**Indeed Job Data Analysis with Data Scrapping**

**Networked Data Modeling and Processing (INFO 5717 - 001)**

(Final Project)

**Team Members**

Venkata Gnaneswari Regati

Email: GnaneswaruRegati@my.unt.edu

Mounika Vunnam

Email: Mounikavunnam@my.unt.edu

Sri Lakshmi Tirupatamma Nallapati

Email: srilakshmiNallapati@my.unt.edu

University of North Texas

Date: July 27, 2023

Table of Contents

[Analysis of the website 3](#_Toc141473378)

[Questions 4](#_Toc141473379)

[Major entities & attributes 4](#_Toc141473380)

[Entity Relationship Diagram 5](#_Toc141473381)

[The MySQL create table statements to create tables in MySQL 7](#_Toc141473382)

[The Python program to establish the database tables in MySQL; 8](#_Toc141473383)

[Add information from the website to your db 10](#_Toc141473384)

[Develop 4 queries and retrieve results from your mysql db 10](#_Toc141473385)

[Python program to visualize the query results: 12](#_Toc141473386)

[Conclusions 15](#_Toc141473387)

### 

### Analysis of the website

We have performed an analysis on job portal website named indeed.com. we can see the sitemap in the below screenshot. Websites have three major sections like finding the jobs, reviews on the companies and Salaries. Finding jobs shows that different jobs available with reference to type of job and locations. Next, we can see Company reviews and indeed have presented the Reviews section in terms of industries, popular companies and comparing different companies in the industry. Lastly, we can observe salaries. Which is the most important part. They have presented in it by searching companies with salaries, finding salaries based on industry, finding the careers and browsing salaries in top companies.

A diagram of a company

Description automatically generated

We have done the web scrapping on indeed.com using Python library named BeautifulSoup. We next utilized its features to extract data from the html content. The major challenge we faced is the website data is dynamic in nature hence we could not able to collect data from the url. Hence we found an approach to collect it. We have downloaded the HTML content from browser into the text files. Hence we pulled those files into the python environment and performed web scrapping using BeautifulSoup object.

A screenshot of a computer

Description automatically generated

### Questions

1. What are the top 10 job titles with the highest average salaries?
2. Which companies have the highest ratings for remote job opportunities?
3. Which locations have the highest number of job listings in the IT industry?
4. What is the average salary range for entry-level positions in New York City?
5. What is the trending job types in the healthcare industry?

### Major entities & attributes

After scrapping the data from the various website we brainstormed and identified five tables required for the analysis. So, we have considered major entities and attributes as

1. JBTTL\_AVGSAL
   1. ID
   2. JB\_TITL
   3. AVG\_SAL
2. COMP\_RATNGS
   1. ID
   2. COMPANY
   3. RATINGS
3. NYC\_AVG\_SAL
   1. ID
   2. TITLE
   3. AVG\_SAL
4. COMP\_LOC
   1. ID
   2. COMPANY
   3. LOCATION
5. TRNDNG\_COMP\_HELTH
   1. ID
   2. TITLE
   3. COMPANY

We have seen that identification column is required and we have added primary key constraint on it for all the tables.

### Entity Relationship Diagram

Next, we have build Entity relationship diagrams for the tables. We can see below screenshots for ER diagrams for all the tables. ER Diagrams provide visual understanding about tables and their relationships with another tables. In this project, since our data set do not provide any foreign key relationships hence we cannot see relations in ER diagram

A diagram of a computer

Description automatically generatedA diagram of a computer

Description automatically generated A diagram of a company

Description automatically generated

Once, we created ER diagrams, we have started creating tables in the database.

### The MySQL create table statements to create tables in MySQL

In this section, we have created database tables we have designed. We can see the screenshots below. We created tables using Python and its libraries on MySQL. Mysql-connection-python library has been installed and used in this project.

We can see below code showing DB connection to the locally installed MySQL database from Python client. Later, we have created new database named INFO5717 for this project. We need to run USE database\_name command to execute queries. Next, we need to create connection and cursor objects to connect to database.

A screenshot of a computer

Description automatically generated

Now, we can see below screenshot showing creating first table in database. We can understood various required datatypes for the table and created it from Python.

Next, we can see processing of data to insert. Here we are translating the data from dictionary to list of tuples. Next, we can see executemany function to insert the data into tables and we commit them using commit function.

A screenshot of a computer code

Description automatically generated

### The Python program to establish the database tables in MySQL;

In this step, we can see COMP\_RATNGS table is created and getting data inserted into it.

A screenshot of a computer program

Description automatically generated

In this step, we can see COMP\_LOC table is created and getting data inserted into it.

A screenshot of a computer program

Description automatically generated

In this step, we can see NYC\_AVG\_SAL table is created and getting data inserted into it.

A screenshot of a computer code

Description automatically generated

In this step, we can see TRNDNG\_COMP\_HELTH table is created and getting data inserted into it.

A screenshot of a computer

Description automatically generated

### Add information from the website to your db

We have added information to the db using Python web scrapping with the help of BeautifulSoup library. We have used html parser as well.

### Develop 4 queries and retrieve results from your mysql db

Once we have loaded all the tables we created, then we performed queries on the database tables. We can see the data is read from database using select queries on each tables. We have next used the retrieved data for data visualization. We have converted the results to list and dataframes for further steps of visualization.

A screenshot of a computer code

Description automatically generated A screenshot of a computer program

Description automatically generated A screenshot of a computer

Description automatically generated

### Python program to visualize the query results:

We have used Matplotlib and Seaborn libraries from Python. With the help of them we plotted all the questions and found the answers to them. We can see answers explained below. We have also added labels and title to the chart for clear understanding and reporting.

Question 1

Here are the answers for the question 1 which we formulated in the beginning.

We have plotted line chart between Job title and Average salary on X-axis and Y-axis respectively. We have seen that Senior Software Engineer title is having highest average salary above all other roles in the IT industry. We can see IT support have least salary.

A graph with a line

Description automatically generated

Question 2

We can see line chart for the remote jobs ratings. In the plot, we can observe that Yardi Systems have very good ratings of 4.2 which highest among all other companies. Most of the companies have ratings around 3, which means average in terms ratings. Scale AI have very poor ratings of 2.1

A graph with lines and text

Description automatically generated

Question 3

We can see the plot of places with Highest jobs in IT industry. We have plotted histogram for the analysis. We can see multiple cities in X-axis and frequencies on Y-axis. Austin, TX is the topmost city for IT jobs in the state of Texas. While other cities have at least 1 job of IT industry

A green graph with black text

Description automatically generated

Question 4

We can see the entry level job analysis in the New York city. We can see multiple job titles from Helper to Doctor. The salaries exists larger difference. We can see Pediatric Ophthalmologist have highest salary of $ 250000 per year. While Associate(entry level) have least.

A graph with blue squares

Description automatically generated

Question 5

Here, we can see the top trending jobs in the Healthcare industry. we can see MNC companies like CVS Health, Nucleus Healthcare. We can see location attribute as well. Here, we see that remote jobs are also visible.

A screenshot of a computer

Description automatically generated

### Conclusions

From our analysis, we can see the results are very nearer to the real time. For example, we can understand doctors’ salary is higher when compared to other jobs. Austin, TX have resulted as huge number of vacancies in the field of IT. We have also, which company is having great ratings in remote work culture. However, we can see that data size is small, we have faced challenges in the data scrapping hence could not collect more data. But we are planning for further analysis and collecting more data to perform clear analysis.