CS 410: Text Information Systems
Course Project

This document is to help explain what has been submitted.

For the course project, we built a Chrome extension specifically for browsing through Campuswire and Coursera posts and pages. Broadly speaking, our codebase consists of three parts: the scraper, the inverted index, and the browser extension. The function of the scraper is to collect text data from Campuswire and Coursera. It outputs the data as a JSON file. The inverted index code takes the scraped text and builds the inverted index which will store the mappings between words and their locations in documents. Using this data structure allows for efficient querying. The browser extension serves as the user-facing end of the application and provides an interface for going through the index. This is what the user ultimately uses to search Campuswire and Coursera for relevant pages according to their query.

The scraper was implemented in Python using the Requests and Selenium libraries. The inverted index was also implemented in Python and it used the Metapy library with Okapi BM25 as the ranking function. The browser extension was implemented in Javascript. Additionally, AWS Lambda was used to run the index builder code inside a Docker container in order to generate the database that users can query.

There are additional README files under the extension, inverted_index, and scraper directories that provide instructions on how to use the code.

Link to the video tutorial is here:
https://drive.google.com/file/d/10W5ZKdIX7_l3IhiENnu01BdIlbzaGYQR/view?usp=share_link

Team contributions:
Andre Roberts: Team captain, browser extension
Caleb Thomas: Scraper
Derek Liu: Documentation, tutorial presentation
Robin Young: Inverted index