Tasks completed so far

Keyword Extraction Implementation:

- Developed a Python script for extracting keywords from course descriptions.
- Implemented two methods for keyword extraction:
    - Traditional NLP techniques using NLTK and Scikit-learn for tokenization, stop words removal, stemming, and TF-IDF based keyword extraction.
    - Advanced NLP using spaCy for extracting key phrases and entities from the course description and title, considering the existence of corresponding Wikipedia pages.
- Integrated text preprocessing steps including tokenization, stop word removal, and stemming.
- Established a connection with the Wikipedia API to validate the existence of Wikipedia pages corresponding to the extracted keywords.

Chrome Extension:

- Started a basic framework for the Chrome extension, including manifest.json, contents.json, popup.html, and popup.js
- designed the extension to successfully read the required content from the Course Explorere webpage and filters the useful information about each course
- Display the extracted content for testing purpose, making sure we only get the useful information.

Remaining tasks

- Optimize keyword extraction code to generate more related words
- Connect Chrome extension to Python script to extract directly from the website

Challenges/Issues being facing

- Keyword extraction problems:
    - Some extracted keywords may related to the course description, but they are too general, such as “Programming”, “Topics”, “Coding”, “Class”. We want to find a way to filter these words out