CS 410 Project Proposal

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.**

   Team GYZ:
   Yuhan Guo (yuhang4) --- captain
   Qi Zeng (qizeng2)
   Haoduo Yan (haoduoy2)

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

   Theme: Intelligent Browsing
   Topic: Chrome extension for entity-based knowledge extension

   Motivation:
   During the webpage reading process, out-of-(user)-vocabulary words are easily encountered even for a well-educated and erudite reader. Inferring the word meaning from the context is not always reliable, especially for some less skilled readers. Jumping out of the current page to conduct searching may disrupt the reading pace and decrease the reading efficiency. And most importantly, readers may be misled by the wrong impression of the particular novel words that they assume they know but actually do not.

   Therefore, we propose to build a Chrome extension that enables knowledge extension for reading. This tool is supposed to help identify the named entity in the current webpage and then provide related links and abstract as reference.

   It is closely related to Theme 1 because it is an intelligent system for less educated people, including those young or eager to learn. It is closely related to the class because it involves the basic text processing techniques and searching techniques.

3. **Briefly describe any datasets, algorithms or techniques you plan to use**

   Datasets: webpages (primarily academic articles)

   Techniques: NER, Key Sentence Identification

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

   We will demonstrate our functionalities on a specific webpage to prove that our approach works as expected.

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

   JavaScript and Python.

6. **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.**

   Chrome extension implementation: 10h
   User interface design: 5h
   Web scraping: 5h
   Data cleaning: 5h
   Named-entity recognition/Key Sentence Identification: 15h
   Searching based on the recognition: 10h
   Filtering based on relevance: 10h