

A Comprehensive Job Search Method: Job Matcher V1

Yaqiong Zhang
July 12th 2018

Abstract

Job matcher V1 is a job recommendation system which gives you recommendations based on your answers to a questionnaire. The questionnaire covers the kind of locations you would like your work to be, the culture of the company and the industry you are interested in. Job matcher V1 will give recommendation based on your answers to the questions and the weights you give to different answers to the questions.

Job match V1 utilizes a dataset downloaded from Kaggle.com that contains 22000 jobs from all kinds of industry. Job matcher V1 takes answers that are numbers and texts. I tried it and it gave me a surprising and meaningful job recommendation.

Table of Content

[Table of Content](#)

[Introduction](#)

[Exploratory Data Analysis](#)

[Location Index](#)

[Distance from Home](#)

[Population Change](#)

[Wrap-up Location Index \(LI\)](#)

[Culture and Industry from Job Description](#)

[A Case Study:](#)

[Combine All the Index Together](#)

[Conclusions](#)

Introduction

Job search website like LinkedIn, Indeed and glassdoor are great resources for people to look for a job. A lot of people like myself, 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 alone 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 other than browsing those websites. To solve this problem, I developed this job matcher prototype. It returns me with the suggested jobs I may be interested in based on the questionnaire I filled in and the weights I gave to each question.

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

1. **What's the location you currently in?**
Answer: 84108
2. **How do you feel about re-location?** (absolutely no, maybe no, Ok, prefer to move)?
Answer: Ok
3. **Do you want to live in a metropolitan region?** from 1(means no)-5(means must).
Answer: 4
4. **Give the keywords you think your future company's culture should be?**
Answer: 'innovative', 'open', 'art', 'ai'
5. **Give the keywords you think your future company's industry should be?**
Answer: 'healthcare', 'programming', 'research', 'analysis'

After finishing this questionnaire, we also give a weight to the question 2 to the question 5. 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 to extract the zip code, the state and the city information from the location column.

Table 1: How the dataset look like after removing the irrelevant columns.

	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

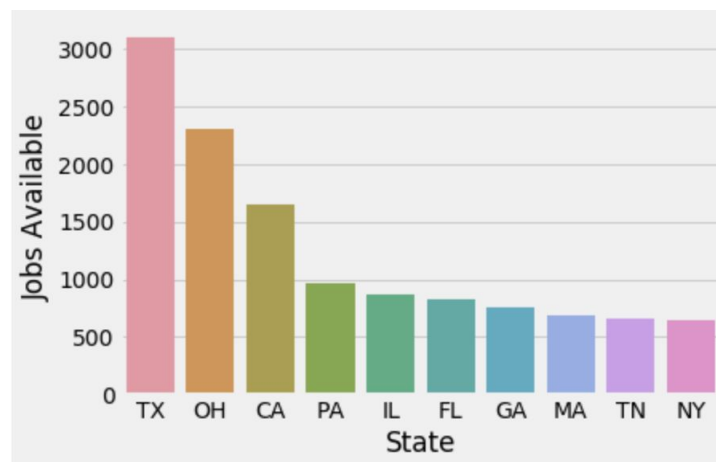


Figure 1: The distribution of jobs in different state.
TX, OH, CA has the largest number of jobs.

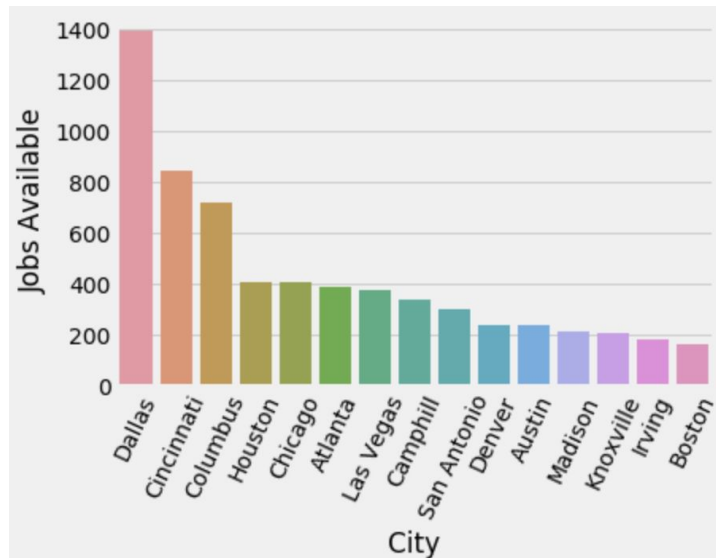


Figure 2: The distribution of jobs in different cities. **Dallas in TX, Cincinnati in OH, Columbus in OH** are the cities with largest number of jobs.

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.

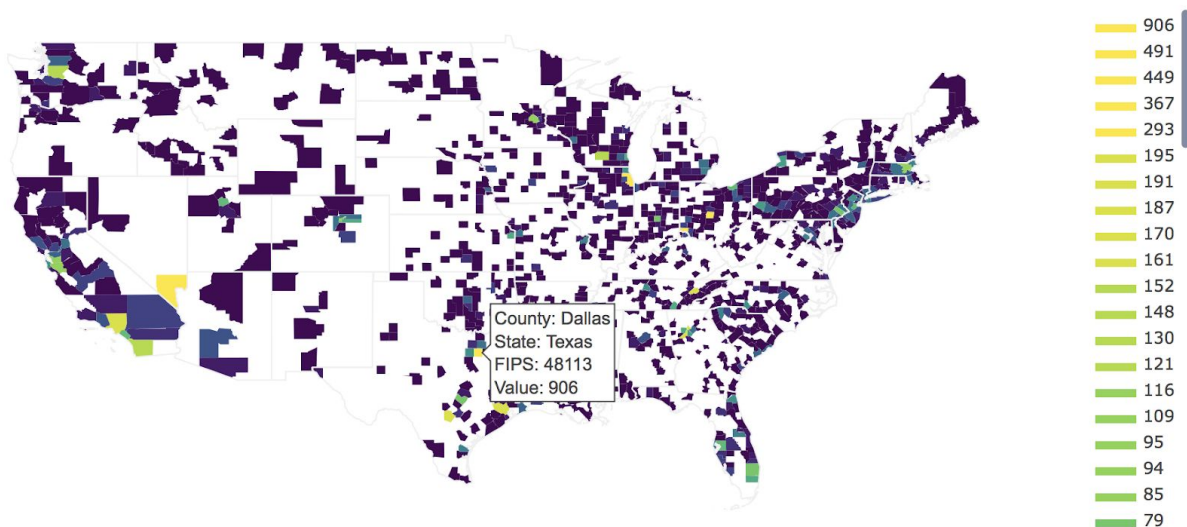


Figure 3: The distribution of jobs in a USA choropleth map. Made with Plotly.

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.



Figure 4: Illustration of possible relation due to different job opportunities.

Distance from Home

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}`. The closer, the higher the index number.

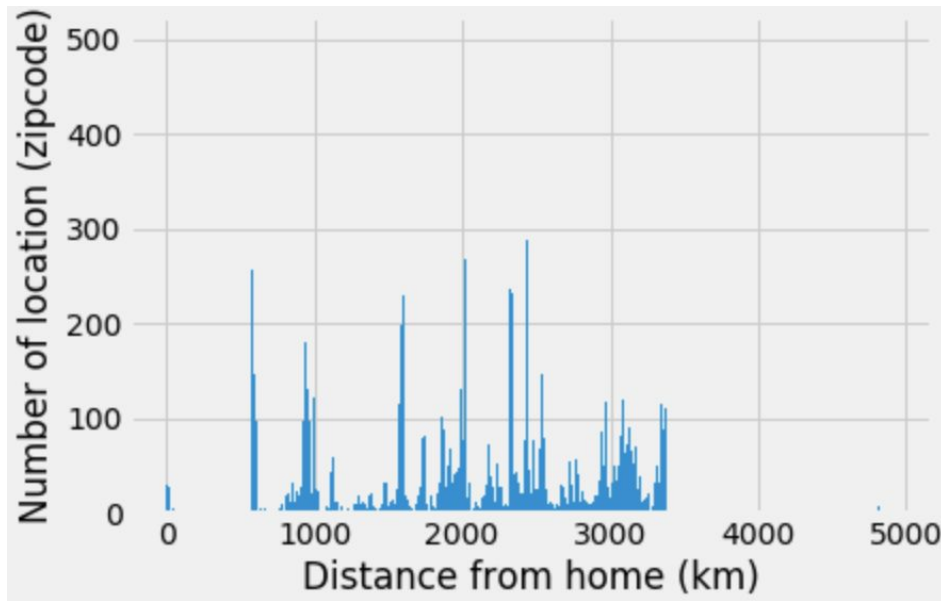


Figure 5: The distribution of the distance from home for the jobs in the dataset.

Population Change

I moved to the northeastern not long ago from Salt Lake City. I noticed that in some of the areas I lived in northeastern, there are empty stores and abandoned houses. This is related with the decrease of populations in those area. So I was thinking that I want to avoid those areas with major population decreases. Thus, I calculated the change of population data by comparing the population of 2000 with the population of 2010. Data source: [US Population By Zip Code For both 2000 and 2010](#).

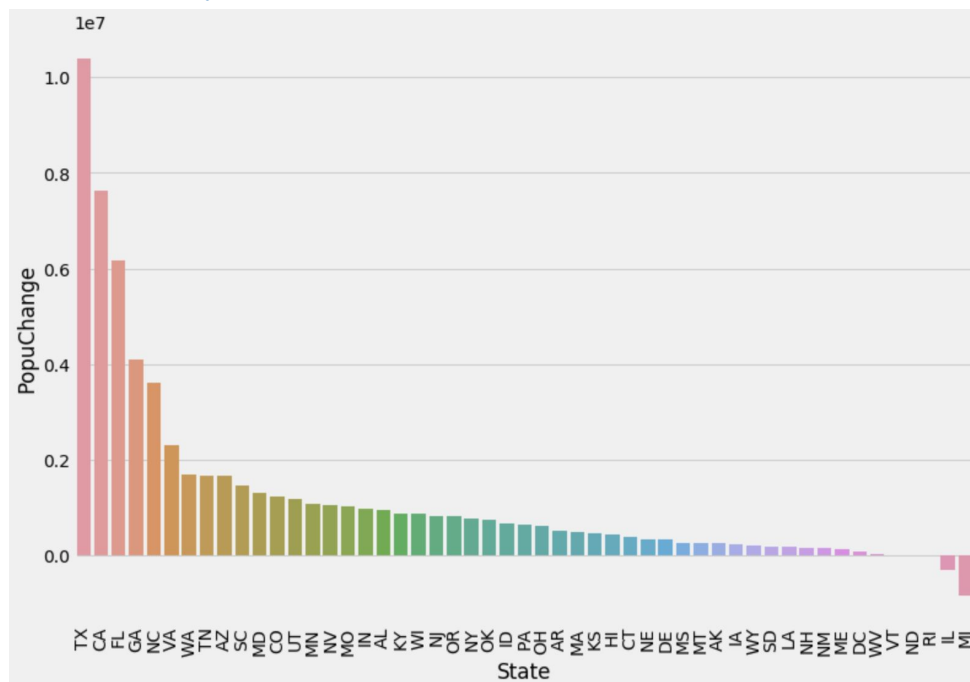


Figure 6: The population change of different states from 2000 to 2010.

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. The

Actually, the relationship between the population change and the economy is controversial. Based on my understanding of population change, I would avoid those areas with extreme increase or decrease of population, while a reasonable increase of population maybe the best scenario.

I divided them with the following categories:

```
popu_diff = {[change_min, change_max): index}  
            = {[., -10000):-2, [-10000, 0):-1, [0,10000):1, [10000,:]:0}.
```

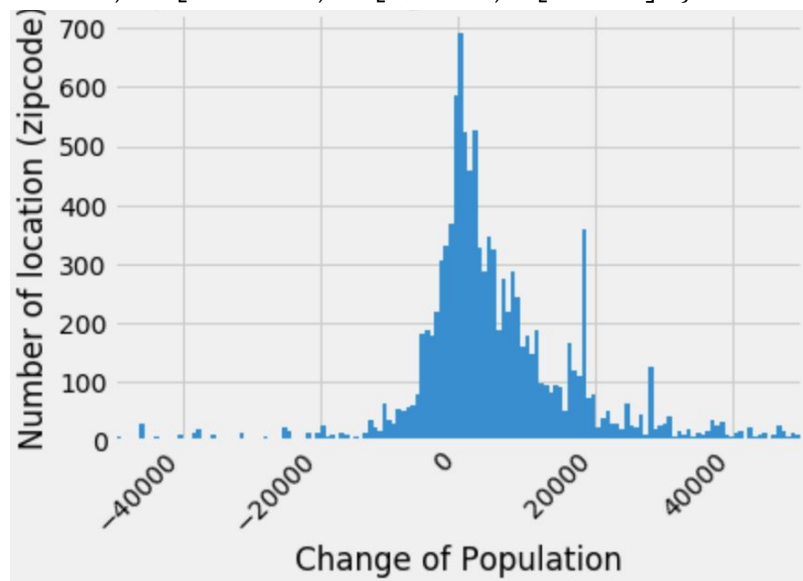


Figure 7: The distribution of population change of the job locations in the dataset.

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

Everyone wants to be happy in their work, this can be strongly affected by whether someone likes the culture of the company or not. Sometime, the culture of the company depends on the industry. For example, government-related big companies makes me think of structured and slow-paced. However, sometimes 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 the company well and read through the job descriptions carefully. 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.

- Tokenize the sentence by separating them into words.
- Remove stopwords, symbols and make all the alphabet in lower-case.
- 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 53713 Affirmative Action/Equal Opportunity Employer/Pre-employment drug testing applies PI94338362 Apply 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 Nouns, 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-words, 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 one ranked as the first one is a software engineer job. For the industry, '**healthcare, programming, research, analysis**', I got a clinical and statistical programming job recommended. I think this makes good sense.



Figure 8: Scattering plot of all the jobs' adjectives (blue dots) in the dataset. The red dot is the target job's adjective keywords projected by the same PCA model used for all the blue dots.

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 Index 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 $FI = CI \times 0.4 + II \times 0.4 - (LI \times 0.05) \times 0.2$. We found Journal Development Editor role of a biomedical journal. And I think it make good sense with the keywords I given to it.

Conclusions

Job matcher V1 is designed and prototyped in this project to give job recommendation to people based on a questionnaire. This questionnaire should take consideration of the location factor, culture and industry of a job. In this project, we utilized a datasets that contains 22000 job descriptions from web scraping. Google pre-trained word2vec word representation was used to transfer the job descriptions in the dataset into vectors, and the distance between the keywords given in the questionnaire and the job descriptions was calculated and ranked. In the meantime, the location factors, including how far is the job related relocation and the population change of the location were taking into consideration when giving recommendation.

Based on my answer to the questionnaire, I got a job recommendation of a journal development editor's role in a biomedical journal in New York City. I personally think is a good answer. I would never think about search for a editor's role before, but this one is related with scientific journal did match my interest of 'research', 'healthcare' and 'art'. The disadvantage of unsupervised machine learning is it's hard to have a direct metrix to give numbers that tells people how good my models is. This is very common for most unsupervised machine learning. However, in large companies, people can judge a unsupervised model by comparing some indirect numbers, such as the increase of sale in this season after using a model, the increase new users in a certain period of time after using a model.

Limitations

A big task of this project is to combine different factors/index together by their weights. In this project, I created a location index, a culture index and a industry index. But when combine the location index with the culture/industry index, it can be tricky because they are very different numbers, in different ranges and accuracy. So it require some normalizations and the normalizations require extra work and experience. This limits the number of different factors to be effectively combined together.

Future Work

I highly encourage you to try out job matcher V1 by entering your answers to the questionnaire. In the future, it would be beneficial to build a graphic users interface of this project or build a stand alone web application or software. It's also worthwhile to connect the data resource of job macher V1 to the job posting website, so it can give up-to-date job recommendations.