# CS410 Project Proposal – Fall 2023

## Topic mining with LLM

1. *What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.*

   My name is Ehsan Sarfaraz and I decided to complete this project by myself. My NetID is ehsans3 (ehsans3@illinois.edu).

2. *What is your free topic? Please give a detailed description. What is the task? Why is it important or interesting? What is your planned approach? What tools, systems or datasets are involved? What is the expected outcome? How are you going to evaluate your work?*

   My free topic is about topic modeling by using a large language model. LDA and PLSA have been used as a technique to model the topics and information under a probabilistic framework. I would like to use LLM to find topics, information related to topic and keywords. This topic caught my attention because my experience and research are in mechanical engineering and as a researcher it is important to extract top details from technical documents. Furthermore, topic modeling can be used to find keywords from lectures to assist students to study. I would like to use the skills that I gained from this course to find topics and top information from research papers, conference presentations, lecture courses, and textbooks.
   For this project, I will use Text Information Systems course lecture transcripts as a dataset and use LLM to find some relevant topics and short information about each lecture. Furthermore, I will use NLP libraries such as NLTK to find keywords from the dataset that I generated from LLM. I would like to see what words from this course are repeated most. For example, are text, mining, topic, feedback, language, likelihood, statistical, probability, IDF, and evaluation are top words that have been discussed in this course? In order to evaluate this approach, I will use my journal paper to realize can system predict the keywords that I provided to publisher.

3. *Which programming language do you plan to use?*

   I will use the Python language for this project. I will use several libraries such as langchain, panda, ChatOpenAI, NLTK, etc.

4. *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.*

   - Research, learn LLM, find relevant libraries, and establish a dataset: 6-7 hours.
   - Generate topics with descriptions from the dataset using LLM: 10-12 hours.
   - Find keyword: 3-4 hours.
   - Documentation and presentation: 3-4 hours.