CS 410 Project Proposal

Intelligent Web Page Information Retrieval and Link Browsing (Intelligent Browsing Track)

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.

   Shi Yao Liu (shiyao3) (Captain)
   Xinrui Zhu (xinruiz4)
   Ryan Cedzo (cedzo2)

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

   Topic: Create a Chrome extension that indexes the current page, scrape the page for hyperlinks and recommend sections and links from the page that may be relevant to the query.

   Problem: Current search capabilities are only limited to exact keywords. In addition, only relevant text results from the current page are returned. By retrieving relevant links from the page in addition to text from BM25, we can provide higher-quality information to the user as well as a starting point for web browsing.

   Relevance: Concepts directly from the course will be applied to a real-life problem. BM25 will be used for ranking, limited web scraping and crawling will be used to extract link information. Methods of augmenting the BM25 algorithm discussed in lectures will also be explored and applied if applicable.

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

   Dataset: Development, testing, and evaluation will be done on a set of Wikipedia pages. The reason is Wikipedia contains a rich amount of text information, a wide knowledge base, and various links to other sites. This provides us with a large amount of high-quality data, making it an ideal proving ground.

   Algorithm: BM25 algorithm. Augmentation techniques and variations introduced in lectures may also be applied if it improves performance.

   Techniques: Segregation page contents, use of hyperlinks, using link descriptions, web scraping, exploring variations of the BM25 algorithm.

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

   Decide on a few Wikipedia pages with a high volume of text data and links to external sources (for example, the UIUC Wikipedia page). As discussed in lectures, empirical

evaluation is always required to some degree. There are three people in the group, each individual selects a page and ranks the links and sections on the page based on relevance to a sample query. The algorithm is then run on the three pages with the same queries. The ranking from the algorithm can then be compared to the ranking we provided.

5. Which programming language do you plan to use?

   HTML, CSS, JavaScript

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.

   This project involves several major components which take considerable amounts of time to be able to obtain accurate results. They are broken down as follows:

   Total 60+ hours

   Dataset: Dataset selection and manual processing. This step involves a lot of manual work. We need to come up with queries and rank page sections and links for each query and split pages logically based on what we expect the user to be interested in. (~10 hours)

   Implementation: UI/UX (interface design, use cases, result presentation), ranking algorithm, web scraping, exploring variations of the BM25 algorithm. This is where the majority of the workload is, the actual implementation of the project. (40+ hours)

   Evaluation: Performance metrics, comparisons, calculating scores. Are the scores reasonable? Do the results make sense? Is it what the user would expect? (10+ hours).