CS410 Course Project Proposal: Improved Keyword Search

Team Name: Blue Team

1) What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.

   This will be an individual project by Kevin Cen, netID: kcen2.

2) What topic have you chosen? Why is it a problem? How does it relate to the theme and to the class?

   I have chosen Intelligent Browsing as my theme. My topic within Intelligent Browsing will be to try and improve keyword search on a specific page. This is a problem because keyword search on a webpage currently only looks for exact matches, and thus often a user will miss similar keywords or themes. Or perhaps the user does not know exactly what keyword he/she is looking for on a webpage and can only enter a related keyword and hope to find relevant matches.

   Thus, because current search capabilities are limited to exact keyword match, I hope to expand upon that capability by allowing users to search over the page using a common retrieval function such as BM25. This relates to the theme and class because it makes browsing more intelligent for the user and implements text/information retrieval techniques learned from this class.

3) Briefly describe any datasets, algorithms or techniques you plan to use

   I plan to use the BM25 retrieval function to try and expand upon the capabilities of keyword search. If time permits, I hope to use other retrieval functions that we have learned in this course.

4) How will you demonstrate that your approach will work as expected?

   I hope to try and demonstrate the approach works by using user feedback - specifically with me as the user making judgements on whether the search results are relevant to the query/keyword. I will test over a number of different documents/websites/topics and then judge the percentage of relevant returned results. Obviously this will take time and is not efficient over a larger scale but for the purposes of this project I believe it will be the best way to judge whether the extension is helpful or not.

5) Which programming language do you plan to use?

I plan to implement this project using Javascript or Python.

6) Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.

   I believe this course project will take me at least 20 hours, which is the expected workload. I did not have much experience coding in Javascript/Pythong prior to this class and have not worked with browser extensions before so I believe learning how to work with those extensions will take a good amount of time (10+ hours). Implementation of the BM25 algorithm will likely take another 6-8 hours and then testing and ensuring the approach works as expected will take another 4-5 hours. This adds up to at least 20 hours, and I hope to try and implement another retrieval function with which to compare BM25 if possible, which will take another 6-8 hours.