Project Progress Report - Developer Team

Team caption: Allan Huang (Shengqi5)

Team member: Zengjie Tang (zengjie3), Danmeng Zheng(danmeng2), Chuching Ho (cch11)

1) Which tasks have been completed?

We have completed data scraping, preprocessing, and storage in the crawler. More specifically, the crawler scrapes data (posts) from the CampusWire CS410 channel, and extracts important information like id, title, description, etc. Then we do the preprocessing of data, including tokenization and stop word removal. We implement an inverted index to support efficient search and retrieval.

Also, we built a web application using Flask to support the initialization of the crawler, including data scraping, preprocessing, and storage.

2) Which tasks are pending?

1. Scraping and storage improvement. We need to utilize more information in the post to support further text retrieval, such as the count of “like” received by the post.

2. Query processing. We need to utilize text retrieval techniques to retrieve the top 5 posts according to the keyword given by the query.

3. UI development of the Chrome extension. We need to implement it to support more efficient interaction with the backend.

3) Are you facing any challenges?

1. Data scraping. We are not very familiar with data scraping. It takes us some time to learn Selenium and use it to interact with web elements, simulate user actions, and extract data from web pages.
2. Web application development. There are several popular web application frameworks in Python, like Django, Flask, etc, but they have different philosophies, levels of abstraction, and use cases. We developed our web application using Flask because Flask is designed to be lightweight and follows a minimalistic philosophy. It provides only the essential components needed for web development.