CS 410 Text Information Systems Final Project Proposal

Team: Ctrl Freaks

Fall 2023

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.**

   Team Members:
   - Blake McBride (blakepm2) - Captain
   - Kaushal Dadi (kadadi2)
   - Rohan Parekh (rohanjp2)
   - Megha Chada (megharc2)
   - Michael Ma (chiuyin2)

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

   We have chosen Intelligent Browsing as our theme for the project. We intend to build a Knowledge Graph Builder Google Chrome extension which will allow users to keep a record of websites they have visited in a similarity graph such that similar webpages (documents) will be clustered together.

   This project directly addresses the hassle of managing disparate information across multiple tabs when performing online research; collecting and visualizing information based on similarity will eliminate the need to keep many tabs open at once while online researching or using third party tools to track one's references. Technically speaking, our application will allow users to find similarities between web pages and create a similarity graph of webpages visited. This will allow the user to not only have a better experience recovering information from their search history, but also identify similarities measured between them, and filter by topic/content.

   This project relates directly to content we have learned in the course regarding information retrieval and text analysis techniques. We will cover this aspect in more detail in the following sections, but our approach will incorporate critical features of

information retrieval such as TF-IDF weighting, ranking algorithms, and measuring similarity between documents.

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

   The dataset would be the web pages browsed by the user, which we web scrap and will be our documents. Some algorithms we plan to implement are TF/IDF to find the most common keywords in the articles and link articles with these keywords. We will also use a ranker algorithm to show the most likely document(s) it is about. We will also store this information in a custom database which we could eventually use to measure the performance of our tool and optimize it in the future.

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

   The final video submission for the course project will give us an opportunity to give a recorded demonstration of the Chrome extension in action. We will do a screen recording where we go over the code at a high level, the documentation we've created, and demonstrate the Chrome extension's functionality and outputs/visualization.

5. **Which programming language do you plan to use?**

   We will use a combination of several programming languages to fulfill the requirements of this project. First, Python will be used for building out the backend of our application given its speed and general applicability for data processing/analysis tasks. We will also leverage JavaScript as a means of building out the front end of our project as a given considering we intend to deliver a Chrome extension.

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.**

   There are 5 members in our group, which means the workload should be at least 100 hours.

   - 5 hours - Learn to how make a Google Chrome extension

- 5 hours - UI Enhancements
- 5 hours - Learn how to webscrape
- 5 hours - Web scrape web pages user has traveled with extension on
- 10 hours - Make a similarity algorithm (TF/IDF + Ranking)
- 10 hours - Visualization of the graph
- 20 hours - Frontend Code
- 20 hours - Backend Code
- 20 hours - Create and link Database

Note: These are estimates and it's very possible that some of these tasks may fall within some range of the numbers we've provided–some taking longer than expected while others may take shorter than expected. But this general breakdown should suffice in justifying the workload of this project for our group.