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

Note:

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 NLP 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[1].

---

[1]More methods will be updated through work.