Project Idea Description

**Abstract**

Entering the job market can be very overwhelming. Digital job boards appear to offer an abundance of opportunities, but most lack an enforced standardized structure for their postings. Even with the available filtering capabilities, students looking for their first jobs are often presented with many jobs they are unqualified or unsuited for alongside real possibilities. Manually parsing through such a crowded digital marketplace to try to understand, evaluate, and prepare materials for each job separately, can feel very tedious and inefficient.

Our project is a tool that will help students position themselves in the job market. Our goal is to process job postings, to then categorize and summarize different segments of the job market. We want to better identify and display information given in each job posting, so that students who use our platform can filter for jobs that meet particular criteria, or get generalized insights about a subset of jobs to see how to become a better candidate. Another possible functionality to pursue is automated generation of cover letters that are tailored to each job.

**Data Backend**

Our job postings are collected from APIs provided by The Muse and Adzuna.

We scraped The Muse’s API documentation page ([collected here](https://colab.research.google.com/drive/1E90-d0EFe6VUDCy1-Uc5CvL9bVo0HaUc?usp=sharing), [stored here](https://drive.google.com/drive/folders/1nR_eUsXeQSBVm5lMa--VZWkrotp-6o0s?usp=share_link)), because it includes a comprehensive list of the exact values available for some of their attributes. We use some of this information now, and may try to incorporate more of it later.

We retrieve the job postings from the APIs and store them into one CSV ([collected here](https://colab.research.google.com/drive/1rUwCaekTr7JhBM_a9gJ4sru0WAL5ifGP?usp=sharing), [stored here](https://drive.google.com/file/d/14I3svPtRJUo5tc6GThOfS18T6wx4x2Yb/view?usp=share_link)), which we read from later to perform analytics and visualization. We will eventually transfer to SQL instead. It’s important to have regularly updatable and accessible data, as the API requests get the most relevant data for a particular point in time.

We have chosen to start by only collecting jobs in the US, given that Adzuna’s API request requires a country. To do this, our API request to The Muse needed an explicit list of US cities, which we processed from the documentation page as described above. One request with all the cities didn’t work, so we ended up doing 4 rounds of retrieval from The Muse, each time collecting n/4 postings from ¼ of the cities in their database. The same process for another country might be more complicated because of language differences, but we may expand globally, in which case we would allocate our Adzuna requests by country. We collect 1000 jobs from each API, not handling the possibility of running out yet.

We mostly store the data from the 2 APIs as they are, figuring out which attributes from one API matched to which from the other. For The Muse, we removed HTML tags from the job description text, and cleaned it further by rearranging around characters and punctuation to have organised data ([here](https://drive.google.com/file/d/1fN4iJ1TiAI_Z8trzvGDdrMlm24czamq3/view?usp=share_link)) that can be used as input for NLP and further sort functions. We alsp used another API (mapbox) to add coordinate features (some of them are wrong, and we need to fix that), from the location recorded in the posting. We still want to clean the other features to be more standardized and categorical, instead of text-based. We also want to add features based on the job description, through a separate process of updating the dataframe ([updated here](https://colab.research.google.com/drive/1bp5dibf_TF1JppYipjFOViYC3tOOIAgh?usp=sharing)).

For the potential cover letter generation, we have collected, as examples, 2 cover letters from resumegenius ([collected here](https://colab.research.google.com/drive/1SCFUkQ6R26AZWY3wt_2SEoQMCycZYMWG?usp=share_link)), some adjectives commonly used during a job application process from resumecompanion and their synonyms from wordhippo ([collected here](https://colab.research.google.com/drive/17s5MvE0Wu2lHC-V4o9YmwV6XUGO-OLef?usp=sharing), [stored here](https://drive.google.com/drive/folders/1v_o3wE-krj49hFY2v-2OWlWK-FZgFgg2?usp=share_link)). For a potential feature where we display insights for each job segment, we may consider expanding our use of Adzuna API’s top companies endpoint ([collected here](https://colab.research.google.com/drive/1e4eHLKewBPx3nfAHrR1ImWtw2olXcRtR?usp=sharing), [stored here](https://drive.google.com/file/d/1MuobQsorjurD-Uac5L__AzbUiImLULcy/view?usp=sharing)).

**Analytics Approach**

We read in the job posting data as described above, and analyze and visualize the data ([done here](https://colab.research.google.com/drive/1syq9HdHCd6bGryiqwgYvtAEZw6vkw7tV?usp=sharing)). We plot based on latitude and longitude (but we don’t yet handle locations where a longitude/latitude isn’t given except storing their info in a list), color based on company (since we noticed some companies appear overwhelmingly), put all the features in a popup, and do nothing else yet. Also there’s some histograms of longitude and latitude distribution, along with a histogram for dates that jobs were posted.

We also filtered the jobs by company and state, which later will be supplemented by enabling queries such as which state or company has the most job offerings for a certain type of job, essentially answering the question of which state or company can my set of skills offer the most utility. ([code found here](https://colab.research.google.com/drive/1OsoW6sbSy73wcv3yMqvOySNSAnt9vnmL?usp=sharing))