Project Proposal for CS 410: Text Information Systems

Group Name: GOAT
Hao Huang, haoh11 (Project Coordinator)
Kaiyuan Wang, kw22

Project Title: Intelligent Course Explorer Extension for Enhanced Academic Understanding

1. What topic have you chosen? Why is it a problem? How does it relate to the theme and to the class?

- A Chrome extension that suggests three relevant Wikipedia pages based on a given course description. Our Chrome extension will operate on the course explorer website, where each university course is displayed with its title and description. Upon activation, the extension will analyze this information and present users with top 3 related Wikipedia links.
- Many students, when exploring potential courses to take, may find it challenging to gain a comprehensive understanding of the course content based solely on the title and brief description provided. Our goal is to bridge this gap and offer students deeper insights into potential courses by connecting them with related topics from a trusted source, Wikipedia.
- The foundation of our project is deeply rooted in the information retrieval techniques we've acquired in the Text Information Systems class. By leveraging methods like BM25, we aim to parse, analyze, and draw relevant connections between course content and Wikipedia topics, epitomizing the practical application of our classroom learnings.

2. Briefly describe any datasets, algorithms or techniques you plan to us

- **Dataset**: While Wikipedia is our primary source, we would not necessarily need a specific dataset. Instead, we'd be querying Wikipedia's API based on the course description.
- **Algorithms/Techniques:**
  - NLP:
    - Tokenization, Stop Words Removal, Stemming & Lemmatization
  - Keyword Extraction:
    - TF-IDF (Term Frequency-Inverse Document Frequency), TextRank

3. How will you demonstrate that your approach will work as expected?

- Conduct user tests where participants provide a course description, use the extension, and then provide feedback on the relevance of the suggested Wikipedia articles.

- Use a set of predefined course descriptions and manually curated relevant Wikipedia links as a benchmark to assess the accuracy of the extension.

4. Which programming language do you plan to use?

- We will use JavaScript for building the chrome extension. Then use python for the NLP and Keyword Extraction since python includes many useful packages.

5. 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.

- Our project requires at least 40 hours (two teammembers) of work. Our task includes building up and setting the chrome extension, implement NLP for word search and comparing, incorporating NLP portion (python) with the extension (JavaScript)

6. Are there potential challenges or limitations?

One potential challenge could be ensuring the relevance of Wikipedia links to the course content. We may also face limitations with the BM25 scoring if it doesn't capture the essence of the course content accurately.