**CS410 Text Information Systems (Fall 2021)**

**Project Proposal:**

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.

* Andrew Son / andrew28 (Captain)
* Sujay Nanjannavar / sujaypn2

1. 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 topic. The issue we would like to combat is the poor results of most search engines outside of Google. Some examples of this include Campuswire, LinkedIn, and Reddit. We plan to build upon LinkedIn job postings by ranking search results based on a query rather than keyword search. This relates to the theme by extending the capabilities of LinkedIn’s keyword search for job postings to allow for better results.

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

We plan to use a basic web crawling technique to collect all outgoing application links on LinkedIn. Then, job specifications such as location, salary, and basic qualifications will be scraped from the original application posting. Once the data has been processed and formatted into a bag of words representation, a BM25 ranking algorithm will be used to rank the 25 documents based on relevance to a search query.

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

In order to evaluate our program, we have collected a sample corpus of 25 documents where each document is a description of a job posting. This collection will be ranked by hand by both members based on three different queries. Then, the collection will be ranked by our ranker with identical queries. The outputs of the ranker will be compared to our ranking to evaluate its effectiveness and accuracy

1. Which programming language do you plan to use?

Python will be the primary language used for this project. It will be used for crawling, scraping, formatting, and ranking of the documents.

1. 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.

Become acclimated with BeautifulSoup, Selenium, and Metapy: 5 hours

Create program to crawl and scraper LinkedIn job postings: 10 hours

Develop script to parse and format raw text: 2 hours

Implement BM25 ranking algorithm for formatted text: 10 hours

Implement a front end to display results of ranking: 10 hours

Evaluation of program (more detail in section 3): 5 hours