CS410 Final Project Progress Report
Annotation Extension for Educational Websites

Group name: Night Crawler
Group members: Jiawei Yuan (jiaweiy3), Ziqi Xu (ziqixu3), Ziyao Zhang (zhang416)

Which tasks have been completed?

According to our schedule, we have implemented a crawler that scraps data from the Campuswire discussion forum to obtain data we need to process. The main information we attempted to grab from each post includes its post id, its title and a main description of the question. We have also done part of the preprocessing work, including tokenization, removal of stop words, and a basic implementation of an inverted index that will make the information retrieval later more efficient. In addition, we have generated a basic outline of our chrome extension and found the asset needed to build the user interface.

Which tasks are pending?

1) Some tweaks on the crawler part that can make the scrapping of information more accurate. We might consider if we want to crawl the answers and discussion threads under each questions as well.

2) Improvement on the data processing part. The discussion posts on the forum not only contain text data, but also contain other forms of data such as images and urls, which make the data cleaning and processing more challenging. Currently our data processing pipeline and the storage of data in the inverted index do not handle those kinds of complicated situations so it seem the processing result contain a lot of unnecessary information. After this we will also need to decide a proper scoring function and finish the remaining part of the classification work.

3) Build the extension & UI part. We need to implement the Chrome extension and connect it with our backend for information retrieval.

4) Testing phrase that allows us to use the feedbacks to evaluate and improve model performance.

Are you facing any challenges?

As mentioned in the previous section, for the crawler implementation, we found that the data of the posts on Campuswire was not so easy to be acquired accurately, and page layout made the crawling job more complicated. When processing the data, there are many encoded images and URLs which makes the data size too huge to handle. Therefore we had to filter these types of data and extract the core meaning of them instead to improve the flexibility of our retrieval model. Moreover, we have limited experience with JS and building extension so the frontend may take more time and effort.