410 Project Proposal

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

What is your free topic? Please give a detailed description. What is the task?

Our topic is to analyze Coursera transcripts to extract topics to extract a list of main topics included in the video. As we know, each course in coursera has a lot of knowledge points. Our app can compute the weight of various important chapter topics in a course, so as to determine some important knowledge points. These analysts' results could help users focus on the key knowledge points of the course.
In addition, our app can also analyze different courses at the same time to reveal whether there are similar knowledge points between different courses, for example, whether there are overlapping knowledge points in applied machine learning and data mining courses. This can help users use their own knowledge to better understand and learn the new knowledge.

Why is it important or interesting? What is your planned approach?

Coursera is getting more and more popular.We plan to apply topic modeling to Coursera transcripts. By this way, our app can help users analyze the main knowledge points of the course. That means users not only can more clearly understand the important theme of the course content, but also capture the similar knowledge points in the different courses that the user has learned. Users can use this analysis result to get a clearer concept of the knowledge focus and structure of the subject they are studying. In addition, if our model or app can be extended to the official Coursera website, the official website can display the summary of the knowledge points of this course in each course, as well as the correlation with the knowledge points of the industry. So that users can have an intuitive understanding of whether this course is valuable to them before deciding to study or not. On the other hand, the administrators of Coursera can easily get the knowledge correlation between different courses, which is of great help to the management and analysis. Therefore, our project is not only very interesting, but also very important.

What tools, systems or datasets are involved?

Tools: Python
System/Platform: streamlit
Datasets: At least 10 course transcripts in coursera.

What is the expected outcome? How are you going to evaluate your work? Which programming language do you plan to use?

The algorithm should be able to find the topics discussed in the Coursera videos. We will watch the Coursera videos and label the topics and then compare them with the outcome of our algorithm. We are going to use Python.

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.

Timeline:

- Finding topics and defining the problem - 10 hours
    - Gathering datasets - 3 hours
    - Research on libraries and tools - 7 hours
- Project implementation - 40 hours
- Testing - 6 hours
- Documentation and presentation - 6 hours