410 Team Progress Report

Team Willis Tower
Zehao Miao(Captain, zehaom2),
Zuhua Cao (zuhuac2),
Fan Wu (fanw8)

In the proposal, we divided our work into three categories, and the corresponding progress are as follows:

Frontend part:

Progress made: Learned how Chrome extension works and finished the setup of the front-end code. Finished the UI design, including the keyword selection component and Add Page button. The frontend project code has been uploaded to the repo under /frontend

Remaining tasks: Integrate the web crawling code to the frontend and link the algorithm. Attach the UI to the corresponding logic. Setup chrome storage methods.

Challenges being faced: Possible UI freeze due to the long processing time of the algorithm. A possible optimization is needed.

Algorithm part:

Progress made: We've implemented a classic BM25 algorithm in JavaScript. In this model, it will have several documents and a query as the input. Then the model will help us to score each document and create an ID for them.
Remaining tasks: The remaining task should be changing it to multiple queries so that it could classify every document based on their scores. Also, we need to connect our algorithm with the frontend and utilize the dataset as our input.
Challenges being faced: The main challenge that we met was writing BM25 in JavaScript since we are not that familiar with this language.

Dataset Testing part:

Compared to the proposal, for the dataset building part, we made some changes. We decided to pick documents from the WIKI pages since it's easy to process. We choose 7 subjects as our base tag to select documents. All document URLs and tags are saved in /dataset/tag_index as reference.
The document text is saved in dataset/doc.txt using beautiful soup. Code is saved in /dataset/test.ipynb
Remaining tasks: Evaluation for algorithm accuracy, accuracy testing.
Challenges being faced: Not sure about how to construct the evaluation part, try to find some template but don't have a clear clue yet.