Course Project Progress Report

CS410 | November 13th, 2021 | Tony Mu, Yang Pan, and Wenjun Peng

1 Completed Tasks

The following tasks have been completed:

- Scrapers for all nine job sites, including Facebook, Amazon, Apple, Netflix, Google, Microsoft, Uber, Lyft, and Airbnb. Each scraper will output two line-separated files: a file that includes job URLs, and a file includes job descriptions. Each scraper takes about two minutes to run and records about 200 job listings.

2 Pending Tasks

Since the text data gathering part has been completed, much of the work left is around the indexing/ranking of the text data. We are considering two approaches: in the first approach, we will implement our own indexing/ranking functions using MeTA and implement our own lightweight Python server to server queries; in the second approach, we are considering using a full text search engine and service such as Solr. The following tasks are pending:

- Tokenization for the job descriptions
- Building an index for the job descriptions
- Build and tune a ranker with good performance for the job descriptions
- Build a server and a web page that will return ranked job listings based on relevance to the input query
- Potentially explore opportunities in utilizing Solr to replace the search and serve part of the stack

3 Challenges

The challenges we have faced so far were all related to the scrapers. We had an issue with the Chrome Webdriver, where the page was not fully loaded when the text data was downloaded. Another issue was with the website's anti-scraping rule. We had to manually throttle our QPS to the page, otherwise we would get incomplete data.