410 Team Project Proposal

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

The theme we chose is Intelligent Browsing. We plan to write a Chrome extension that supports the user to classify their default bookmark list. For some users, they might have a large number of sites bookmarked but don't want to label it every time. This extension is designed to solve such problems. It will be able to sort the bookmark list with intelligence. The user will input a label list, and the extension will sort bookmarks automatically based on the given label list. Such problems are related to document classification and text mining, which is the main topic in this class.

As for the dataset, we didn't find a useful one to work on, so we plan to make a sample dataset by ourselves. We will pick 7 keywords and choose the top 10-15 results from the Google search using those keywords, then form a set. We assume only those top results will be considered relevant. Our test will be based on this dataset to check the performance of our algorithm, and test the functionalities. The main algorithm we utilize is BM25. The program we designed will use each element in the label list as a query, and then the program will call BM25 to rank the documents. It will pick the top 10 results to form a folder with the corresponding keyword.

For the evaluation part, we will manually label a dataset of common web pages, and perform the classification on the pages using our extension. We will then compare the results given by our extension and their correct labels and compute the accuracy of our algorithm. The programming language we plan to use should be HTML, CSS, JavaScript (React Frontend).

Last, we will mainly separate our workload into three parts: frontend, algorithms and testing.

[Total 20 hrs] The front end part includes two tasks: web crawling and user interface design. For the web crawling of documents, we need to find a proper strategy to crawl general web pages, and then store the crawled pages in the browser's local storage. This design will take about 8 hours, considering the fact that we spent 4 hours on MP2.1 to crawl a specific type of page, and the crawling of general pages will be more complex. For the user interface design, this will be a normal web application design, which will take about 12 hours considering the UI logic implementation and styling.

[Total 21 hrs] The algorithms part has 4 tasks: algorithm choosing, coding, debugging and connecting. The first one for the algorithm should be choosing a proper algorithm for bookmarks classification, which will take about 4 hours to search and understand the current techniques for text classification. Then we need to spend about 8 hours writing the code part for this algorithm, then we should spend 4 hours debugging. After that, it will take about 5 hours to connect our algorithms with our frontend part. It will take 21 hours to complete by adding them up.

[Total 20 hrs] The dataset/testing part includes two tasks: dataset forming and accuracy testing. The dataset forming task requires us to look for accessible web pages and label them with our testing classification keywords. Given the dataset size mentioned before, this task will take about 15 hours, considering the time to look for useful web pages and read/label them. The second task requires accuracy checking and fine-tuning the parameters of our algorithm, which will take about 5 hours.