26 November 2023

Assignment - Job Market Analysis

# Project AIM

Analysis of job market data to explore the open positions related to the jobs in the field of data or machine learning in a particular region. While it is not compulsory to do so, you could use web scraping techniques introduced in this course to collect data from these websites. Also, note that you are not limited to the sites mentioned above. Identity one job that you aspire to transition into and identify the skills required for the job.

## Objectives & Assignment Description:

### Collect and use job market data to explore data related positions or machine learning related positions in a particular region (e.g., US, India).

### You will have to use appropriate keywords (for e.g., “data scientist”, “data engineer”, “ML engineer”) to extract information on the positions available on websites of your choice.

# Procedures:

## Data Collection Method

The data collection involves collecting the data from any one of the websites like Indeed, Glassdoor, Naukri, Monster. We also must use one keyword to extract the required jobs from website. For this assignment I choose the following.

## Website – Indeed

## Keyword – Machine Learning

### Robot text file from indeed does not block /jobs and /viewjob URLs for user agent \*. We also inspected the page to find corresponding html tags, classes and other attributes to extract the data.

### **Data discovery**: I have selected the indeed website for data extraction. Challenges faced are the following. The website uses Cloudflare for protection against DDoS mitigation, a Web Application Firewall, API protection, etc. But this also make sure we do not get the data by using simple requests library. I have also tries using special headers. But this was also not successful. Hence, I have used selenium library to open the webpage using Chrome driver from selenium library. Each webpage opening takes around little more than 3 seconds to complete. Sometimes more than 3 seconds I have also used beautifulsoup library as I found it easier to navigate the complex html page. I also found ways to do pagination and access pages of data. This makes up the discovering phase of the Data collection. This solution arises another issue of time taken for extraction of data. I must use python library concurrent futures (<https://docs.python.org/3/library/concurrent.futures.html>) to achieve parallel execution of certain function to extract more data in less amount of time. For getting details about each job, we need to extract /viewjob page for each job entry in jobs page (index page with list of jobs). Pagination should also be done using start and limit parameter to extract the data. We also had filters like location, programming language, job type for extracting information in detail. The website shows a bug where we are not able to access the list beyond 1000. This issue is solved by using more filter to reduce the total result set.

### **Data structuring**: For each job we extracted the details like, title, time of posting, company, location, metadata (includes job type, salary range and other special details), partial job description and link to full job details page. From individual page, we have extracted the programming languages and full description of the job.

### **Data Cleaning**: Here we encoded the data for job type, programming languages into Boolean form where we have True (1) for presence and False (0) for absence. We have also removed the duplicates and aggregated the columns for the duplicate assuming same jobs have same title, company, date posted, company rating, full description. For presence column it was a simple Boolean and (sum) and for metadata it was concatenation. We also had to extract data again in case we were getting invalid description.

### **Enriching Data:** Enriching the data is done by using the using nltk to extract major programming language tokens. We also did the same for job type from metadata. We have full-time, contract, fresher, internship, part-time, temporary, permanent, freelance as job types and these categories can be overlapping. We also extracted salary from metadata and separate it into maximum, minimum and normalized salary along with intended. Here normalized salary means salary calculated for entire year. We also had to create category for jobs as job title can be very similar but different when be compare. We have identified major clusters using K-means for the title and used this as a reference to create our major category of jobs.

|  |
| --- |
|  |

## Market Data Visualization

### Titles

|  |  |
| --- | --- |
|  |  |

### We can see that top categories are engineer, developer, analyst, software engineer, data scientist, manager, technology related, AI, Data science, consult, machine learning etc. In comparison to salary, other than positions like management and leadership, most of the jobs are well distributed in the category. But we can see some jobs in data scientist, machine learning and artificial intelligence as the highest salary job.

### Locations where jobs are offered.

|  |  |
| --- | --- |
|  | A screenshot of a graph  Description automatically generated |

### We can see most jobs have demand in Bangalore, Pune, Mumbai and Hyderabad. We can also see salary also follows the same order.

### Job Types

|  |  |
| --- | --- |
|  |  |

Note that pie chart only contains the top combination of the job types. We can see that full time employee is having highest demand. Part time is only having 10th position in demand.

### Company Offers.

|  |  |
| --- | --- |
|  | In the dataset, a significant number of companies are observed to have a limited job offering, often less than 10 positions. This trend persists as we narrow down our focus to the top 100 companies. It is noteworthy that there is a total of 1549 companies in the entire dataset out of which thousands of companies have only very few openings.  Furthermore, the salary distribution across different companies indicates a variation in compensation. Notably, only a few companies offer the highest-paying positions, and these tend to be specific roles with limited availability. |
| A screen shot of a computer screen  Description automatically generated | A screen shot of a graph  Description automatically generated |

### Skills

### A graph of a number of data Description automatically generated with medium confidence

### The top demanded skills are Python, SQL, Java, R, JavaScript, C, Go, React etc.

### Rating

### A table with text and numbers Description automatically generated

### We can also see that the Turing, CoinDCX, Curl Analytics, Anlage Infotech, TAO The Automation Office, SymphonyAI, Scienaptic Systems, Infiniti Research, Digital Green, Yara, Sophos, Salesforce, Google, Elixir Consulting, EXCELUS LEARNING SOLUTIONS PVT LTD, Bain & Company, Adobe, SAP, NVIDIA, NEC Software Solutions and Microsoft is having the highest ratings from other users. This might show the trust these company have with the public.

## My Ideal Job

### My ideal job has following preferences.

#### High Salary

#### Good Company - High rating

#### Better Location

#### Scope for improvement.

#### Skills – Python, R, SQL, Go, Machine learning, etc.

I will prefer the Bengaluru as the preferred location as this location shows highest demand in job and highest variance in the salary. This indicates the potential for having highest promotions and salary hikes which shows scope of improvement in the job position. The same location also has relatively cool climate and I personally prefer this place as my relatives are also in same city. I will conclude that the ideal job is Software Engineer, Machine Learning, Google Assistant (ID:3958) from Google. But If I ony consider the jobs for which salary is given, then I will choose Data Science - AI/ML with Big data Architect (ID: 1097) from Anlage Infotech company.