Author: Kadin

Task: Scraping Job Information from JobsInMaine

Script: PhaseTwo/Scraping/jobsinmainescraper.py

Data: PhaseTwo/Data/maine_jobs.csv

Methodology:

To do this project I started out writing code that extracted all the information from the page that I needed to fulfill the attributes in the Job and Employer entities.

I inspected the page to find what tags to pull from and then made a loop for each href that referenced a job application.

Partway through I started to use Sonnet 4.5 to streamline the process and it generated the code that I needed to use to extract information.

At the very end, I found that extraction was taking far too long because I was visiting a new webpage for every job being extracted so I pasted in my code to Sonnet 4.5 and asked it to make it quicker. It polished it into it's final form where it runs operations concurrently and keeps a cache of employer webpages that I visit multiple times.

To collect the data, I looked through the jobsinmaine page to se when the job postings went to. Around page 35 is where they were pushing 3 months old so that is where I set the scraper to scrape to, to avoid unnecessary old job postings.

In total, ~1000 jobs were scraped.