Bryson Li (brysonl2)
CS 410 Project Proposal: Intelligent Browsing
Fall 2022

This project that I am proposing is to create an extension to index the current page and allow users to search over the page using a common retrieval function (ie. BM25), possibly also extracting some of the topics from the page. This project is related to text retrieval (scraping from web pages, BM25) and text mining (possibly extracting topics or something in that area like sentiment analysis). I am planning on doing this as a solo project, so much of the time will be going through tutorials on how to make extensions and working on some smaller versions of the project to test functionality. I know we do have a similar extension for cs 410 that does this, but I am planning on my extension being a simpler version that at least does something extra (ie. possibly extracting topics from the page or something else).

To be able to accomplish this I am planning on using a simple python-based server, that will be started locally, and the extension will directly contact that server (most likely using javascript).

The datasets that I will use are whatever pages that the user chooses. The programming languages that I will use are python and javascript (and possibly more as needed, as I investigate these further). The algorithms I will use are BM25, but I may choose a different one as I am developing it (for now I will stay with BM25)

Here are some tasks that I will need to complete in order to finish the project (Exact details for each of these tasks may vary as I look into them)

1. [3-5 hours] simple proof of concept - python program with python server - scrape from selected webpage(s) and send to local python server
2. [1-2 hour] look into building extensions
3. [5 hours] simple extension (javascript) - take current webpage send to server (local server for simplicity of debugging)
4. [3 hours] add search functionality (if the extension part was done well then this will mostly be for using BM25/search algorithm and formatting the results page, etc.)
5. [5 hours] add additional functionality [ie. topic extraction or sentiment analysis after the user has chosen/ saved multiple articles]

To demonstrate that my approach will work as expected, I will need to be able to start up the python server locally, and configure the extension to point to that server. and then I will need to show the user saving a few articles, and then I should be able to search for articles (on a different page) and get results that use the search algorithm. Also there should be some additional information based on the additional functionality that I choose (for example: topic extraction).