@illinois.edu)

Github repo: https://github.com/alexlin1822/410CourseProject.git

Progress Made Thus Far

- We met over Zoom to discuss progress and plan for next steps (3 hours)
- We have gathered course transcripts from Coursera (3 hours)
- We did research on libraries and tools (15 hours)
- We wrote a program to parse the transcripts (7 hours)
- We used gensim, spacy, and nltk to find topics in the documents (15 hours)

Remaining tasks

- We need to build a web app using streamlit, reformat the transcripts, and refactor the code to use multiple documents to form a corpus (12 hours)
- Use pyLDAvis to visualize topics (3 hours)
- We need to perform testing (6 hours)
- Write documentation, prepare powerpoint slides, and record presentation (6 hours)

Challenges and Issues

We have faced some challenges and issues while working on this project. First, we have to collect the data from Coursera. Also, we have to clean the data and filter out the stop words using libraries. We have also faced some dependency issues while working on this project. Also, we spent a lot of time figuring out which libraries we should use to do topic modeling. We also spent time reviewing the LDA algorithm.