Team captain and lone team member: ajfitch3
Team Name: Andrew

Project Proposal

I chose Theme 1: intelligent browsing. For my project, I will design and implement a chrome extension that enables search within a webpage using the Okapi-BM25 ranking function. This is a problem because current search functionality within a webpage on google chrome is limited to exact key word matching. There are some cases where a user may want to do a more approximate search within a document using a complex text retrieval algorithm. This relates to the theme of the class because it involves mining data from a webpage, indexing textual data, and then running it through a text retrieval ranking function.

To implement this project, I plan on creating a chrome extension where the client-side front end is in HTML, CSS, and JavaScript and the backend is a google cloud function implemented in python. The front-end will be responsible for parsing the webpage and extracting the text as strings (I will probably treat each sentence as an individual document), and then sending the document to the google cloud function. Then, google cloud function will be responsible for indexing the documents (sentences) and ranking the documents (using BM25). The client will communicate with the backend via HTTP protocols. The Google Cloud Function will rely on metapy to implement the indexing and the search algorithm.

To demonstrate that this approach works as expected I should be able to use my extension while viewing a webpage, enter my search, and then the highest-ranked sentences that best match my search should be displayed to me in order. To evaluate this approach more thoroughly, I could choose a document and manually rank sentences for a given query and then see how the search function performs against my manual ranking. I also should evaluate my

chrome extension on several different types of webpages to ensure that it can parse any webpage for text data effectively.

Overall, I expect this project to take me at least 20 hours. Setting up the chrome extension on the front end and providing a nice UI will take at least 5 hours (I've made a chrome extension before so some of this will be spent doing research). Adding logic to effectively mine text data from webpages and split the text data into documents delineated by sentences will likely take another 5 hours. Setting up the google cloud function, with indexing and ranking will take me around 10 hours (I've never used GCP before so I imagine there will be a bit of a learning curve). If I have extra time, I can expand this project by adding more search options (like allowing users to modify a document from a sentence to a paragraph or allowing users to customize the retrieval function they use). However, I expect the initial chrome extension will take more than enough time to fill the allotted 20 hours.