CS410 Project Progress Report – Fall 2023

Ehsan Sarfaraz

11/18/2023

1. *Which tasks have been completed?*

    - Research to find a way to download cs 410 transcript as a dataset.
    - Use several of my research papers as a dataset.
    - Convert docx to txt format and clean data for analysis.
    - Research and learn to find libraries to use LLM.
    - Write a script code to use course transcript and technical paper to communicate with LLM to extract topics and description.
    - Used panda data frame to save data in csv format for next analysis.

2. *Which tasks are pending?*

    - Next steps will be to use extracted topics to find keywords.
    - Convert csv to text file.
    - Remove stop words from data.
    - Use the maximum likelihood formula to find word distribution in the document.
    - High probability words will be considered as key words.
    - Evaluate the result with my expectation keyword that I provided to the publisher.

3. *Are you facing any challenges?*

    I enjoyed working on this project because I wanted to use the skills that I learned from this course to use it in different fields. As a researcher it is important to find the right keyword as a query to find relevant research papers within your search criteria. Since I worked on this project independently, I spent time learning LLM, clean data and how to download transcripts from coursera. Extracting a topic from a text document with LLM also was challenging that I faced so far.