Course Project Progress Report

Shai Yusov (syusov2@illinois.edu)

System: EducationalWeb System
Subtopic: Identifying in-demand skills

1. Progress Made Thus Far

My project is about automatically identifying in-demand skills. The original plan was to scrape and analyze job postings for software engineers in New York, NY, from the job board Indeed and then extract the top skills from that dataset.

So far, I have written, tested, and executed a component to scrape job postings for software engineers in New York, NY, from Indeed, to clean the scraped the data, and to save all processed job postings to a file; I have thoroughly investigated various algorithms and approaches to extracting the top keywords, and thus most in-demand skills, from the dataset; I have written and tested a component to ingest the dataset and extract the top skills using a keyword extraction algorithm; and I have compared and instrumented various approaches to extracting the top skills to achieve the best results.

I have taken care to isolate the interfaces of these components from the implementation details as much as possible so that the functionality will be generic and could potentially integrate with other tools and systems in a straightforward manner.

2. Remaining Tasks

The primary tasks of this project are complete.

3. Challenges and Issues

When scraping job postings from Indeed, the primary issue I have faced is how to successfully extract all the needed pages in a timely manner and without being throttled. Specifically, my use case involves scanning job results pages, then obtaining job post links from that page, and so on. If executed in serial, this process would be rather slow, so I parallelized this IO-intensive process using a pool of threads that fetch and extract information from pages in parallel. This led to a significant speedup. I also had to account for various failures so I implemented retry-able mechanisms to ensure graceful error recovery. Overall, I was able to download all the data.

Another issue I am facing is that, so far, the keyword extractions algorithms with the highest subjective quality of results are also the ones taking the longest. Specifically, it seems that an implementation of TextRank produces fairly relevant results but takes a little longer to finish, while other algorithm implementations of RAKE, YAKE, and TF-IDF produce somewhat more diluted results but in a shorter amount of time.