Job matcher V1
– A comprehensive way of search for a job
July 12th 2018
Yaqiong Zhang
Abstract

Introduction

Job search website like Linkedin, Indeed and glassdoor are great resources for people to look for a job. A lot of people start their job search with those websites by entering their keywords and the location they want to be. However, job search with those websites can be time-consuming when we are not very sure about the keywords and the location. When I put some non-specific keywords, and search national-wide in Linkedin, it returns me a long list of jobs. Let along I have several ideas about the keywords I want to use, they represent different possible jobs that may match my educational background most, or match my current interest most or sometimes, I just want to explore the possibilities. On the other hand, I don't want to spend too much time in reading those lists of jobs. After all, I've heard that I should spend more time talking with people than browsing those websites. To solve this problem, I developed this job matcher prototype. It returns me with the suggested jobs I may interested in based on the questionnaire I filled in.

The questionnaire is expendable. But for this prototype, I put those questions in it:

1. What's the location you currently in?
   A: 84108
2. How do you feel about re-location, from hate, not very hate, Ok, preferred?
   A: Ok
3. Do you want to live in a metropolitan region, from 1(means no)-5(means must)?
   A: 4
4. What kind of population change from -2 (means big decrease), -1, 0, 1 2 (means big increase) is acceptable?
   A: 0, 1
5. Give the keywords you think your future company's culture should be?
   A: 'innovative', 'open', 'art', 'ai'
6. Give the keywords you think your future company's industry should be?
   A: 'healthcare', 'programming', 'research', 'analysis'

After finishing this questionnaire, we also give a weight to the question 2 to the question 6. After those step, the job matcher should give you a suggested job for you from the database. After you read the suggested job, you can give feedback to the job matcher. You can say, it's not the type of job I want, or you can say give me more like this! The job matcher should adjust their calculation if necessary and give you another suggested job.

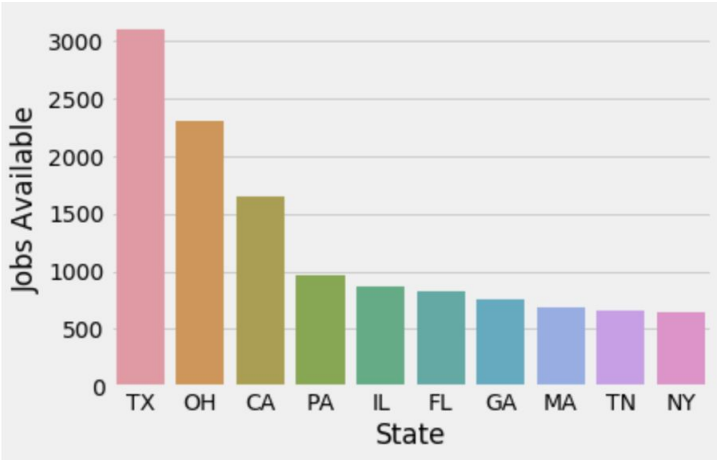
Exploratory data analysis

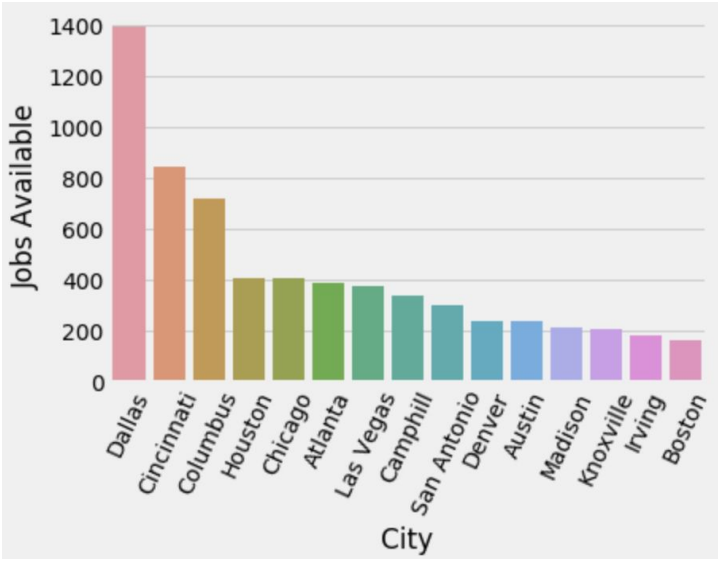
Because extensive web scraping is strongly prohibited in those major job posting website. I found an alternative dataset from Kaggle.com with 22000 jobs. Data source: kaggle dataset. Note that this dataset contains web scraping data from Monster.com, so data is not cleaned and does

not have labels on it. And this dataset contains a wide variety of work such as IT engineer , clerks and room attendant. So we can give different tastes of recommendations.

The first thing I do to clean the dataset is to drop some un-relevant columns for this project, such as url, salary and id. Then I used regular expression and other things to extract the zip code, the state and the city information from the location column.

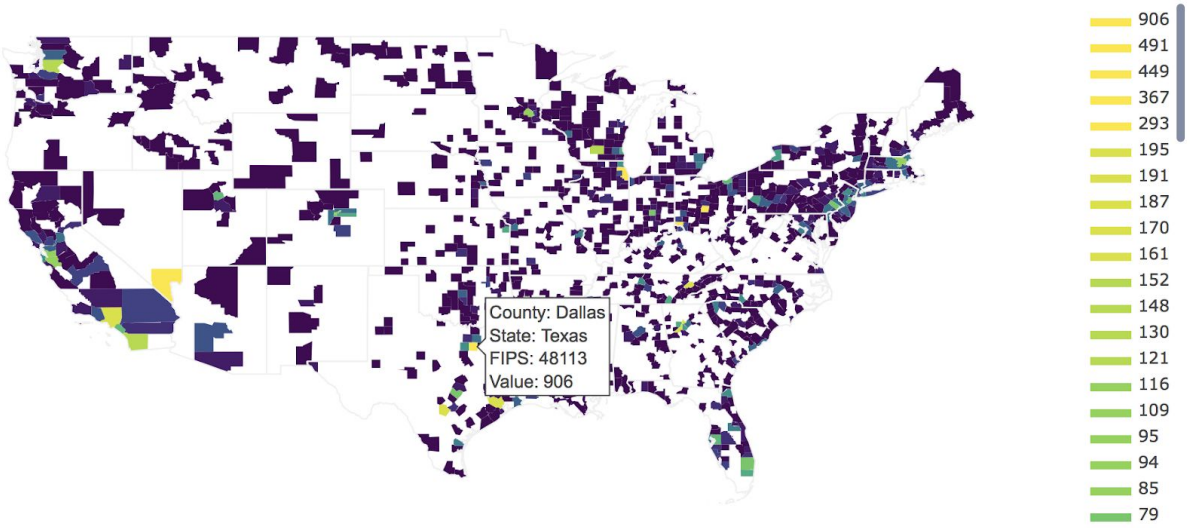
| | job_description | job_title | location | organization |
|---|---|---|---|---|
| 0 | TeamSoft is seeing an IT Support Specialist to... | IT Support Technician Job in Madison | Madison, WI 53702 | NaN |
| 1 | The Wisconsin State Journal is seeking a flexi... | Business Reporter/Editor Job in Madison | Madison, WI 53708 | Printing and Publishing |
| 2 | Report this job About the Job DePuy Synthes Co... | Johnson & Johnson Family of Companies Job Appl... | DePuy Synthes Companies is a member of Johnson... | Personal and Household Services |
| 3 | Why Join Altec? If you're considering a career... | Engineer - Quality Job in Dixon | Dixon, CA | Altec Industries |
| 4 | Position ID# 76162 # Positions 1 State CT C... | Shift Supervisor - Part-Time Job in Camphill | Camphill, PA | Retail |



Notice that Texas is the state which have the largest number of jobs and dallas is the city which has the largest number of jobs. I think this can't represent the real situation of the job market since New York City should at least be in the list. This plot and result only represent the result of the dataset.

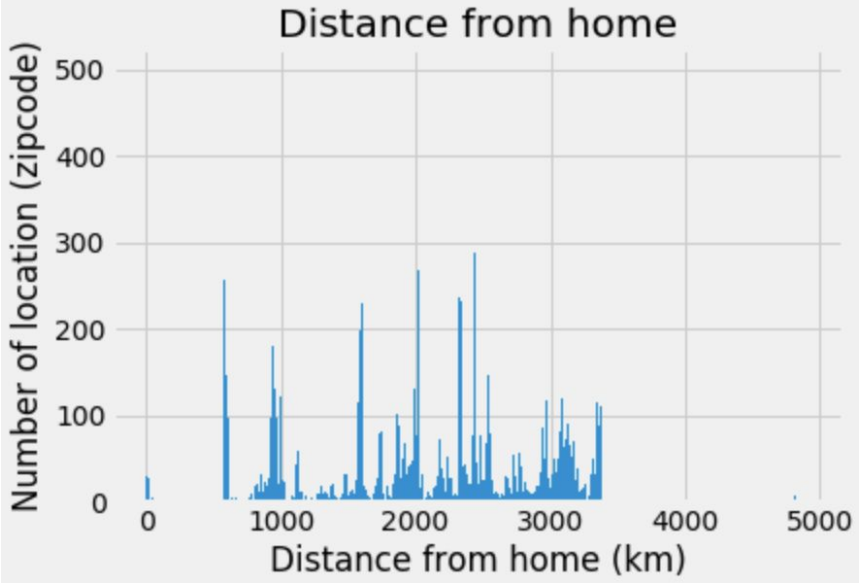
Another interesting way to visualize the distribution of job in the USA is using the choropleth map. I used Plotly to make this interactive plot. Actually, I was having a little trouble making this plot until I realized that this plot use FIPS (The Federal Information Processing Standard) code instead of zip code. So, after I convert the zip code to FIPS code, I get this choropleth map in a breeze.



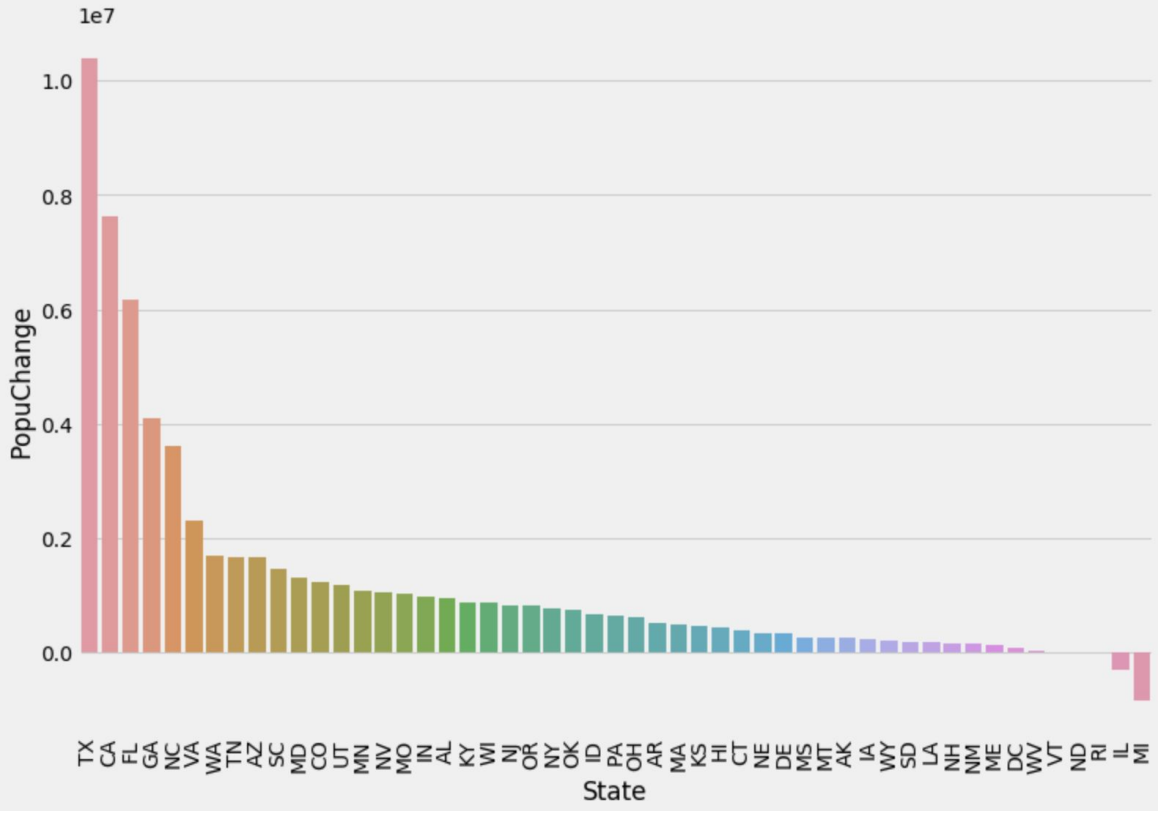

Location index

Different people have different preference about the location. For me, if I found one job that matches exactly what we want, then I will go for it wherever it is. But most of the time, I will still think about the location. I can ask a lot of questions, such as, how far away from my home? Is it a place close to cities? Is the place has a healthy population change? Here, we try to answer some of the questions by merge datasets together.

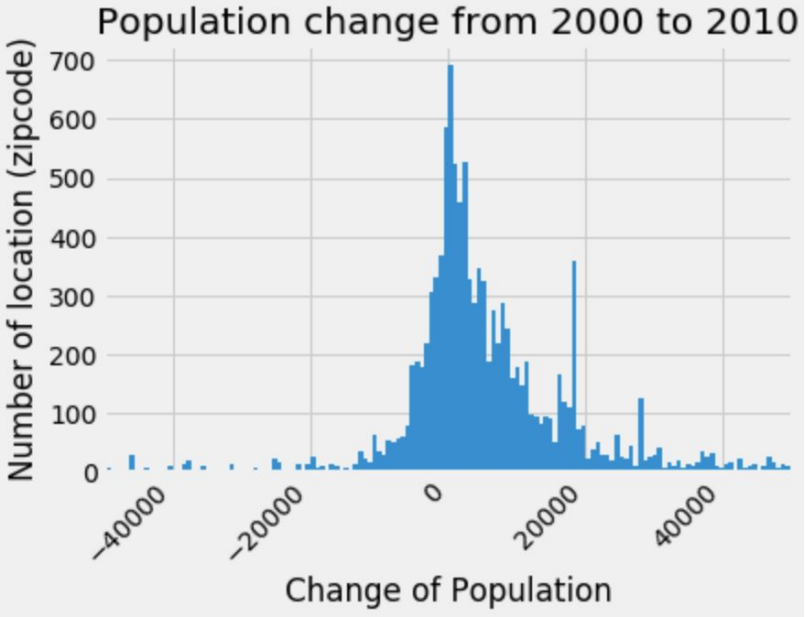
First, the distance between the job and the current location was calculated. This numbers is stored in a new column named “distance_home”. I set my current location as Salt Lake City, 84108. From the distribution of “Distance from home”, we divided the jobs into four categories: *distance_home = {[dis_min, dis_max):index} = {[0:1000):3, [1000,2000):2, [2000,3000):1, [3000,:]:0}. T*he closer, the higher the index number.




Second, we get the change of population data by comparing the population of 2000 with the population of 2010. If there is an increase of population, it may indicate that that area have a good amount job opportunity and overall healthy economy. I use this data source: [US Population By Zip Code For both 2000 and 2010]



The histogram is the distribution of population change of the locations in the jobs in the dataset. The majority of the locations have an increase of population, this is true when we look at the distribution of population change in all the locations in the USA. Based on my understanding of population change, I divided them with the following categories: *popu_diff = {[change_min, change_max):index} = {[:, -10000):-2, [-10000, 0):-1, [0,10000):1, [10000,:]:0}.* This is based on that extreme increase and decrease of population are not good, while a reasonable increase of population maybe the best scenario.




Wrap-up location index (LI)
*LI = A*popu_diff - B*distance_home. Let's say, A is 50% and B is 50%, they have equal weights.*

Culture and Industry from Job description text analysis
Everyone wants to be happy in their work, this can be strongly affected by whether someone likes the culture of the company or not. The culture of the company can depends on the industry. For example, government-related big companies makes me think of structured and slow-paced. But two companies in the same industry can still be very different. The same with industry. For example, “data science” can be very different from one position to another. Those difference require us know about the company and read through the job descriptions. In this project, we will automate the culture and industry matching by job description text analysis.
The text was cleaned first with the following steps. 1. Tokenize the sentence by separating them into words. 2. remove stopwords, symbols and make all the alphabet in lower-case. 3. word stemming. We did the step 3 and found it’s very strong (e.g. motivated was stemmed into motiv). And I didn’t do stemming for this project to ensure the text is readable.

A case study:
*'The Wisconsin State Journal is seeking a flexible and motivated reporter/editor to lead its three-person business desk. We’re looking for an experienced journalist able to spot trends, be a*

*watchdog and reflect the Madison area's vibrant entrepreneurial community. This is a hybrid reporting and editing position that calls for a fleet-footed, multimedia storyteller able to maximize the newspaper's online presentation while also editing two sections a week. Candidates must have strong news judgment, be well versed in business news and trends and be able to quickly prioritize coverage.At least five years' experience reporting or editing for digital and print platforms desired. To be considered for the position, applicants must apply online.\xa0As part of your online application, please attach five samples of your work or links to five recent stories.Wisconsin State Journal, 1901 Fish Hatchery Road, Madison, WI 53713Affirmative Action/Equal Opportunity Employer/Pre-employment drug testing appliesPI94338362Apply Here'*

In the above list, after the text analysis finding the adjective and nouns, we found that **''flexible', 'motivated', 'entrepreneurial', 'vibrant'** are good indicator of the culture of the work and they are adjectives. Other words: **'multimedia', 'online', 'digital', 'recent'** are more indicating for the industry of the work. In the list of Noun, the following words: **'business', 'trends', 'reporting', 'news', 'reporter', 'journalist'** are great indicator of the industry of this job

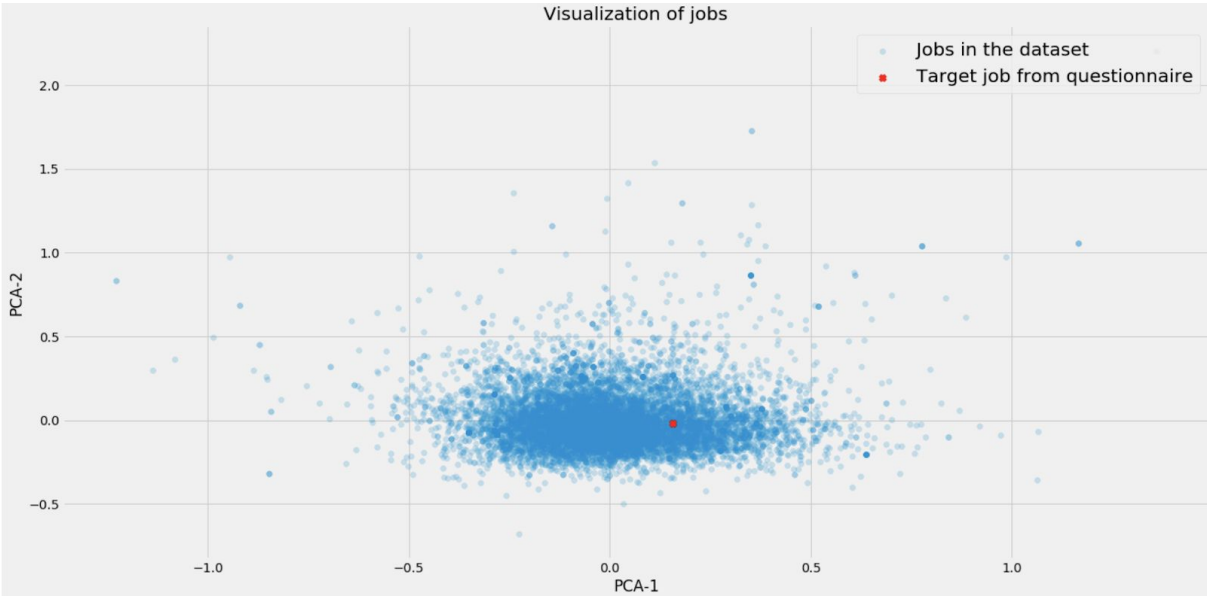
In order to do machine learning on texts. We need to find word representations (word embeddings). One commonly used word representations is bag-of-word, and another commonly used one is frequency–inverse document frequency (TF-IDF). They take the word's frequency of appearance in the library into account. But both methods do not carry the meanings of the words. So I started by using word2vec pretrained vectors to represent words. Word2vec models were trained by the the text environment of a word or we call it "skip-gram" model, which predict the words appears around a word in a window sized text of a sentence.

Google word2vec is trained by Google News dataset (about 100 billion words). The model contains 300-dimensional vectors for 3 million words and phrases.

We transfer the a job description to a vector and create the scattering plot using the following steps:

1. Each keywords (adjectives or nouns) were embedded using the google word2vec.
2. Average of all the keywords' vectors of a job description.
3. The dimension of all the job description vectors was reduced to 2 dimension by principal component analysis (PCA).
4. The same procedure was used to get the target job's vector and use the same PCA weights to reduce the dimension and plot it in the scattering plot.

Recommendation was given by calculating the distance between the target job and the available jobs in the dataset. Give the keywords for culture: **'innovative','open','art','ai',** I got some recommendation jobs. The first one is a software engineer job.The for the industry, **'healthcare', 'programming', 'research', 'analysis',** I got a clinical and statistical programming job recommended.




Combine all the index together
To combine the location index, culture and industry distance together, we need to give weights to each of the factors. For example, if location is very important, just put a higher weights on it.

Let's say:

- Location factor 20%
- Industry factor 40%
- Culture factor 40%

And we will calculate the most related job by using these weights. Location factor is a little bit tricky, since it's not in a vector format as the other factors.

For industry factor and culture factor, I will have two vector add together with equal weight (50%). Let's define the final index FI = CI x 0.4 + II x 0.4 - (LI x 0.05) x 0.2

We found Journal Development Editor role of a biomedical journal. And I think it make good sense with the keywords I given to it.

To continue...