Project Proposal

The names and NetIDs of our team are Phyllis Wang (pjwang3), Eshia Rustagi (eshiafr2), and Felicia Wang (felicia5). Our captain is Felicia Wang.

We chose Theme 1 and our chosen topic is an extension that allows users to index the current page and utilize a text retrieval function to search for a phrase, and have relevant results and their area on the page displayed to them. This extension would be an alternative to using Ctrl-F on a page, as that function is limited by only returning exact matches to the query, and only displaying results by order of appearance. These limitations mean that a user would have to remember the exact phrasing of a section of text on the page in order to effectively find it; otherwise, a vague/general query may return way too many results to effectively look through

This extension relates to the theme of 'Intelligent Browsing' in that it gives more functionality to browsing and gives users a better experience and more utility when using the web, through features that more intelligently take information given by the users to return results that are more likely to be what they are looking for. To achieve this, we plan to use indexing and text retrieval, which directly relate to their respective topics taught in the class, as well as utilizing the functions covered during lectures.

The datasets we will be using are primarily Wikipedia articles and PDF documents, and the user can enable the extension once they are on these pages. The data will be collected when the user arrives on the page. In our implementation, we plan to use BM25 as our main algorithm to rank the documents, as it is one of the best search algorithms and commonly used by search engines. We also have experience using this algorithm in our homework assignments, which is why we chose this algorithm. In our project, each 'document' will be a paragraph on the page, so that we can return the most relevant paragraphs to the user for their search query. We will also explore using TF-IDF and bag-of-words techniques in order to gauge the relevancy of these documents. And finally, if we have time, we may attempt using word2vec for semantic analysis, so that we can return documents that are similar to the search query, and not just exact matches.

In order to demonstrate that our approach will work as expected, we plan to create a mock web page with a known set of both synonymous and unrelated terms such that when we search for a word or phrase, we can confirm the accuracy of both the results that are provided as well as those that are not provided. In addition, we may also test our extension on a couple short Wikipedia articles and PDF documents to ensure that it also works as intended on text we cannot control. We intend to show these tests in our final demonstration video.

We plan to use Javascript for the functionality of the extension, such as processing user query input and returning results to the user. We would also use HTML/CSS for styling of the extension. For backend, such as retrieving text data, indexing, and the text retrieval algorithm, we plan to use Python.

An outline of the main tasks and an approximation of how much time each will take:

1. Understand the development guides and learn to develop a basic extension (12 hrs)
2. Build basic extension, which will include a popup with search bar and table with results (6 hrs)
3. Index page functionality (upon entering Wikipedia page/clicking on extension) (10 hrs)
4. Access and collect data for current text document (specific Wikipedia page) (3 hrs)
5. Implement search algorithm, display results (BM25) (18 hrs)
6. Experiment to find optimal algorithm parameters (3 hrs)
7. Create test pages to prove correctness of approach (9 hrs)
8. If we have extra time, we would attempt the following tasks to improve our project:
    a) Learn how to use the word2vec algorithms and incorporate it into our ranking algorithm (5 hrs)
    b) Implement semantic analysis using word2vec, such as searching for similar words in query (12 hrs)

The tasks we expect to complete by the project deadline (steps 1-7) will have a total workload of **61 hours** for the 3 of us. If we have extra time to implement our stretch tasks, then our total workload would be 78 hours for the 3 of us.