Mining the US Technologist data

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

1 PROBLEM STATEMENT/MOTIVATION

(like homework 1, what knowledge and how would you apply that knowledge, what is interesting that you hope to find) There are many reports saying 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 setsa are more popular/unpupular, through 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.

2 LITERATURE SURVEY

(previous work) describe and cite.

3 PROPOSED WORK

E.g., what do you need to do for data collection, preprocessing (cleaning integrating, transforming, etc.), process for derived data, design, evaluation. Describe how it is different than what has been done previously from your literature survey (or if replicating).

- Data scraping: Screaping HTML data from the web and store them in seperate json file
- 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

4 DATA SET

(make sure you have the data set!). Provide URL and details about the data set (similar to homework 1, chapter 2, etc.)

- H1B data scraped from I scraped from open data online.[1] The data set is mainly from the United States Department of Labor (DOL) on how many H1-B petitions were filed and approved, with detailed information such as Employer, the title of the job, city/state locations, along with base salary, and submission and acceptance dates.
- Skill set (Internal data)
- US Bureau of Labor Statistics[2]
- Data mined and stored on my machine

5 EVALUATION METHODS

E.g., metrics, existing solutions, …
here are many ways we can evaluate the results. Comparing time series analysis data within each job we can compute the increaing or decreasing sides of each job. Location analysis on which job are located where with repect to other jobs and locations.

6 TOOLS

- Beautifulsoup for mining
- Python library for all the computing and analysis
- JSON for data store

7 MILESTONES

What you plan to have done by when
The Following table is my work plan for carrying out the project.

Date	Work Plan
Week 1	Submit the Research Proposal.
Week 1	Submit the Research Proposal.
Week 1	Submit the Research Proposal.
Week 1	Submit the Research Proposal.
Week 1	Submit the Research Proposal.
Week 1	Submit the Research Proposal.
Week 1	Submit the Research Proposal.
Week 1	Submit the Research Proposal.
Week 1	Submit the Research Proposal.
Week 1	Submit the Research Proposal.

8 PART 1 PROPOSAL FEEDBACKS FROM CLASSMATES

Pratik Prasanna Raghavendra

Interesting topic, Abulitibu. Just curious, what would outlier data look like in your case? Also, how do you detect and eliminate them, since for example, really high income may simply be an outlier for individuals, but an important point for other employers?

[1] *https://h1bdata.info/index.php*.

[2] *https://www.bls.gov/developers/api_python.htm*

CSCI 5502, Boulder, University of Colorado
© 2022 ACM. 978-x-xxxx-xxxx-x/YY/MM…$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

Rohit Kharat

I find this a unique topic and interesting to analyze in today's scenario. I liked that you are using web scraping, which I enjoy; using the BeautifulSoup library for web mining is fun. I am not familiar with the SpaCy library, something you can shed a light on. I also liked the idea of using location information for your analysis. I would recommend geolocation API something which I used earlier in my projects. I would also recommend doing feature reduction or getting the important features in your proposed work for focusing on only the crucial attributes.

Abhinav Gupta

This looks like an unusual topic! Curious to see what your data cleaning and preprocessing would look like. As your job skill data is internal, I don't know how it looks like, but I have a couple of questions. First question is that technologies are evolving and higher demand will always be for the new tech. As that tech, perhaps, is not even in the market yet, how would you predict the demand? Second question is if you would be considering countries and degrees as visa approval ratio varies highly on these factors. Final question is that as per uscis.gov - "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", so wouldn't your dataset be corrupted too?

Varun Manjunath

An interesting topic. But is often the case that H1b workers are paid less as H1b is a nonimmigrant visa and is temporary in nature. So the point where you perform analysis on the dataset to check whether H1b workers are paid less is not very useful. Rather a more interesting trend would be to check which location hires the most H1b candidates. Dataset description with a bit more depth would be great in your case. Also, are you planning an inner join query to combine the datasets? Would suggest you remove certain outliers in the combined dataset before performing K-Means clustering.

Danielle Aras

This is a very interesting and relevant topic. I agree that data cleaning is an important part of your project as you are merging different data sources with different features. I would be curious to know more details about the data sets like the size and date range. Technology is a fast moving field so the most current data will be the most useful for job seekers; looking at which jobs are trending up or down is a great idea.

I agree with other posters that the H1B visa data will introduce some challenges to your analysis. Given the trend of employers looking to take advantage of the system to under-pay workers, the popularity of H1B workers in a particular job may not actually correspond with a genuine demand that is not met by domestic workers. An interesting "spin-off" question would be creating an algorithm to detect companies abusing the system based on the availability of candidates in their area and H1B salaries.