Mining the US Technologists Data

Abulitibu Tuguluke

(Group 5)
DHI Group Inc., CU Boulder

August 7, 2022

Outline

Before we start

- Why we choose to mine the technologist data? "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.
- What is a 'Technologist'? The simplest definition of a technologist is an expert in a particular field of technology.
- What is a '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.

Outline

SpaCy POS

	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
⋮	⋮	⋮	⋮	⋮	⋮

Algo

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.
⋮	⋮

Outline

U.S. Department of Labor data

Table: Deparrtment 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

U.S. Department of Labor data

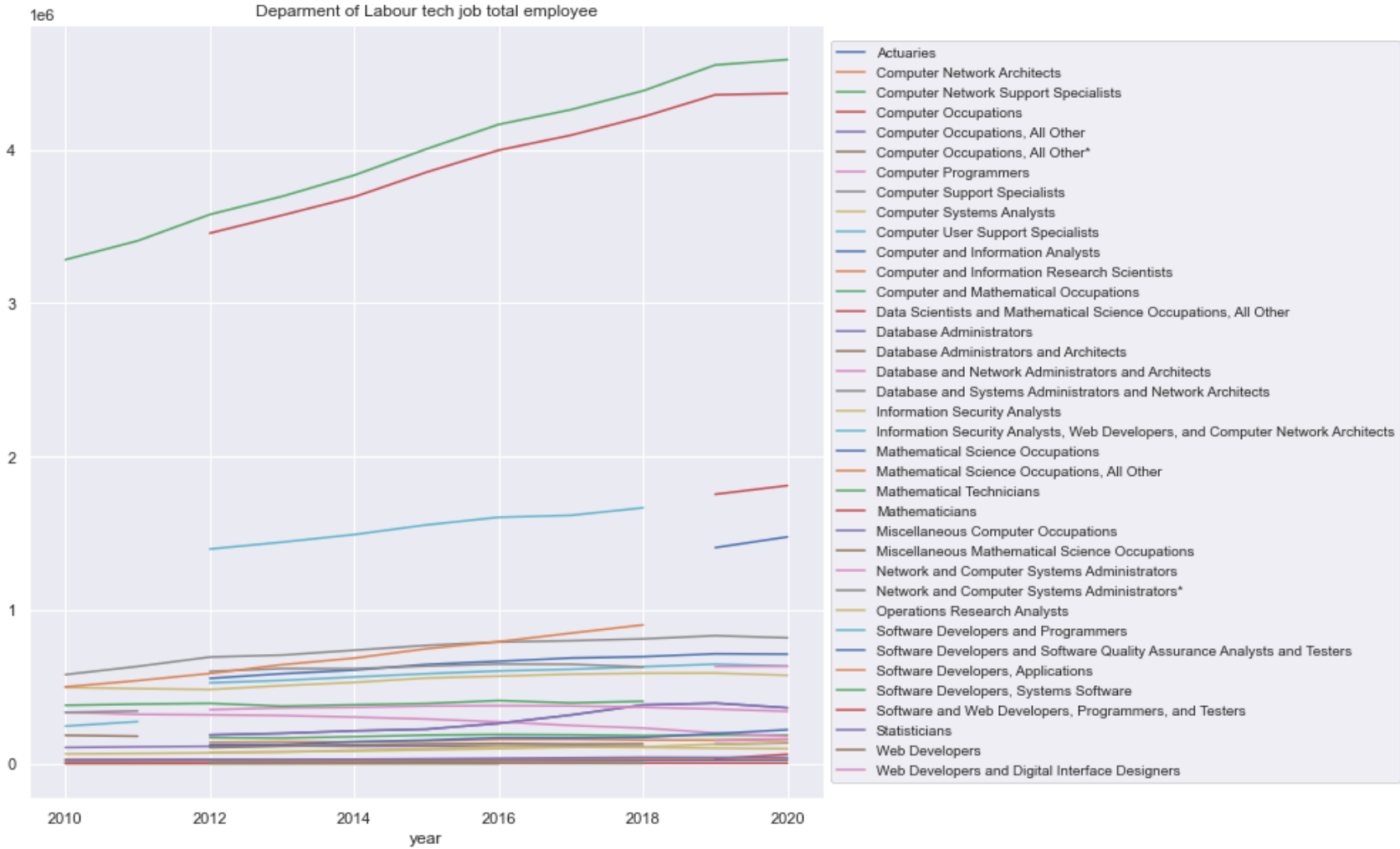


Figure: Deparment of Labour tech jobs number of employee.

U.S. Department of Labor data

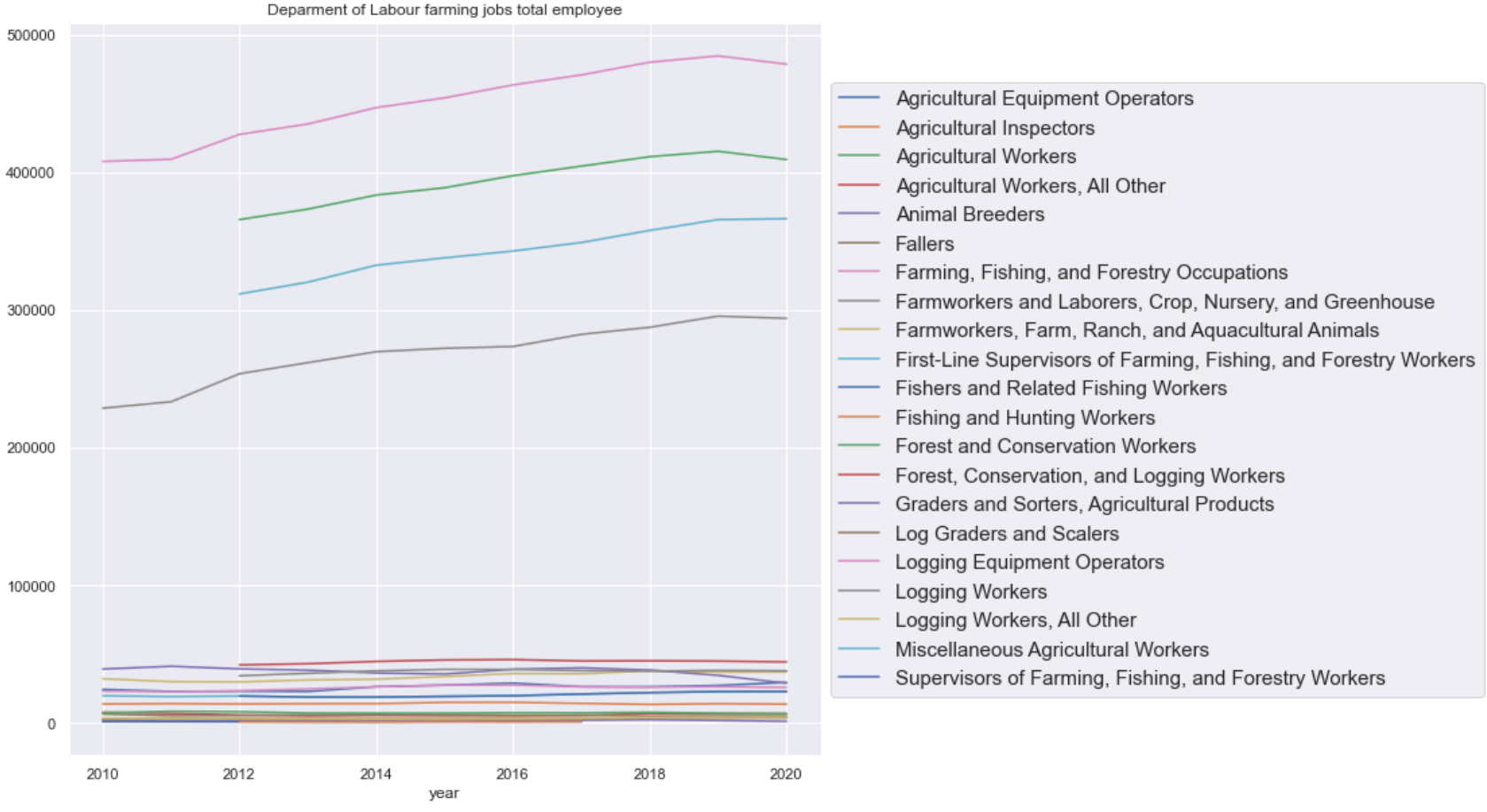


Figure: Deparment of Labour farming jobs number of employee.

U.S. Department of Labor data

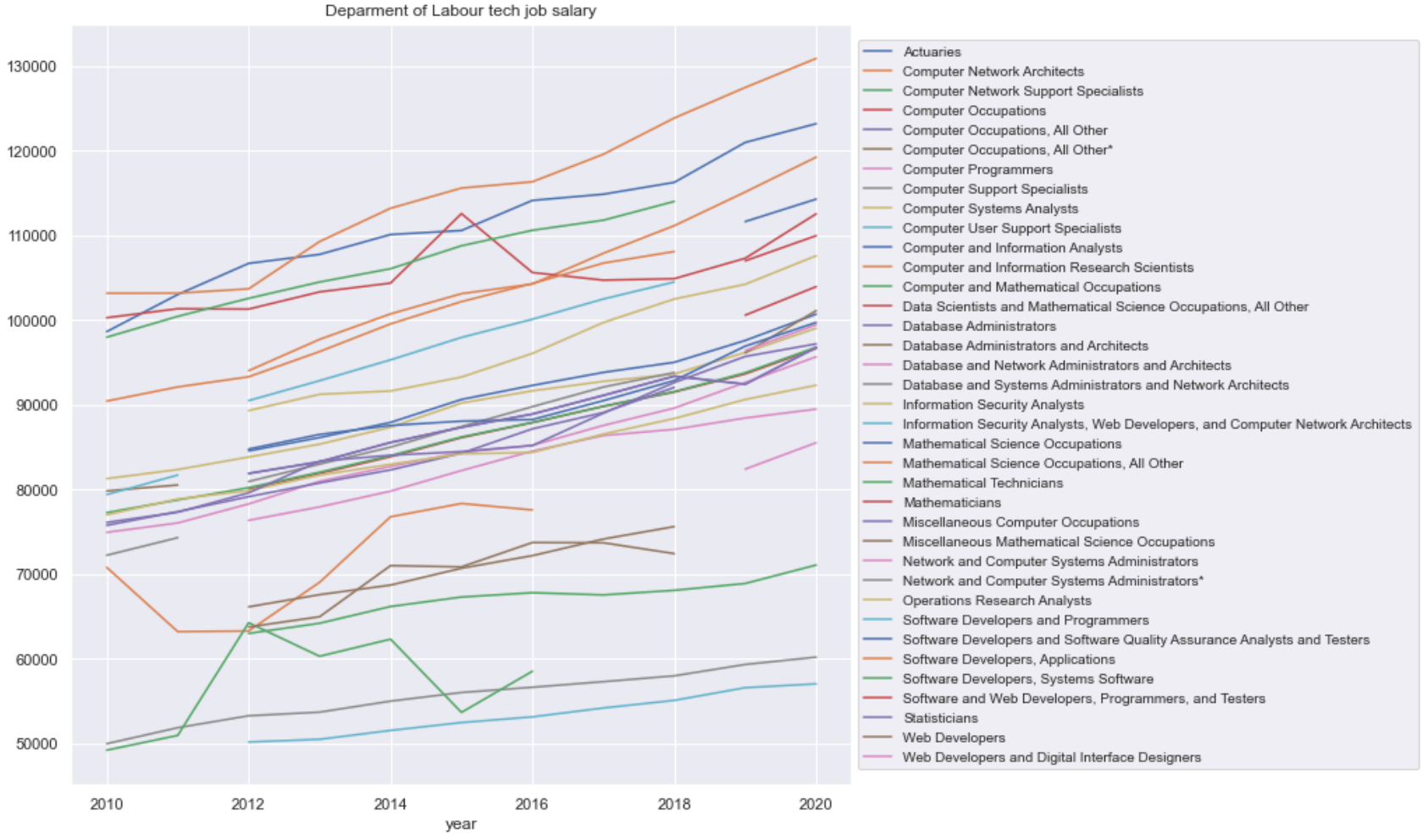


Figure: Deparment of Labour tech job salary.

H1b data

Table: 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

EDA

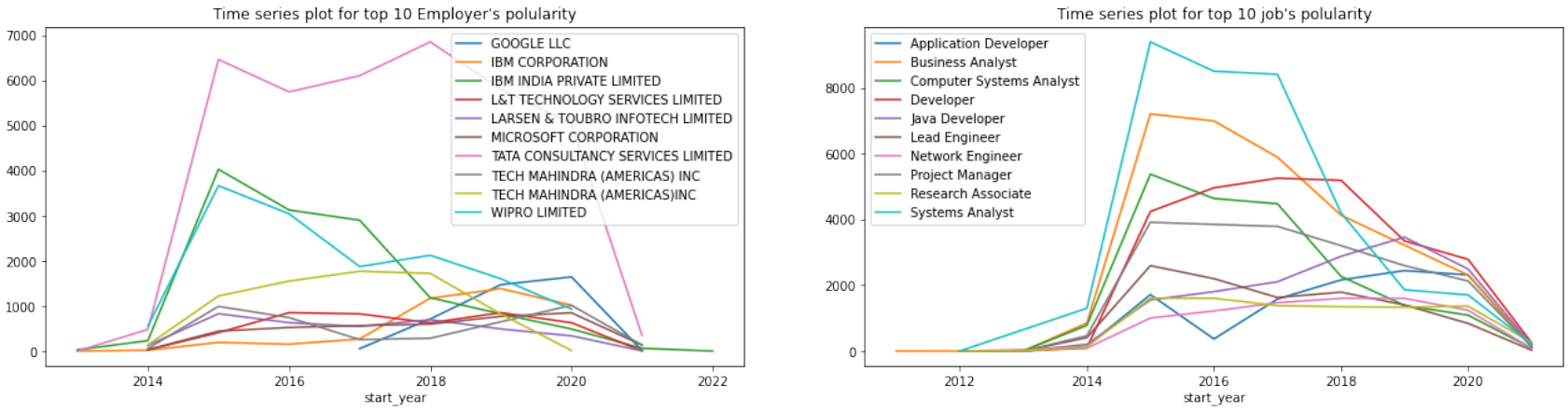
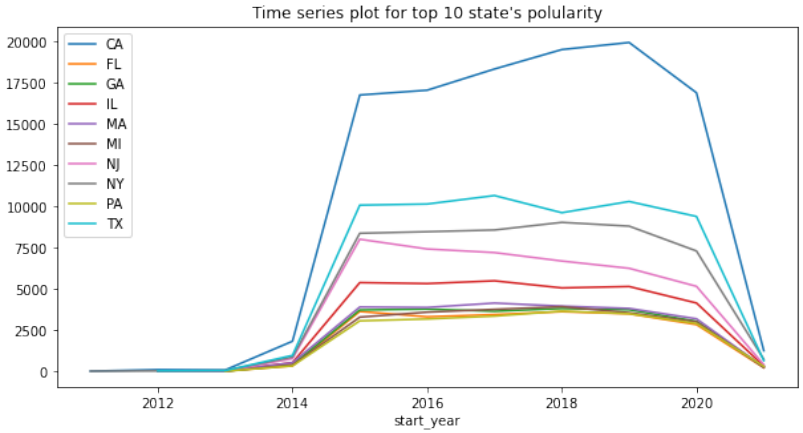


Figure: Top 10 employers and top 10 jobs.

EDA



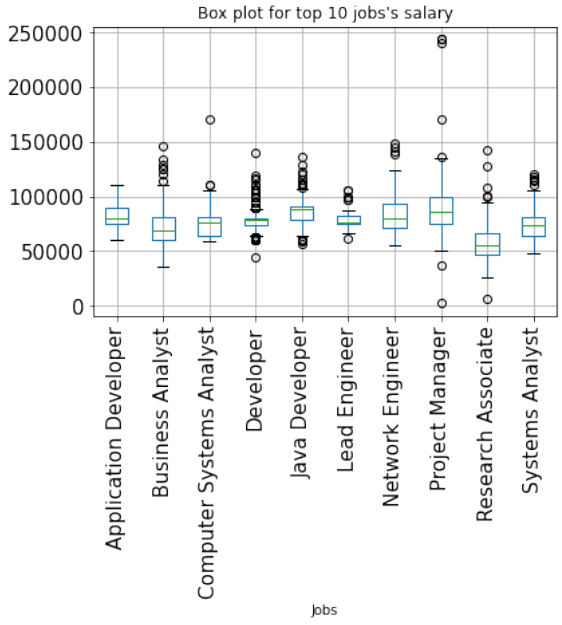


Figure: Top 10 states and top 10 jobs salary.

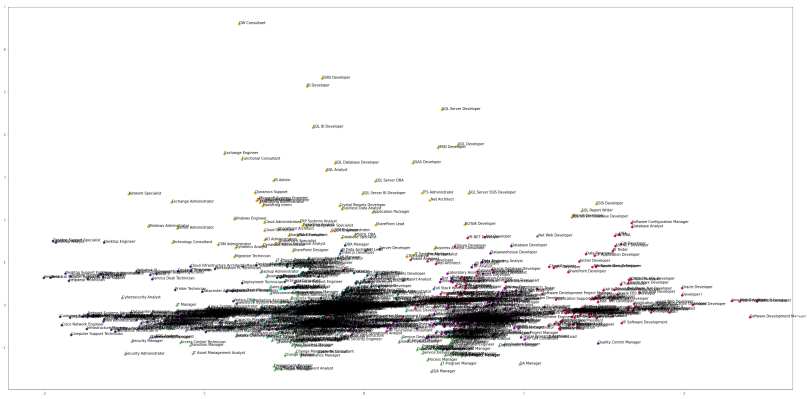
Outline

Data cleaning

Table: 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

K-Mean



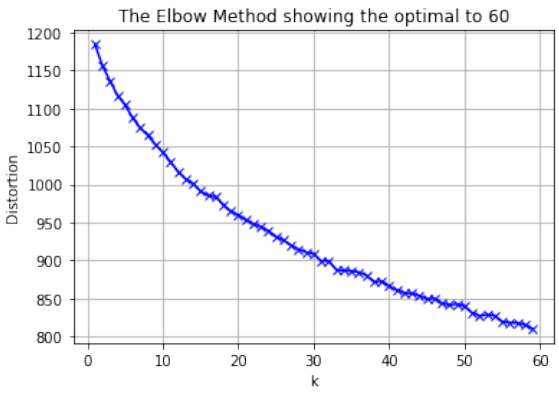


Figure: PCA + K-mean + Elbow.

Outline

Trending Skills

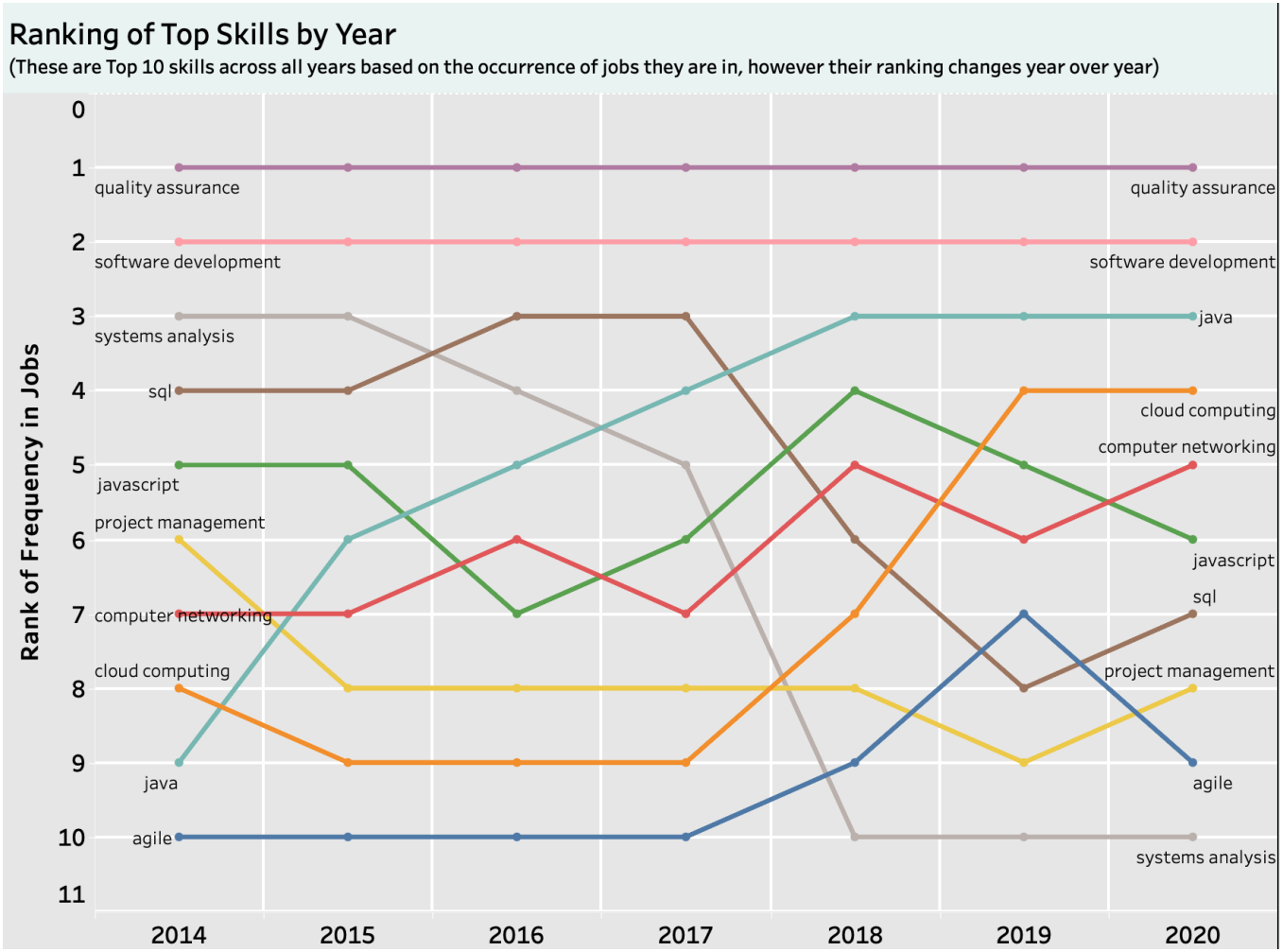


Figure: Skill 'trendiness' throughout the years

clustering

- 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 developmen 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.

Outline

Application

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.

Outline

Knowledge graph

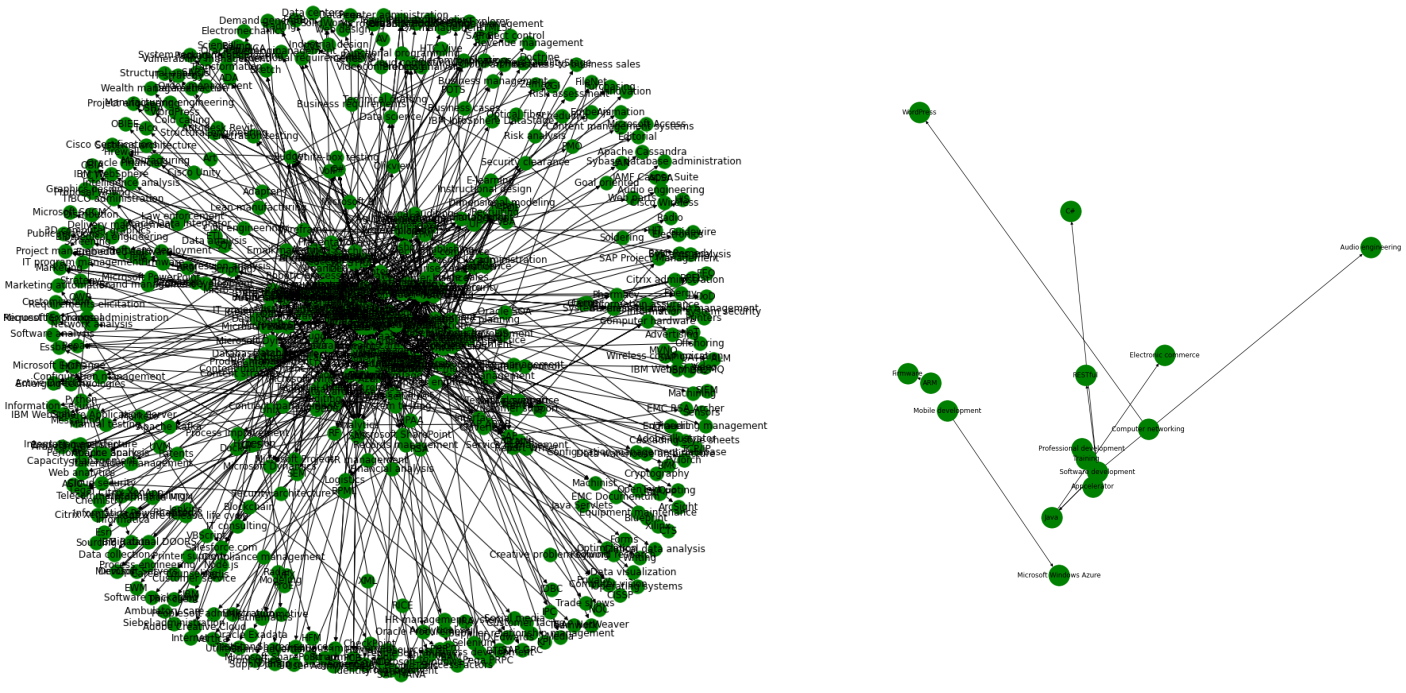


Figure: knowledge arrow

MORE DATA. FROM SUPPLY SIDE

Outline

What did we do and learn?

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.

Outline

Application

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

Questions?

Thank you