LiveDataLab System Extension - Progress Report

1. Progress Made: After looking more into our project and concept, we decided to implement our system extension in a slightly different order than initially proposed. We began by going to Coursera and obtaining all the lecture transcripts for the weeks relevant to this course's MPs. From there, we copied them all into separate .txt files in VS Code and began writing our functions to parse the transcripts. Before ranking them, the lecture transcripts had to be parsed to remove timings and stop words. Each of these tasks was done by a separate function and the output of each function was stored in a different .txt file. Transcripts are in the “Parsed\_Lectures” folder after the timings are removed and in the “SWRemoved\_Lectures” after the stop words are removed. The initial set of stopwords was obtained from nltk.corpus, however, we extended the set of stop words to be more specific to this class and the words we wanted to remove. After going through this process with the lectures, we began a similar process with the MP instructions by placing the READMEs of each MP into separate .txt files. We once again began determining which words to add to our set of stop words, however, due to the nature of the instructions and their inclusion of code, we must figure out a different way to remove stop words that we have not implemented yet. To summarize, thus far we have implemented all lecture parsing required prior to implementing the ranker function and are over halfway through doing the same for the MP instructions.
2. Remaining Tasks: Our next task is creating the BM25 ranker and running it for each MP description using the parsed lecture files. Once we have ranked which lectures are most relevant for each MP, we will organize them into the top three most relevant lectures for each MP. After that, if we have time, we will work on creating the UI display. If possible, we would like to add this extension to the actual site, but for now, we are planning on creating a mock display of LiveDataLab. This mock display will include individual tiles for each MP, and when clicked, will have hyperlinks to the top three most relevant lectures. When the MP’s tile is clicked, we will have it access the document that stores the three most relevant lectures to create that display. We will test this display with different search queries and documents to see if the UI ranks and displays the correct results.
3. Any Challenges/Issues: At this point in our project we encountered different challenges. Our first problem came up when we were processing our data. Since a part of our project was finding keywords in the lecture notes, we had to turn all of the lectures into text files. When we transferred the notes into files, we found that there were time stamps for the dialogue throughout the file. In order to fix this we wrote code that parsed the files and created new files that did not have the time stamps. The next challenge we faced was deleting stop words such as “the” from the files. We used a library that was supposed to have all of the stop words in the English language but we found that after we ran our code, there were still many common words left. We fixed this by creating a new set of stop words that were missing from the original set and adding it to the set. Our code used this set to filter out these words. We also added words that were common in CS 410 that would not count as keywords in the context of this course, such as “Coursera” and “Due”. Our current challenge is that when implementing the ranker function, we don’t know how to actually test the function. In previous MPs that required us to implement/use ranking functions, the actual execution of the function was done through the main function which we did not write. As a result, we are struggling to interpret how to test the BM25 ranking function on our text files. A future issue we will face is figuring out how to access/connect to LiveDataLab. More specifically, once we have the lectures and MPs sorted and ranked, how do we get it displayed on LiveDataLab or make it an extension of the site?