LinkedIn Job Analysis

SANDIP SHAW
24070243048
ARITRIKA DUTTA
24070243008
DEBJYOTI SENGUPTA
24070243013

1.Summary

This project involves analysing job listings scraped from LinkedIn by combining and refining multiple sources to uncover patterns in job roles, companies, locations, and required skills. The cleaned and enriched dataset is stored in MongoDB to enable efficient querying and support deeper insights through further analysis or visualization.

2.Objectives

- To integrate and clean job listing data from multiple sources.
- To identify trends in job titles, companies, locations, and required skills.
- To store the final processed dataset in MongoDB for efficient querying and future analytics or visualization.

3.About the Dataset

The dataset used in this project was sourced from Kaggle and contains job listings scraped from LinkedIn. It is a large-scale dataset with over 1.3 million job records and is spread across three CSV files.

Dataset Files

1. linkedin_job_postings.csv
   Contains metadata about each job posting, such as title, company, location, and timestamps.
2. job_summary.csv
   Includes textual summaries or descriptions of each job posting.
3. job_skills.csv
   Lists the skills extracted or mentioned in each job listing.

Common Key Column

All three datasets are linked using the common field: job_link (unique identifier for each job posting).

Features / Columns Overview

Column Name	Description
job_link	Unique URL of the job posting
last_processed_time	Timestamp of the last time the job was processed
got_summary	Indicates if the job summary was successfully fetched (t or f)
got_ner	Indicates if Named Entity Recognition (NER) was successfully applied (t or f)
is_being_worked	Indicates whether the job is currently under processing
job_title	Title of the job position
company	Name of the hiring company
job_location	Location where the job is based
first_seen	Date when the job was first discovered or posted
search_city	City used as a search filter to find the job
search_country	Country in which the job was searched
job_level	Seniority level of the job (e.g., Entry, Mid, Senior)
job_type	Type of job (e.g., Full-time, Internship)
job_summary	Detailed description or responsibilities of the job
job_skills	Key skills required for the job

4.Tools and Technologies Used

To successfully carry out the analysis of LinkedIn job listings, the following tools and technologies were utilized:

1. Programming Language

- Python: Used for data manipulation, analysis, and visualization due to its wide range of powerful libraries and simplicity in handling large datasets.

2. Python Libraries

- pandas: For data cleaning, transformation, and manipulation of tabular data.
- matplotlib & seaborn: Used to create visualizations like bar plots and histograms to better understand trends in job titles, companies, and locations.
- wordcloud: Used to generate word clouds from job summaries to visualize the most frequent terms in job descriptions.
- re (Regular Expressions): For text processing and cleaning job summary and skills fields.

3. Database

- MongoDB: A NoSQL document-oriented database used to store the final cleaned and enriched dataset. MongoDB allows efficient storage, retrieval, and querying of large volumes of unstructured or semi-structured job data.4. Development Environment
- Jupyter Notebook: An interactive environment that was used for writing code, running analysis, visualizing results, and documenting the workflow in one place.

5. Data Source

- Kaggle: The datasets used in this project were obtained from Kaggle and contain over 1.3 million job postings scraped from LinkedIn.

5.Project Workflow

1. Data Loading & Merging
   Imported three CSV files and merged them using the common job_link column to create a unified dataset.
2. Data Cleaning
   Removed missing values, standardized columns, and ensured consistency across job titles, skills, and summaries.
3. MongoDB Integration
   Connected to MongoDB and inserted the cleaned dataset into a collection for efficient querying and future use.
4. Exploratory Data Analysis (EDA)
   Analyzed job titles, companies, locations, and skills to identify key trends.
5. Visualization
   Created bar plots and a word cloud to visualize the most common roles, companies, and keywords in job descriptions.

6.MongoDB Integration

To efficiently store and manage the large job listings dataset, MongoDB—a NoSQL document-oriented database—was used. The cleaned and merged dataset was inserted into a MongoDB collection using Python.

MongoDB was chosen for its scalability and flexibility in handling semi-structured data. This setup enables fast querying and paves the way for future analytics or dashboard integration.

Key steps:

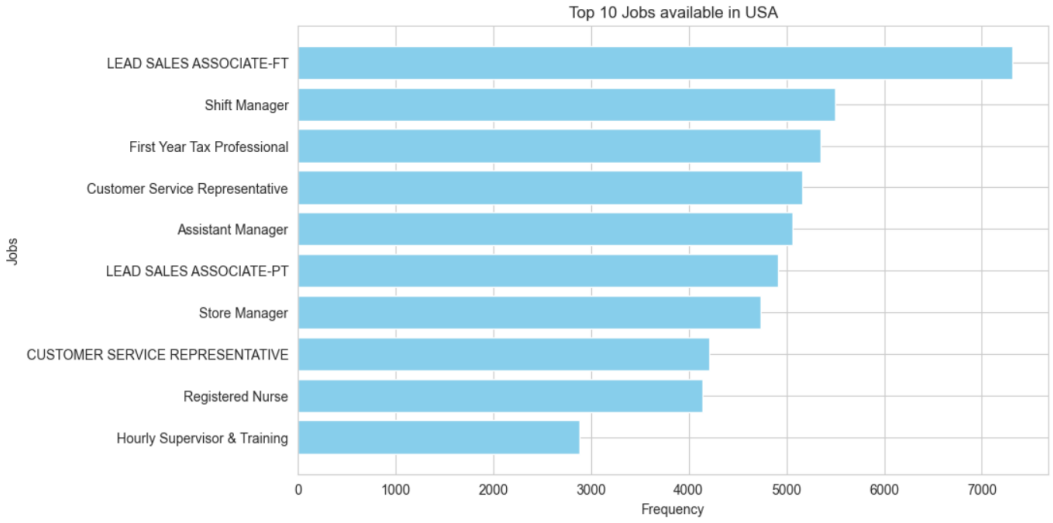
- Connected to MongoDB using pymongo
- Inserted records as documents into a collection
- Verified successful storage and tested sample queries

7.Exploratory Data Analysis (EDA)

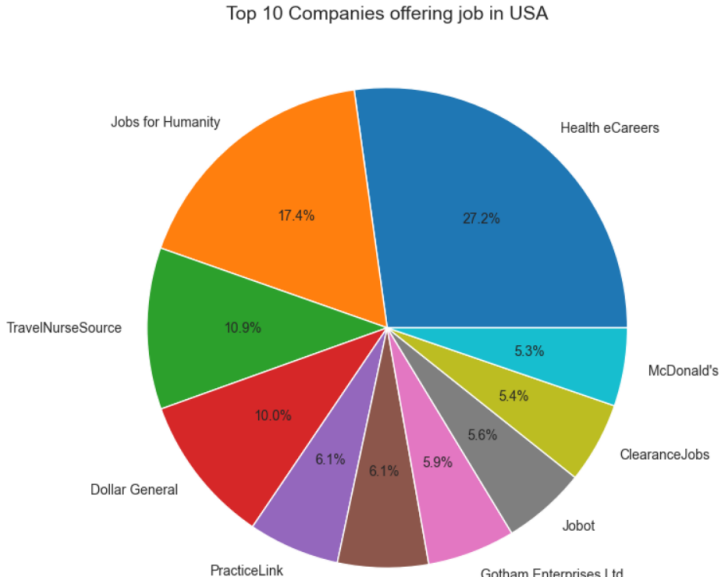
EDA was performed to uncover trends and patterns in the job market data. The goal was to identify the most common job titles, hiring companies, popular job locations, and in-demand skills.

Key insights explored:

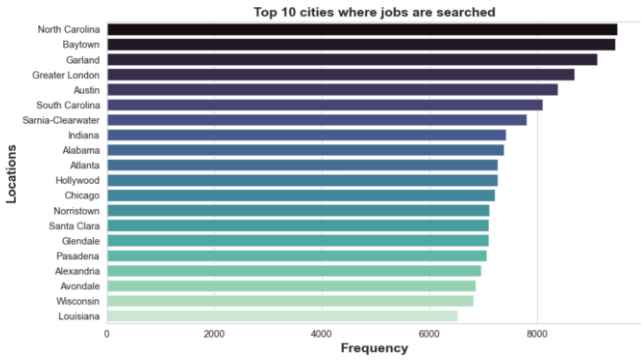
- Top Job Titles: Identified the most frequently posted job roles.

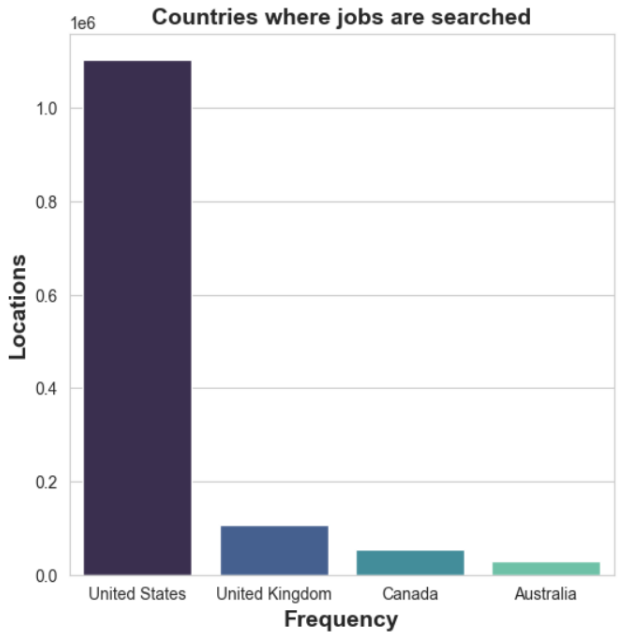


- Hiring Companies: Analyzed which companies posted the most jobs.

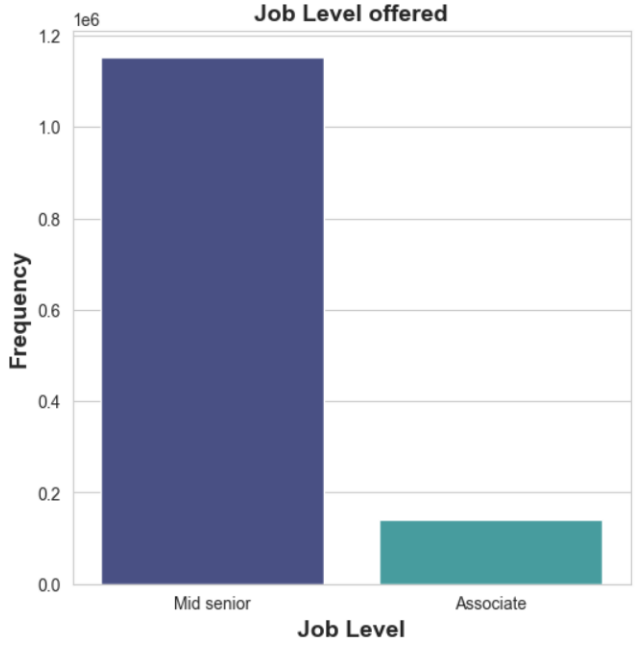


- Job Locations: Found the cities and countries with the highest job availability.

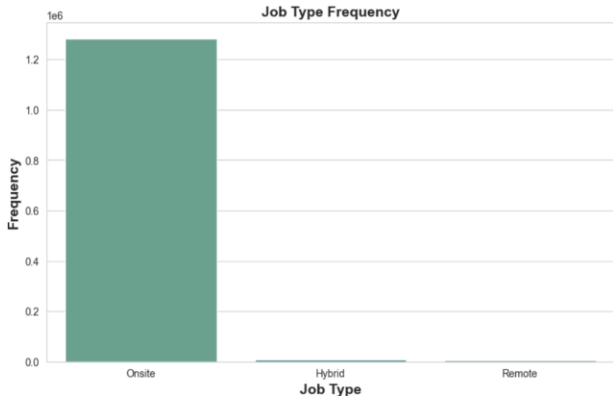




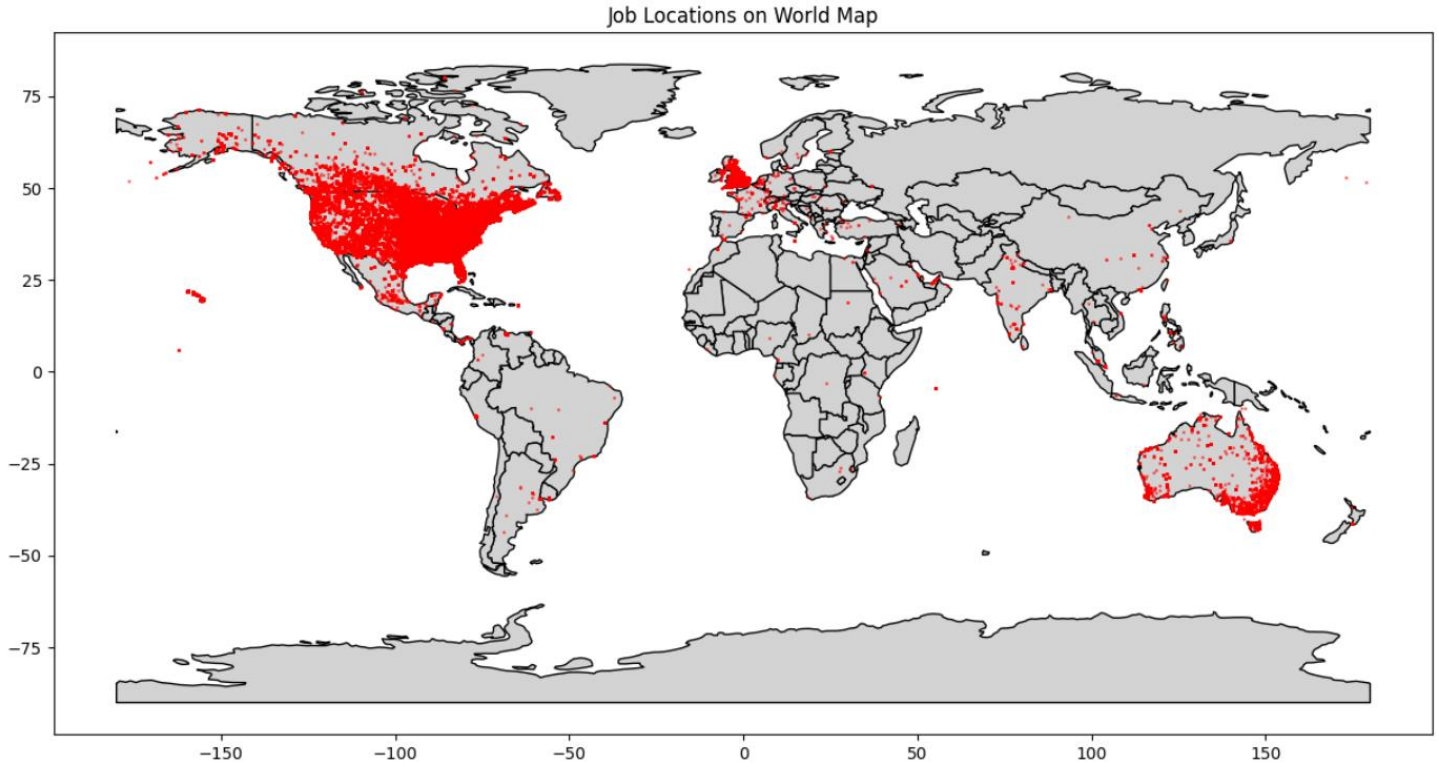
- Job Level Offered:



- Job Type Offered



Job Locations Visualization



Topic Modeling Results Using Latent Dirichlet Allocation (LDA)

In this section, we present the results of the topic modeling process using Latent Dirichlet Allocation (LDA) on the job summaries. Topic modeling helps in identifying underlying themes or clusters in a set of text data. By applying LDA to the job listings, we were able to extract meaningful topics that can be used to categorize and better understand the job descriptions.

Understanding the LDA Output

The results of the LDA model are shown below. Each line represents one topic, which is characterized by the most frequent keywords and their corresponding weights. The weights indicate the importance of each keyword within a topic, with higher weights suggesting greater relevance. The topics are indexed by numbers (0, 1, 2, etc.), which act as identifiers for each theme.

Example of LDA Output:

```
(0, '0.023*"pay" + 0.013*"insurance" + 0.012*"medical" + 0.012*"hour" + 0.011*"healthcare"')
(1, '0.024*"status" + 0.018*"disability" + 0.018*"employment" + 0.016*"gender" + 0.014*"protect"')
(2, '0.039*"care" + 0.038*"patient" + 0.020*"health" + 0.014*"nursing" + 0.012*"clinical"')
(3, '0.015*"client" + 0.010*"manage" + 0.010*"business" + 0.009*"manager" + 0.009*"financial"')
```

Interpreting the Topics

Based on the top keywords and their weights, we can interpret the following topics:

- **Topic 0: Pay & Benefits**
    - Keywords: pay, insurance, medical, healthcare, hour
    - This topic is likely related to job listings discussing pay rates, hourly jobs, and benefits such as medical insurance and healthcare. These may correspond to entry-level or service industry positions.
- **Topic 1: Equal Opportunity & Compliance**
    - Keywords: status, disability, employment, gender, protect
    - This topic focuses on job postings related to non-discrimination and equal opportunity, often found in legal or HR-related sections of job descriptions.

Significance of Topic Modeling

This approach is powerful because it allows us to categorize and understand job postings without relying on manually labeled data. The identified topics provide insights into job trends, such as the prevalence of certain industries (e.g., healthcare) or language (e.g., French-speaking regions). These topics can be used to:

- **Assign a dominant topic to each job post**: Based on the topic with the highest probability, we can classify each job listing into a specific category.
- **Visualize trends**: By visualizing the distribution of topics across different job listings, we can better understand which sectors dominate in a given dataset (e.g., healthcare, retail, business).
- **Search and filter**: Users can filter job listings based on the identified topics, making the search more relevant.
- **Build recommender systems**: Topic modeling can power recommendation engines that suggest jobs based on the user's selected topic.
- **Create dashboard tabs by topic**: A job dashboard could include different sections for each topic, making it easy for users to explore jobs in specific categories.

Job skills Word Cloud

Created Word cloud to better understand the most required skills

8.Challenges Faced

- Handling Missing or Incomplete Data: Some job records lacked summaries or skills, which required filtering and cleaning.
- Merging Large Datasets: Joining 1.3 million records from multiple files was computationally intensive.
- MongoDB Insertion: Inserting a large volume of documents into MongoDB required optimization to avoid performance issues.
- Text Cleaning: Standardizing unstructured text fields like job summaries and skills required regular expressions and manual adjustments.

9. Key Findings / Insights

Most In-Demand Skills:

- Python, SQL, and Machine Learning emerged as the top skills in demand across job listings.
- Other frequently required technologies included AWS, TensorFlow, Power BI, and Tableau, highlighting the emphasis on cloud platforms and data visualization.
- Soft skills like communication and teamwork were also mentioned but less frequently than technical skills.

Companies Hiring the Most:

- The analysis showed that certain large tech companies and consultancies had significantly more job listings than others.
- For example, companies like Accenture, Deloitte, and Amazon appeared frequently, indicating high demand for data professionals in consulting and cloud services.

Trends in Job Locations and Levels:

- Remote roles and hybrid work models are becoming more common, although cities like New York, San Francisco, and London still lead in job volume.

- Job levels were diverse, but there was a strong demand for mid-level positions like Data Analysts and Machine Learning Engineers, with fewer entry-level and senior roles comparatively.
- Roles like Data Scientist, ML Engineer, and Data Analyst are the most advertised job titles.

10.Conclusion

This project successfully analyzed a dataset of job listings to uncover valuable insights about the current demand in the data job market. By extracting top skills, identifying key hiring companies, and understanding location and level-based trends, the project helps job seekers and recruiters alike to stay informed.

Accomplishments:

- Identified top in-demand skills and tools in the industry.
- Highlighted hiring patterns across companies and cities.
- Provided a snapshot of role levels and distribution.

Future Improvements:

- Deploying an interactive dashboard (e.g., using Dash, Streamlit, or Power BI) for dynamic job market analysis.
- Predicting job trends using time-series models to forecast which skills and roles will grow in demand.
- Implementing NLP models to match resumes with job descriptions for better candidate-job fit and automated screening.

11. References

- Dataset Source:
  Kaggle Job Listings Dataset (Insert specific dataset link here)
- Official Documentation:
  - Python Docs

    - Pandas Documentation
    - MongoDB Documentation (if used)
- Other Resources:
    - Stack Overflow (For troubleshooting and code-related queries)
    - Medium articles and blogs (For job trend insights and feature engineering inspiration)