Mining the US Technologists Data

Abulitibu Tuguluke

DHI Group Inc., CU Boulder

July 3, 2022

Outline

Title

Mining the US Technologists Data

Outline

Team member

Group 5: Abulitibu Tuguluke

Outline

Description

There are many reports stating US is (or will) face a technology workers (Technologists). Our study 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 try to understand which technologist job are in increasing/decreasing demands, which skill sets are more popular/unpupular, through clustering, time and geo-location analysis. 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 skillset guidance, answers the questions such as whether foreign workers are getting paid the same amount as their US colleges.

Outline

Prior Work

There is already a skillsets collection been done on tons of job descriptions for technologist through data mining. We have 1432 job titles associate with perticular skill sets. Unfortunately, most of the online literatures are focused on salary analysis, which motivates job seeker, where we will do our own salary report as well.

Papers

- Curriculum Vitae Recommendation Based on Text Mining by Honorio Apaza Alanoca, Americo A. Rubin de Celis Vidal, Josimar Edinson Chire Saire

Outline

Datasets

- H1B data scraped from $https://h1bdata.info/index.php$
- Job skill set (Internal data)
- US Bureau of Labor Statistics
  $https://www.bls.gov/developers/api_python.htm$
- Data mined and stored on my machine

Data Description for scraped for H1b techonlogist data

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

Data Description for US Bureau of Labour statistics[1]

	occ_code	occ_title	group	tot_emp	emp_prse	h_mean	a_mean	mean_prse	h_pct10	h_pct25	h_median	h_pct75	h_pct90	a_pct10	a_pct25	a_median	a_pct75	a_pct90	annual	hourly	year	occ_group	area	area_title	area_type	naics	naics_title	i_group	own_code	o_group	jobs_1000	loc_quotient	pct_total	prim_state
count	13208	13208	23	13208.0	13208.0	13208	13208	13208.000000	13208	13208	13208	13208	13208	13208	13208	13208	13208	13208	795	58	13208.000000	9731	2658.0	2658	2658.0	2658.0	2658	2658	2658.0	2658	0.0	0.0	0.0	1329
unique	1507	1272	2	8454.0	226.0	3806	5951	NaN	2224	2858	3523	4241	4626	3707	4729	5615	6447	6756	1	1	NaN	5	NaN	1	NaN	NaN	1	1	NaN	5	NaN	NaN	NaN	1
top	39-7010	Tour and Travel Guides	major	6940.0	0.5	*	*	NaN	*	*	*	*	*	*	*	#	#	#	True	True	NaN	detailed	NaN	U.S.	NaN	NaN	Cross-industry	cross-industry	NaN	detailed	NaN	NaN	NaN	US
freq	13	22	22	9.0	480.0	795	58	NaN	795	795	795	795	795	58	58	75	172	571	795	58	NaN	5717	NaN	2658	NaN	NaN	2658	2658	NaN	1578	NaN	NaN	NaN	1329
mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.125901	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2015.652332	NaN	99.0	NaN	1.0	0.0	NaN	NaN	1235.0	NaN	NaN	NaN	NaN	NaN
std	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.118357	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2.762959	NaN	0.0	NaN	0.0	0.0	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN
min	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.100000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2011.000000	NaN	99.0	NaN	1.0	0.0	NaN	NaN	1235.0	NaN	NaN	NaN	NaN	NaN
25%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.500000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2013.000000	NaN	99.0	NaN	1.0	0.0	NaN	NaN	1235.0	NaN	NaN	NaN	NaN	NaN
50%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.800000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2016.000000	NaN	99.0	NaN	1.0	0.0	NaN	NaN	1235.0	NaN	NaN	NaN	NaN	NaN
75%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.400000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2018.000000	NaN	99.0	NaN	1.0	0.0	NaN	NaN	1235.0	NaN	NaN	NaN	NaN	NaN
max	NaN	NaN	NaN	NaN	NaN	NaN	NaN	25.200000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2020.000000	NaN	99.0	NaN	1.0	0.0	NaN	NaN	1235.0	NaN	NaN	NaN	NaN	NaN

[1] Internal jobs data will not be shared at any stage/part of this project.

Outline

Proposed work

- Data scraping: Screaping HTML data from the web and store them in seperate json files locally
- Data cleaning: Probably the most important part of the project
- Data preprocessing:
- Data integration: Integrate internal data with mined H1b data
- Data mining and analysis

Outline

List of tool(s)sets

- Beautifulsoup for mining
- Python library for all the computing and analysis: notably NumPy, Pandas, SK-learn,SpaCy, etc.
- JSON for data store

Outline

Evaluation

There are many ways we can evaluate the results. Clustering Performance Evaluation Metrics like Silhouette coefficient is a good start when applying clustering algorithm, for time series, all the error evaluations. Besides,comparing time series analysis data within each job we can compute the increaing or decreasing sides of each job along with correlations. Location analysis on which job are located where with repect to other jobs and locations. Finally ,we are looking forward to apply all the relavent techniques we are currently learning in the class[2].

[2] More methods will be updated through work.