

Mining the US Technologist data

Abulitibu Tugulu
DHI Group Inc.
6465 Greenwood Plaza Blvd
Greenwood Village, CO 80111
abtu8803@colorado.edu

ABSTRACT

Several studies have reported the US will face an even bigger tech talent shortages in the near future, which will translate into revenue decreases. In this report, we will study the US technologist data through various mining techniques, by first introducing the US department of labor data along with H1b record (due to its partially reflective and openness). We then introduce the technologist concept along with data scrapping technique we applied to obtain the dataset. After carefully processing the data with exploratory analysis, we will try to understand the technologist trends such as: Which job title is becoming more popular and which one is not. The only machine learning algorithm for this project will be semi-unsupervised clustering, through which we will try to answer the utmost question of which tech skill set is becoming more demanding, and which can help our ultimate aim of recommending the right skill sets for the future tech job-seekers.

Our study shows that on top of tech talent demand in US are on the rise, is cloud engineering and project management skills will be in huge demand, while SQL and system analysis are in decline. Java still grows strong as python has not yet replace its spot, and Software developers should not worry that data scientists will make their job obsolete, at least not yet. The report will end with conclusion along the future work that can be study along the way.

1 INTRODUCTION

1.1 Why we choose to mine the technologist data

It is no secret that the new competition for this century is the 'hunt' for talents., "by 2030, more than 85 million jobs could go unfilled because there are not enough skilled people to take them, according to the latest study conducted by Korn Ferry."¹ "The shortage has been amplified by an insufficient number of U.S.-citizen computer science college graduates, restrictive (until recently) and limited H-1B visas for experienced IT professionals to fill the immediate

gaps", echoed the Tech Serve consulting.² Dice.com's internal analysis already showed that we are going through technology workers (Technologists) shortage, and that is not due to the 'Great Resignation' but due to market demand. Our report will mine through the US H1B data set along with Dice.com's job and candidate profiles (along with skill sets for each job title), to better understand which technologist job (e.g.: Software developer, Data scientist) are in increasing/decreasing demands, which skill sets (e.g.: Python, C++ etc.) are more popular/unpopular, through time-series. Our goal is to better understand the US technology market and what we can do to help people who try to get into tech market, by providing them with skill-set guidance along with professional training advice.

The motivation for this report came from a conversation with one product manager stating that there are no analysis on technologist supply and demand, even on salary data! The product analysts stated they don't have the data, and the R&D team think it is a analyst job and should not 'waste' time on. I decided to try that as a personal project. As I dive in more, I realized the official data set provided from department of labor is vague and uninformative, most of the job titles are lumped together, and focusing only on salary, hence no detailed information based on skills and locations. I then decided to mine other open source data, while adding relative skill set by mining separate job description data set.

My philosophy to tackle this problem/project is to use a simple algorithm on multiple complicated sets, and not the other way: some complex algorithms on a nice set instead. Personally, I had some 'fatigue' issues with words like 'Deep' and 'Neural' for now, both at work and school, so I've decided to do some old fashion mining by not using heavy GPU computing with 'sexy' machine learning titles, and just have some miner's fun with my hardhat and shovel.³

1.2 What is 'Technologist'?

The simplest definition of a technologist is an expert in a particular field of technology. At Dice, we focus more on an employee's job title and summarizing skill sets to define whether a candidate is technologist or not. Engineers and applied scientists are definitely in that category, unfortunately, a musician or a cleaning crew are not. Yet, if his/her resume has certain skill sets that can be applied to solve scientific problem. e.g.: programming, algorithms, him/her may still be considered as a technologist. For bookkeeping purpose, we currently consider 1468 job titles as technologist, for example: Oracle DBA, Ruby Developer, Research Analyst and Systems Engineer.

¹<https://www.kornferry.com/insights/this-week-in-leadership/talent-crunch-future-of-work>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CSCI 5502, Boulder, University of Colorado

© 2022 ACM. 978-x-xxxx-xxxx-x/YY/MM...\$15.00

DOI: 10.1145/nmnnnnn.nnnnnnn

²<https://www.techservealliance.org/news/the-state-of-the-technology-talent-shortage/>

³I will leave that to CSCI 5922-Neural Networks and Deep Learning, this fall.

1.3 What is 'Skill set'?

As the name suggests, it is the skill that associate with job title of a candidate. Although a individual, such as myself, would like to fit everything single skill that she has ever grasped into a resume, we only consider certain words that are associated with the job title. Take Oracle DBA for example, we tried to limit the particular tech words that are relatively 'unique' or necessary for the job function, so it has 'IT management; PL/SQL; SAN; Unix; Statspack; OEM; Performance tuning; Oracle; Oracle database administration; AWR' as its skill sets. Other skills such as excel and word, we do not consider particularly outstanding compare to OEM, in other word: it is a skill everyone has, which makes them less relevant. Other examples are:

- Ruby Developer: Software development; React.js; Ruby on Rails; Ruby; JavaScript
- Systems Engineer: Computer networking; Systems engineering; Quality assurance

Next, we will introduce the previous work that has been done to 'stem' the skill set for each technologist title.

2 PREVIOUS/RELATED WORK

I was always interested in information retrieval and knowledge graph, this project gives me the perfect opportunity to pursue those. Before we can join the skill-sets with H1b data, we need to first 'extract' the keywords. What I did was using POS tagging to set up skill keywords and trained on 10000+ job title data set, so that to get the labeling for new data set working more accurate. Afterward, we clustered multiple job description for certain job titles and found the common skill words for each job. SpaCy's Pipeline component for part-of-speech tagging⁴ comes in really handy for the data preparation. What we did was to manually add a 'Tag' button for SMEs to spot the 'tech' words from our training set in HTML, resembling the following table:

	Sentence	Word	Tag	POS	Job description #
0	1	Softwares	Tech	NNPS	description: 1
1	2	SQL	Tech	NNP	description: 1
2	3	Java	Tech	NNPJ	description: 1
3	4	Web	Tech	NNS	description: 1
4	5	Azure	Tech	NNP	description: 1
⋮	⋮	⋮	⋮	⋮	⋮

Once we got enough samples, we apply the Viterbi algo to the future data set and mine the list of words for each job post, we then trim the list to get the overall skill-set for each title, like the following for sample job titles.

Job	Skills
Net Application Developer	Microsoft technologies;Software development;C#;HTML;Quality assurance;ASP.NET;Visual Basic .NET;.NET;Agile.
Android Developer	Software development;Java;Mobile development;Quality assurance;Android development.
⋮	⋮

Once we got the whole table for all job titles, we are ready for the major task for this project, getting the data, which will be explained in details in the next section.

3 DATA SET

3.1 U.S. Department of Labor data

What better resource to go to than the Department of Labor's statistics, After consulting with a product manager who points to the open API department provided, I decide to mine their data first. Shockley though, their data does not give any insight on specific job title but a very general one. I've chosen to call the data between year 2010 and 2021 so that to better match with H1b's time series, and here's what we get, after thorough analysis of 14027 data entries,with total 34 columns:

Table 1: Department of labour data

	count	unique	top
occ_code	14027	1507	29-2010
occ_group	9731	5	detailed
occ_title	14027	1272	Tour and Travel Guides
group	46	2	major
tot_emp	14027.0	8936.0	11860.0
annual	852	1	True
hourly	62	1	True
year	14027.0	NaN	NaN
area	2658.0	NaN	NaN
area_title	2658	1	U.S.
prim_state	1329	1	US
emp_prse	14027.0	229.0	0.5
mean_prse	14027.0	NaN	NaN
a_mean	14027	6155	*
a_median	14027	5808	#

There are 1272 job titles from department of labor open data, among them, the only code (15-XXXX) that resemble our definition of technologist is :

- Computer and Mathematical Occupations 15-0000

Unfortunately, only 37 of them instead what we proposed of roughly 1400+ is publicly stored, they are:

- 'Computer and Mathematical Occupations',
- 'Computer and Information Research Scientists',

⁴<https://spacy.io/api/tagger>

- 'Computer Systems Analysts', 'Computer Programmers',
- 'Software Developers, Applications',
- 'Software Developers, Systems Software', 'Database Administrators',
- 'Network and Computer Systems Administrators*',
- 'Computer Support Specialists',
- 'Information Security Analysts, Web Developers, and Computer Network Architects',
- 'Computer Occupations, All Other*', 'Actuaries', 'Mathematicians',
- 'Operations Research Analysts', 'Statisticians',
- 'Mathematical Technicians',
- 'Mathematical Science Occupations, All Other',
- 'Computer Occupations', 'Computer and Information Analysts',
- 'Information Security Analysts',
- 'Software Developers and Programmers', 'Web Developers',
- 'Database and Systems Administrators and Network Architects',
- 'Network and Computer Systems Administrators',
- 'Computer Network Architects', 'Computer User Support Specialists',
- 'Computer Network Support Specialists',
- 'Miscellaneous Computer Occupations',
- 'Computer Occupations, All Other',
- 'Mathematical Science Occupations',
- 'Miscellaneous Mathematical Science Occupations',
- 'Database and Network Administrators and Architects',
- 'Database Administrators and Architects',
- 'Software and Web Developers, Programmers, and Testers',
- 'Software Developers and Software Quality Assurance Analysts and Testers',
- 'Web Developers and Digital Interface Designers',
- 'Data Scientists and Mathematical Science Occupations, All Other'.

This dataset does not provide any details on what do these occupations include. We tried some online resources⁵ (including DoL's own mapping data from 2018⁶) to try mapping the 1400+ tech job into these 37 categories that yields horribly inaccurate result, which mean we either has to add a new attribute by clustering the 1400 into 15, or again write a new scrapper that mines Glassdoor's data (This should not end up being a scrapper project), even that does not guarantee a good accuracy since all are self-reported, hence more biased than H1b data. Due to the time limit, we have to reconcile that.

Although, the discovery does not mean DoL's data are completely worthless, there are some good insights that confirm our believe that tech talents are in rising demand. Take Figure 1 for example: The top 2 lines represent 'computer occupation' and 'software developer' professions, their growing rates are even higher than the rest of technologists, it is a clear indication of higher demand for such labor, as opposed to the Farming jobs, which stays relatively the same, in some cases, even declined, as shown in Figure 2. If we look at the salary data for technologist in Figure 3, we noticed the significant salary increase over the year, take into account the inflation and supply, it only amply the demand side of the tech market, which forces the employers to hand out competitive salary. As oppose farming job market shown in Figure 4, where it did not even cover the inflation over a decade.

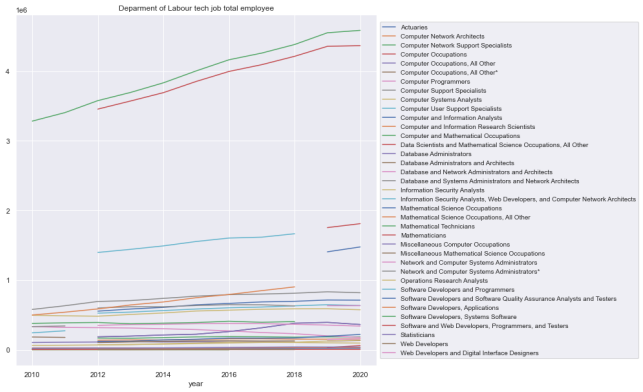


Figure 1: Department of Labour tech jobs number of employee.

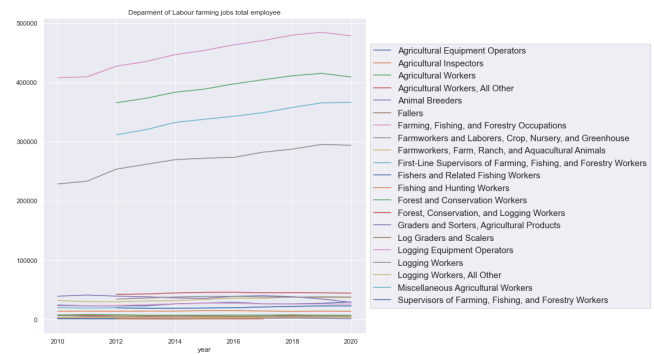


Figure 2: Department of Labour farming jobs number of employee.

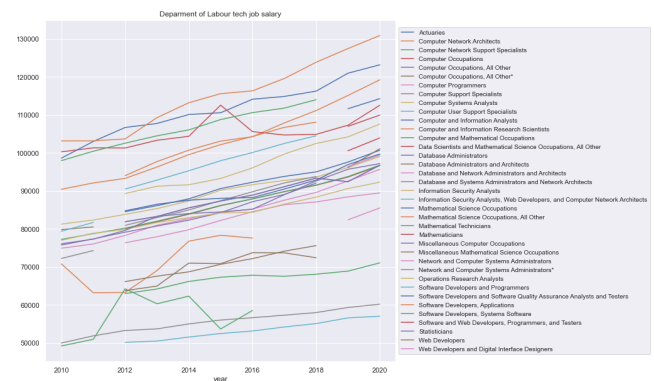


Figure 3: Department of Labour tech job salary.

All of these give us more reasons to find our own data, as an immigrant worker myself, the second 'gold mine' to dig is H1b, which will be the main focus of the next section.

⁵https://occupationdata.github.io/apst_mapping.pdf

⁶<https://www.bls.gov/soc/2018/home.htm>, only 20+ were able to matched, that leaves 1380+ unmatched still

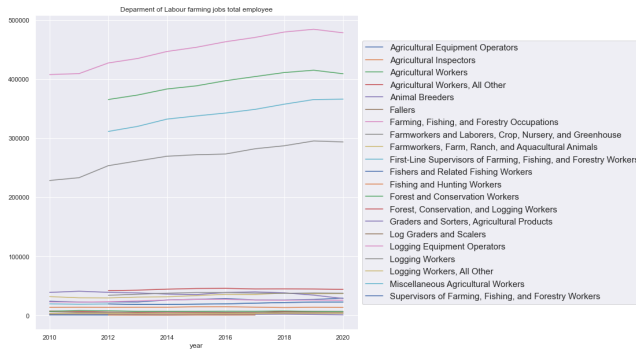


Figure 4: Department of Labour farming jobs salary.

3.2 H1b data

H-1B is a visa in the United States that allows U.S. employers to temporarily employ foreign workers in specialty occupations (mainly tech). Although, some may argue that "The H-1B temporary visa program has been exploited and abused by employers primarily seeking to fill entry-level positions and reduce overall business costs"⁷, that does not necessarily mean these jobs are not in demand, quite the contrary, it is due to market demands. After all, it is not a charity for foreign workers, it is a sign of demand for particular skill-sets here in this country, unfortunately, some company tries to find loopholes in the process.

According to H1b law code, the occupation as requiring theoretical and practical application of a body of "highly specialized knowledge in a field of human endeavor", including but not limited to 'biotechnology, chemistry, computing, architecture, engineering, statistics, physical sciences, journalism, medicine and health, economics, education, research, law, accounting, business specialties, technical writing', and it requires the applicants to have a bachelor's degree or its equivalent as a minimum to practice in that field. This is a clear indication that it is a technologist candidate, which is another reason I chose such a open source as my project data. We will now move on how to collect such a data from web source.

3.3 Data scrapping

Python beautifulsoup is the go-to package for data types of online HTML in my opinion. From the following url link:

```
url = "https://h1bdata.info/index.php?em=&job=%s&city=%s&year=%d"
```

we can tailored the job and city and year as parameters, the HTML table has the following six attributes:

- EMPLOYER : String, Nominal
- JOB TITLE : String, Nominal
- BASE SALARY: Float, Interval
- LOCATION: String, Nominal
- SUBMIT DATE: (date) String, Ordinal
- START DATE: (date) String, Ordinal

We only set the job variable as tech job titles we summarized earlier, and the raw data looks like the following:

Next, we will look at the exploratory data analysis.

⁷<https://www.uscis.gov>

Table 2: Scrapped data as off July 15th

	count	unique	top	freq
Dice_job_title	2495079	1180	Systems Analyst	198860
EMPLOYER	2495052	66795	TATA CONSULTANCY SERVICES LIMITED	157580
JOB TITLE	2467304	1770	SYSTEMS ANALYST	198719
BASE SALARY	2467304	37871	60,000	146505
LOCATION	2467304	10474	NEW YORK, NY	136642
SUBMIT DATE	2467304	2749	03/13/2015	14268
START DATE	2467304	2898	09/01/2015	32532
Job_Title	2495079	1180	Systems Analyst	198860

3.4 EDA

Figure 5 shows the top 10 companies that submit H1b over the year, and we can clearly spot the visa 'abusers'. But what is more interesting is that it has some bit tech company's on the list, I do not consider it as a sign of abuse, due to the 'tech' nature of IBM, Google, and Microsoft, quite the opposite, it re-affirmed the thesis of this project: tech skills are in high demands.

Figure 6 shows that the top job title filed for the visa, and it is

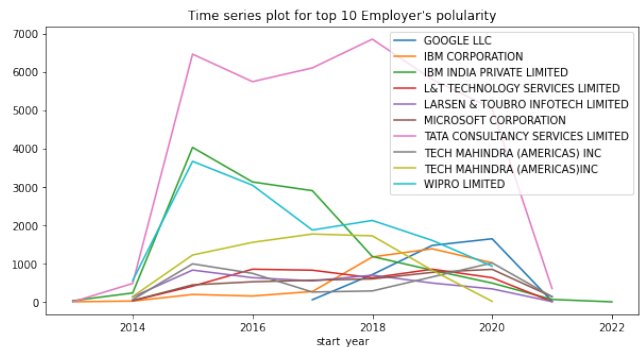


Figure 5: Top 10 companies filing H1b visa.

clearly the 'trendy' titles.

Figure 7 indicated the states that have the most H1b submission,

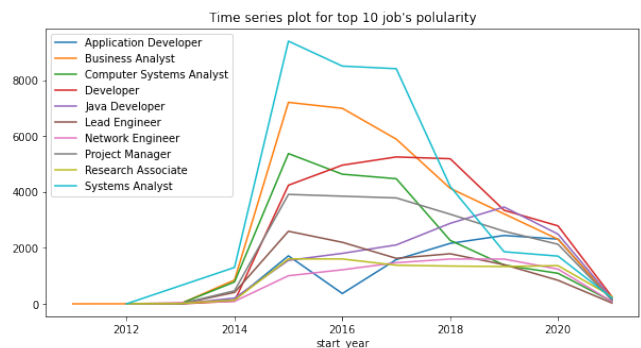


Figure 6: Top 10 job applied for H1b visa over 10 year period.

not very far from the talent pools.

Two more information we can read from these plotting. First, it is a clear indication that after 2016 elections, there are some declined on H1b application due to policies. Second, the outlier data between

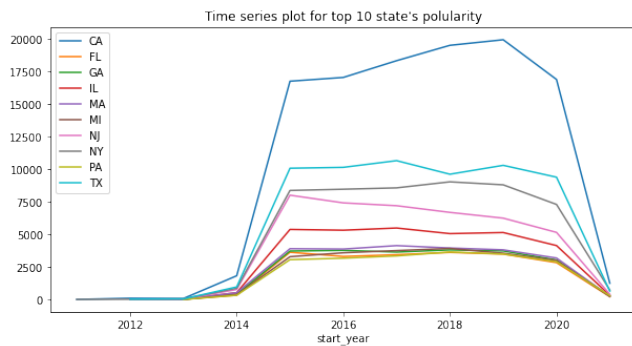


Figure 7: Top 10 states filed for H1b visa.

year 2012 to 2014, and again from year 2020 to 2021. Shows that we do not have enough data on those particular years, and not because the titles are dying. If we focus only on year between 2015 to 2019 and do linear regression, we can easily see the demands should in fact go up.

Last but not the least, the salary data from Figure 8 and Figure 9 may help us understand where are the talents needed and preferable jobs that are out there. In addition, we can spot some of the last

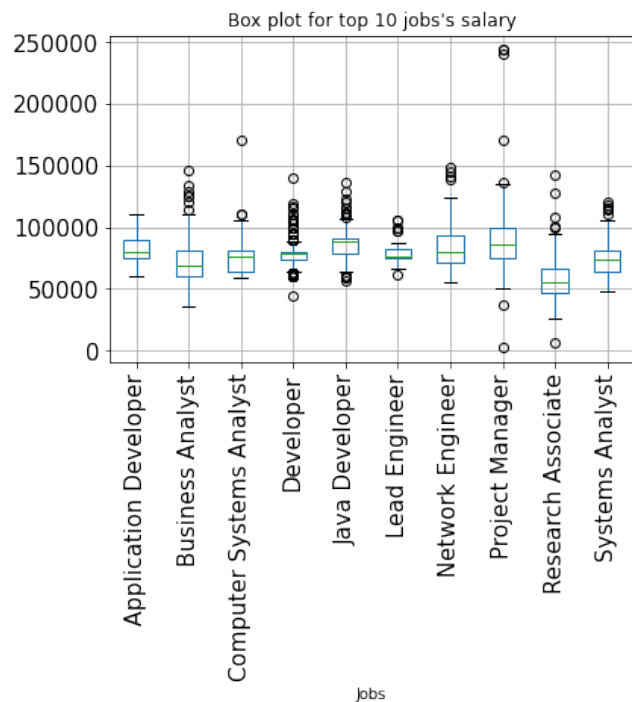


Figure 8: Top 10 job salary box plot.

popular job titles associated with H1b application as in Table 3 The least popular states as shown in Figure 10.

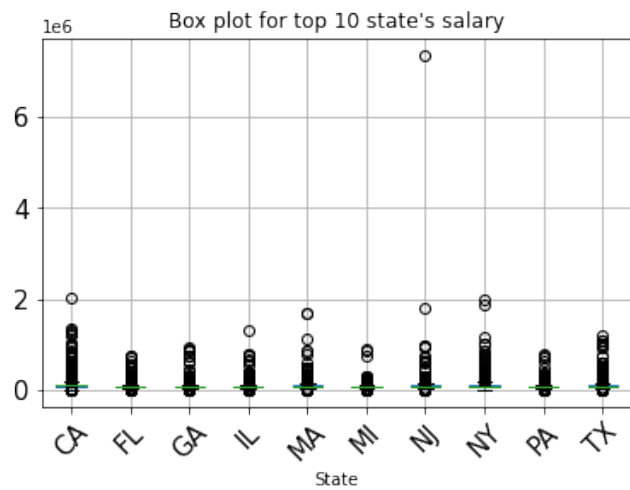


Figure 9: Top 10 state salary box plot.

Table 3: Least 'popular' technologist job titles in H1b data

	Dice_job_title
300	Dynamics Functional Analyst
282	Director Infrastructure
354	F# Developer
1076	Vulnerability Management Analyst
1067	Vertica DBA
63	Avaya Engineer
255	Desktop Support Analyst
291	Disaster Recovery Specialist
345	Epic Consultant

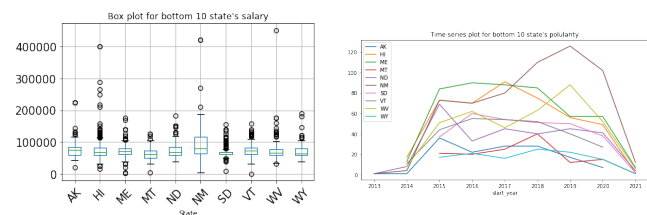


Figure 10: Bottom 10 states for hiring tech talents.

4 MAIN TECHNIQUES APPLIED

Once we are done with data scrapping using beautiful soup and analyzed with pandas along matplotlib, we can move on to the most important part of the project: data cleaning.

4.1 Data cleaning

This is by far the most time-consuming process, yet we can not ignore. Duplicity is a tricky concept for this data set, due to the highly likely possibility of a company has multiple submissions on the same day at the same city. Yet, we decided to drop the duplicates, because we believe a single hire is already a strong signal for that particular job title, and we want to avoid the unnecessary noise while clustering.

Outlier detection is another process that needed to be carefully orchestrated, not much so for the job title, but for the salary data. As we already noticed in Figure 9, that some of the salaries are just too high for it to be considered worth averaging, we decide to ignore any salary that approximate 1 million, but do little for the lower end of the tile, in order to spot the salary abuse case. In the end, we ended up having a fewer number of data point (70% shrinks), as shown in Table 4.

Table 4: New h1b data after data cleaning process

	count	unique	top	freq
Dice_job_title	603967	1107	Systems Analyst	35605
EMPLOYER	603967	66538	TATA CONSULTANCY SERVICES LIMITED	36885
JOB TITLE	603967	1655	SYSTEMS ANALYST	35605
BASE SALARY	603967	37791	60,000	23256
LOCATION	603967	10453	NEW YORK, NY	34957
SUBMIT DATE	603967	2747	03/16/2018	2377
START DATE	603967	2897	10/01/2020	17065
Job_Title	603967	1107	Systems Analyst	35605

4.2 Data processing and integration

After cleaning, we can then combine the skill-set data with H1b data, Pandas merge and join function can do the trick (we also need astype method to convert the objects into the right data type, along with splitting the location string and date string to extract the state and city information for H1b starting year info).

```
pd.merge(df_project, jobs, left_on= ['occ_title'],
right_on= ['Job_Title'],
how = 'outer')
```

The data frame in the end would look like: The reason for that is to associate each title with particular skill set that we clustered in the previous work. Now we are ready to cluster the data we have.

4.3 Machine learning

No project seems complete without the magic hand of machine learning nowadays. Although I personally believe we are still at baby stage of vector computing, as Judea Pearl called it: the first ladder of causality: Association⁸, it is still worth the effort since it suite well with more data collection with better computing resource. Yet, we can not naively think it as a black box just to get the job done, we are at the stage of realizing that iterative methods can only you so far, to quote François Chollet "Training ever bigger convnets and LSTMs on ever bigger datasets gets us closer to Strong AI – in the same sense that building taller towers gets us closer to the moon." Next, we consider a unsupervised learning for the model, since we have no idea which skill-set should go to which cluster.

⁸The Book of Why: The New Science of Cause and Effect. 2018.

Table 5: Data frame(pandas) for the New h1b with integration on skill sets

row_id	employer	location	job_title	base_salary	submit_date	start_date	end_date
1	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
2	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
3	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
4	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
5	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
6	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
7	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
8	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
9	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
10	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
11	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
12	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
13	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
14	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
15	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
16	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
17	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
18	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
19	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
20	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
21	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
22	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
23	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
24	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
25	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
26	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
27	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
28	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
29	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
30	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
31	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
32	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
33	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
34	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
35	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
36	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
37	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
38	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
39	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
40	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
41	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
42	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
43	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
44	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
45	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
46	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
47	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
48	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
49	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014
50	NET APPLICATION DEVELOPER	NEW YORK, NY	NET APPLICATION DEVELOPER	60000	01/18/2014	01/18/2014	01/18/2014

4.3.1 *K-means*. Word2Vec algorithm is a natural language processing technique invented at Google. It maps words to vectors of real numbers, or in other words, embedding the word into vectors. We would be using sk-learn package mainly, to first vectorize the skill-set string of words, with the help of cluster's **KMeans** package to cluster the texts. The logic here is use the vector representations of words learned from word2vec, and semantically put similar words close to each other in a 2-D vector space. Using K means algorithm, we then can group words from the text into similar groups and analyze the results.

We use Principal Component Analysis (PCA) for the purpose of dimensionality reduction, in order to map the work into 2-D space. A example of 5 clusters is shown in Figure 11

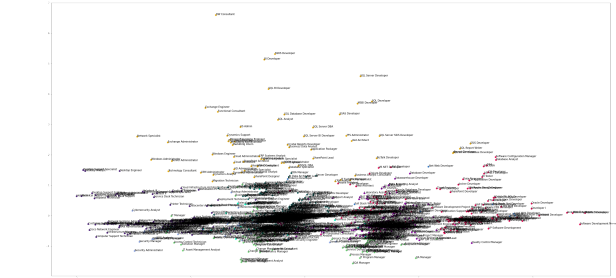


Figure 11: Clustering by job skill sets

4.3.2 *Evaluation of the clustering*. The elbow method helps to choose the optimum value of k by fitting the model with a range of values for the number of clusters. We can consider it as a 'cost'-function, we define it as We defined it as and ϵ which is sum of squares distance between data points and respective centroid of cluster to which the data point belongs, we stop when the ϵ (iterations) converges. But, as we see in Figure 12, there is no 'optimal' number of neighbor that we can choose from. One of the reasons for that is the tech skills for the job are 'convoluted', meaning

the distance between each job title are closer than we first vision. Yet, we can still borrow from department of labor data set when categorizing the tech job titles, and set the k equals to 10.

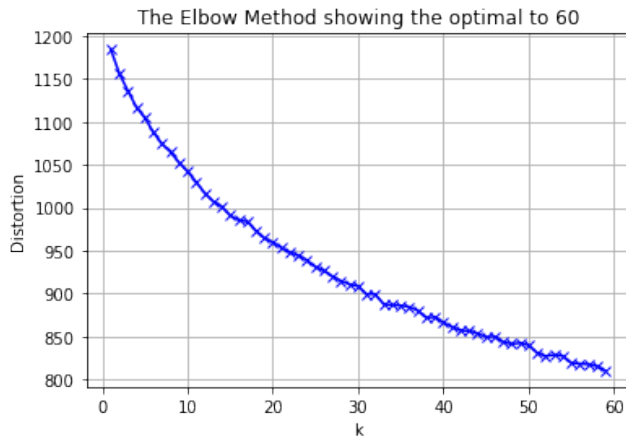


Figure 12: Clustering Elbow method by job skill sets

Training on all H1b data set does not make much difference, as shown in Figure 13, it take longer to train while the results are the same, yet the distance are way higher.

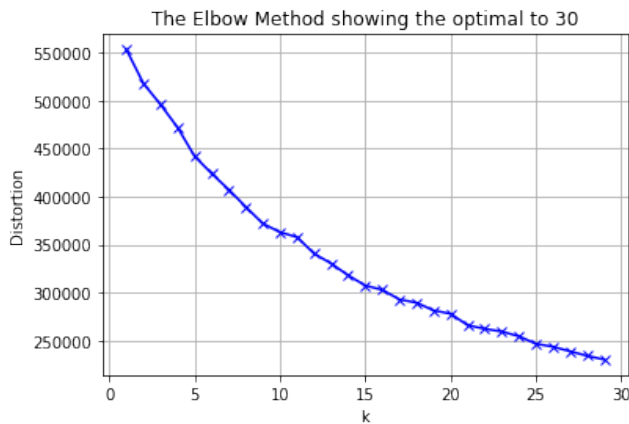


Figure 13: Clustering Elbow method using all H1b data

5 KEY RESULTS

5.1 Trending Skills

We successfully mined the top 10 job titles that are in demand, they are: System analyst, Developer, Business Analyst, Computer System analyst, System Administrator, Project manager, Java Developer, System engineer, Application Developer. Lead Engineer. As opposed to media's champion: Data scientist and ML engineer. Next, we will explore the skill trend. Figure 14 shows us a clear image of what are the top skill that dominate each year's skill. QA and Software developer ranked 1 and 2 with no decline in sight. System analyst,

on the other hand, is becoming less popular over the year, probably due to job automation, so is SQL. Cloud computing is on the rise, along with agile, it is a clear sign that cloud engineering and project management will be in huge demand in the future. Next, let us take

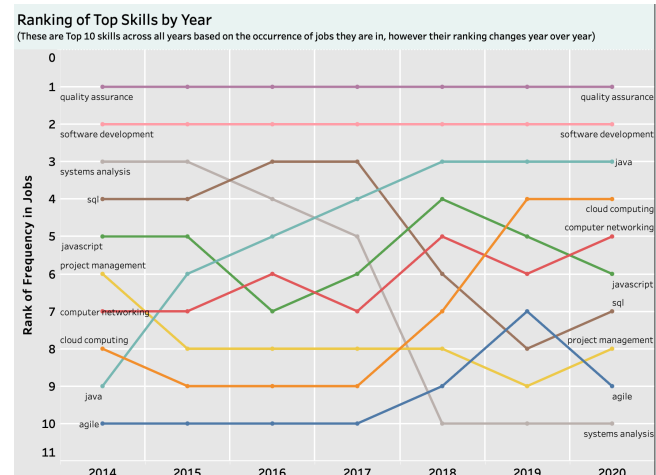


Figure 14: Skill 'trendiness' throughout the years

a look at the clustering results.

5.2 Clustering skill set

The Following are most frequent skills belong to each (total 10) cluster:

- Cluster 0
 - sap
 - sap fi
 - sap fico
 - sap grc
 - sap mm
 - sap pp
- Cluster 1
 - data warehouse
 - database
 - etl
 - microsoft sql server
 - microsoft ssis
 - quality assurance
 - software development
 - sql
- Cluster 2
 - computer networking
 - cyber security
 - dod
 - firewall
 - information security
 - policies
 - quality assurance
 - security clearance
- Cluster 3
 - analytics
 - database
 - health care
 - marketing
 - oracle
 - quality assurance
 - recruiting
 - research
 - software development
 - training
- Cluster 4
 - amazon web services
 - automation
 - cloud computing
 - computer networking
 - java
 - microsoft windows azure
 - python
 - quality assurance
 - sales
 - salesforce.com
 - software development

- Cluster 5
 - angularjs
 - cascading style sheets
 - html
 - html5
 - javascript
 - php
 - quality assurance
 - react.js
 - software development
 - ui
 - web development
- Cluster 6
 - brand
 - business development
 - customer relationship management
 - marketing
 - microsoft windows
 - quality assurance
 - recruiting
 - sales
 - salesforce.com
 - social media
- Cluster 7
 - cisco
 - cisco certifications
 - computer hardware
 - computer networking
 - microsoft windows
 - quality assurance
 - switches
 - technical support
- Cluster 8
 - agile
 - budget
 - it service management
 - planning
 - program management
 - project management
 - project planning
 - quality assurance
 - recruiting
 - training
- Cluster 9
 - .net
 - agile
 - automated testing
 - automation
 - c #
 - java
 - quality assurance
 - software development
 - software engineering
 - sql

We can easily guess that cluster 0 is SAP related, 1 is database. 2 is cyber-security, 4 is marketing, 5 is web developer, 7 computer is hardware, 8 is project management, 9 is software engineering. Only cluster 3(health-tech) and 6 (Fin-tech) are a bit hard to define. Overall results for unsupervised learning is beyond our early expectation.

6 APPLICATIONS

There are infinite fields that our knowledge gain from this project be applied to. First, is the bigger picture of technology demand in the US. For any recruiting company, the huge demand of particular job titles are the first thing they should pay attention to, since it is consider the 21st century capital, and need to competed for.

Second is for job market on which skill-sets are in need so that job-seekers can have a better vision. Words like ‘A.I.’ and ‘Machine learning’ are too vague to be focused on, while skill-set from our cluster study shows that:

Third, there can be a better job category mapping within the knowledge gained, how and why certain jobs should be labeled as the same title with text clustering and Fuzzy matching. Many applications can be achieved with the future work on the same topic, due to the time limit for this project, we will sketch out the big picture for the future study in the next section.

7 FUTURE WORK

7.1 Knowledge Graph

A knowledge graph, also known as a semantic network, represents a network of real-world entities—i.e. objects, events, situations, or concepts—and illustrates the relationship between them. This information is usually stored in a graph database and visualized as a graph structure, prompting the term knowledge “graph.”⁹ Many web companies has their own graph base. For example, Google¹⁰ uses it to enhance search engine results with information gathered from a variety of sources. We can absolutely do the same for technologist skills.

Figure 15 show the over all skill set connection training from the skill-set, we can definitely see the connection between each skill, and that is a ‘gold mine’ by itself, from this ‘mine’, we can easily map out the skill set for each job so that to find the ‘shortest path’ between each skill, e.g.: if a job is looking for a JAVA programmer, we can visualize a C++ person’s resume, and see it is a good fit if not enough talent is found around that area. Figure 16 is a good demonstration on how close some skill with eachother.

7.2 Supply and demand of talent

At Dice, we have almost all the tech job posts along with candidates’ profile, and yet we do not have a good study on supply side of the market, if by mining with geo-location data, we can map out the talent pool around certain title demand, e.g.: if a company in Wyoming is looking for a particular skill set job and our database show no candidate in that area, we can recommend the profile close to that zip code, with the same or similar skill, a win-win situation for both.

7.3 Semi-supervised clustering

As we see in Figure 12, there is not a definite cluster number that satisfies a good classification, and by introducing a semi supervised learning cluster may give us a satisfying result. As shown in Table 6, we can perhaps fine tune the clusters by first label some of the titles, so that the other ones can ‘follow’ the lead by learning the similar skill.

Table 6: job title mapping

2018 SOC Code	2018 SOC Title	2018 SOC Direct Match Title	Job Title
0 15-1211	Computer Systems Analysts	Information Systems Analyst	Information Systems Analyst
1 15-1211	Computer Systems Analysts	Systems Architect	Systems Architect
2 15-1212	Information Security Analysts	Network Security Analyst	Network Security Analyst
3 15-1231	Computer Network Support Specialists	Network Support Technician	Network Support Technician
4 15-1231	Computer Network Support Specialists	Network Technician	Network Technician
5 15-1232	Computer User Support Specialists	IT Support Specialist	IT Support Specialist
6 15-1241	Computer Network Architects	Network Engineer	Network Engineer
7 15-1243	Database Architects	Data Architect	Data Architect
8 15-1243	Database Architects	Data Integration Specialist	Data Integration Specialist
9 15-1243	Database Architects	Database Developer	Database Developer
10 15-1244	Network and Computer Systems Administrators	LAN Administrator	LAN Administrator
11 15-1244	Network and Computer Systems Administrators	Network Analyst	Network Analyst
12 15-1244	Network and Computer Systems Administrators	Network Systems Administrator	Network Systems Administrator
13 15-1252	Software Developers	Software Systems Engineer	Software Systems Engineer
14 15-1253	Software Quality Assurance Analysts and Testers	Software Quality Engineer	Software Quality Engineer
15 15-1254	Web Developers	Web Architect	Web Architect
16 15-1254	Web Developers	Web Developer	Web Developer
17 15-1255	Web and Digital Interface Designers	Digital Designer	Digital Designer
18 15-1255	Web and Digital Interface Designers	Web Content Specialist	Web Content Specialist
19 15-2031	Operations Research Analysts	Operations Analyst	Operations Analyst
20 15-2041	Statisticians	Biostatistician	Biostatistician
21 15-2051	Data Scientists	Data Visualization Developer	Data Visualization Developer

⁹IBM blog

¹⁰https://en.wikipedia.org/wiki/Google_Knowledge_Graph

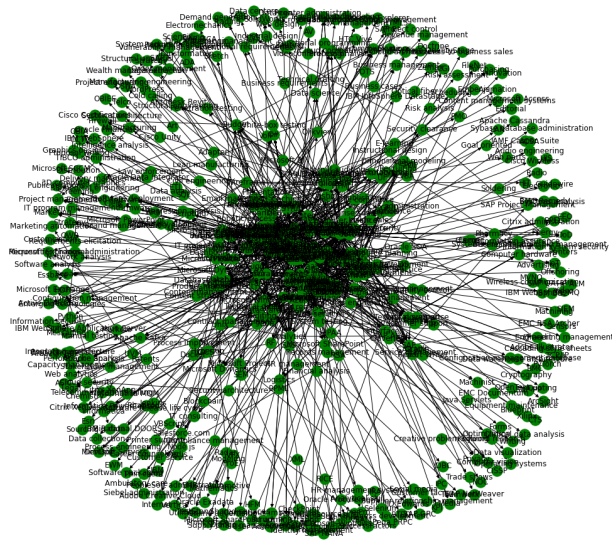


Figure 15: Overall connection of skillsets

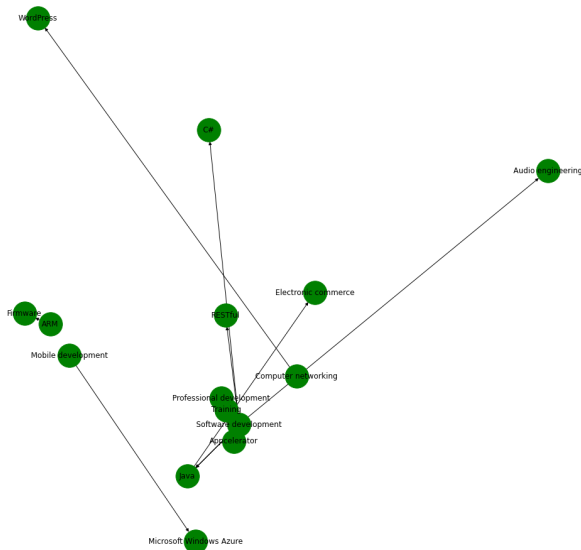


Figure 16: Directed graph between certain skill.

7.4 Skill-sets Prediction

One cool feature of time-series data is to predict using regression, yet our data collection failed at obtaining the recent H1b data to be able to study the growth rate of each skill set, in order to predict the future popular skill, this will help a company who focus on

employee training to start the right path early in the game, it would also help a candidate to navigate his/her tech journal by betting on the right road map.

8 CONCLUSION

In this project, we study the mining of US technologist data by combining US department of labor data along with web scraped H1b data set, after the introduction of word tagging to get skill set data, we clean and analyze the processed and joined data to later clustering the job skill data set. Since there is no clear technologist mining or analysis report online, our objective is to fill the blank page with knowledge of technologist's trench in the US studying time series and Geo data. Our study shows that on top of tech talent demand in US are on the rise, is cloud engineering and project management skills will be in huge demand, while SQL and system analysis are in decline. Java still grows strong as python has not yet replace its spot, and Software developers should not worry that data scientists or ML engineers will make their job obsolete.

Our study's limitation is the availability of lasted H1b data set along with non guidance on job title mapping table. In the first case: we are not able to do a deep time series analysis and prediction, in the latter, we do not have optimal clustering number for the unsupervised learning. We are looking forward to the future development of this topic with a better data set collection, the study of each topic can be its own project: Knowledge graph for each skill set, supply and demand of each job tile in the US hiring market, semi-supervised clustering with title mapping, and dynamic prediction for each skill and title.

BONUS MATERIAL: VISUALIZATION

We used Tableau¹¹ for our final visualization for this project. It contains the following interaction information:

- (1) A US map with H1b filtered by state and year.
- (2) Base salary distributing of most popular jobs filtered by state and year.
- (3) Time series of skill ranking over year.
- (4) Most frequent skills by 10 clusters.

REFERENCES

- [1] Enghin Atalay, Phai Phongthientham, and Sebastian Sotelo. Mapping Text to Occupational Characteristics; Mapping Job Titles to SOC Codes; Mapping Job Titles to OCC Codes. (????), 11.
- [2] Adil Bahaj, Safae Lhazmir, and Mounir Ghogho. 2022. KG-NSF: Knowledge Graph Completion with a Negative-Sample-Free Approach. (July 2022). <http://arxiv.org/abs/2207.14617> arXiv:2207.14617 [cs].
- [3] Aurélie Breidenbach, Caroline Mahlow, and Andreas Schreiber. 2021. Implicit Gender Bias in Computer Science – A Qualitative Study. (July 2021). <http://arxiv.org/abs/2107.01624> arXiv:2107.01624 [cs].
- [4] Marco Capó, Aritz Pérez, and Jose A. Lozano. 2018. An efficient K -means clustering algorithm for massive data. (Jan. 2018). <http://arxiv.org/abs/1801.02949> arXiv:1801.02949 [cs, stat].
- [5] Jens-Joris Decorte, Jeroen Van Haute, Thomas Demeester, and Chris Develder. 2021. JobBERT: Understanding Job Titles through Skills. (Sept. 2021). <http://arxiv.org/abs/2109.09605> arXiv:2109.09605 [cs].
- [6] Nicolas Gutierrez and Manuela Wiesinger-Widi. 2016. AUGURY: A time-series based application for the analysis and forecasting of system and network performance metrics. In *2016 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*. 351–358. DOI: <https://doi.org/10.1109/SYNASC.2016.062> arXiv:1607.08344 [cs].
- [7] Patrick Hochstenbach, Herbert Van de Sompel, Miel Vander Sande, Ruben Dedeker, and Ruben Verborgh. 2022. Event Notifications in Value-Adding Networks. (Aug. 2022). <http://arxiv.org/abs/2208.00665> arXiv:2208.00665 [cs].
- [8] Bin Ji, Shasha Li, Jie Yu, Jun Ma, and Huijun Liu. 2022. Win-Win Cooperation: Bundling Sequence and Span Models for Named Entity Recognition. (July 2022). <http://arxiv.org/abs/2207.03300> arXiv:2207.03300 [cs].
- [9] Daesoo Lee. 2022. Better Reasoning Behind Classification Predictions with BERT for Fake News Detection. (July 2022). <http://arxiv.org/abs/2207.11562> arXiv:2207.11562 [cs].
- [10] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. (Sept. 2013). <http://arxiv.org/abs/1301.3781> arXiv:1301.3781 [cs].
- [11] Sukavanam Nanjundan, Shreevishesh Sankaran, C. R. Arjun, and G. Paavai Anand. 2019. Identifying the number of clusters for K-Means: A hypersphere density based approach. (Dec. 2019). <http://arxiv.org/abs/1912.00643> arXiv:1912.00643 [cs, stat].
- [12] Zachary A. Pardos and Andrew Joo Hun Nam. 2018. A Map of Knowledge. (Nov. 2018). <http://arxiv.org/abs/1811.07974> arXiv:1811.07974 [cs].
- [13] Jože M. Rožanec, Elena Trajkova, Inna Novalija, Patrik Zajec, Klemen Kenda, Blaž Fortuna, and Dunja Mladenčić. 2022. Enriching Artificial Intelligence Explanations with Knowledge Fragments. (April 2022). <http://arxiv.org/abs/2204.05579> arXiv:2204.05579 [cs].
- [14] Nina Smirnova and Philipp Mayr. 2022. Evaluation of Embedding Models for Automatic Extraction and Classification of Acknowledged Entities in Scientific Documents. (June 2022). <http://arxiv.org/abs/2206.10939> arXiv:2206.10939 [cs].
- [15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. (Dec. 2017). <http://arxiv.org/abs/1706.03762> arXiv:1706.03762 [cs].
- [16] Yinyi Wang, Zitao Wang, Weijian Sun, and Wei Hu. 2022. Enhancing Document-level Relation Extraction by Entity Knowledge Injection. (July 2022). <http://arxiv.org/abs/2207.11433> arXiv:2207.11433 [cs].
- [17] Ziyang Wang, Wei Wei, Chenwei Xu, Jun Xu, and Xian-Ling Mao. 2022. Person-job fit estimation from candidate profile and related recruitment history with co-attention neural networks. (June 2022). <http://arxiv.org/abs/2206.09116> arXiv:2206.09116 [cs].
- [18] Natnael A. Wondimu, Cédric Buche, and Ubbo Visser. 2022. Interactive Machine Learning: A State of the Art Review. (July 2022). <http://arxiv.org/abs/2207.06196> arXiv:2207.06196 [cs].
- [19] Chuhui Xue, Jiaxing Huang, Shijian Lu, Changhu Wang, and Song Bai. 2022. Contextual Text Block Detection towards Scene Text Understanding. (July 2022). <http://arxiv.org/abs/2207.12955> arXiv:2207.12955 [cs].
- [20] Xingzhi Zhou and Nevin L. Zhang. 2022. Deep Clustering with Features from Self-Supervised Pretraining. (July 2022). <http://arxiv.org/abs/2207.13364> arXiv:2207.13364 [cs].

¹¹<https://public.tableau.com/app/profile/abulitibu.tuguluke/viz/H1BApplicationsv2/H1BJobs?publish=yes>