



UNIVERSITY OF MARYLAND, BALTIMORE COUNTY
MPS DATA SCIENCE, FALL 2023

DATA 603
FINAL PROJECT PRESENTATION
GROUP #7

ANALYZING JOB MARKET TRENDS USING LINKEDIN JOB LISTINGS DATA

Professor – Dr. Najam Hassan

Project Team:

Saivarun Kotha – FW61697
Varun Aditya Madala – ZW71217
Sai Manvitha Nadella – FT44056





Introduction

- In today's dynamic job market, understanding trends and demands is crucial for professionals and organizations.
- Our project aims to analyze 1.3 million job listings scraped from LinkedIn in 2024 to gain insights into job market dynamics, identify skill demands, and enhance job recommendation systems.
- This project is not only academically relevant but also provides practical insights that can benefit professionals, recruiters, and policymakers.
- Through this analysis, we seek to contribute to the field of data science and provide actionable insights for stakeholders in the job market.
- We are also developing a job recommendation system to develop a job (content-based) recommender system that accurately matches job seekers' skills with the most appropriate LinkedIn job postings. In doing so, to provide job seekers with actionable information and strategic guidance.



Objectives

Analyze Job Market Trends

Goal: Identify the most in-demand job titles and industries in different cities or countries.

Why?: Lack of comprehensive data on current job market trends hinders informed decision-making for job seekers and employers.

Company Hiring Patterns

Goal: Determine the top companies hiring for specific job positions.

Why?: Understanding which companies are actively hiring for certain positions can help job seekers target their applications effectively.

Skill Mapping

Goal: Utilize skills data to determine the most sought-after skills in different job categories.

Why?: Identifying the skills in high demand can help educational institutions and training programs tailor their offerings to meet market needs.

Job Recommendation System

Goal: To alleviate the challenges faced by job seekers in the current job market by providing them with personalized and relevant job recommendations.

Why?: Motivated by the increasing complexity and competitiveness of the job market, exacerbated by economic uncertainties and layoffs. There is a clear need to bridge the gap between job seekers' skill sets and the requirements of job postings.



Data Sources and Collection

The dataset was obtained from Kaggle, a platform for data science and machine learning datasets. The dataset contains 1.3 million job listings scraped from LinkedIn in 2024, providing a rich source of information for analysis.

Data Sets


job_skills.csv (Includes skills and job id)

job_summary.csv (Includes summary of each job id)

Linkedin_job_postings.csv (Includes entire description of job posting)

Data Collection Method

The dataset was originally collected using web scraping techniques to extract job listings from LinkedIn. The data includes information such as job title, company, location, skills required, and job description.



Data Sources and Collection

job_skills.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1296381 entries, 0 to 1296380 Data columns (total 2 columns):
#      Column      Non-Null Count  Dtype
---  ---
0      job_link      1296381 non-null  object
1      job_skills      1294346 non-null  object
dtypes: object(2)
```

job_summary.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1296381 entries, 0 to 1296380 Data columns (total 2 columns):
#      Column      Non-Null Count  Dtype
---  ---
0      job_link      1296381 non-null  object
1      job_summary    1294346 non-null  object
dtypes: object(2)
```

linkedin_job_postings.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1348454 entries, 0 to 1348453 Data columns (total 14 columns):
#      Column      Non-Null Count  Dtype
---  ---
0      job_link      1348454 non-null  object
1      last_processed_time  1348454 non-null  object
2      got_summary    1348454 non-null  bool
3      got_ner        1348454 non-null  bool
4      is_being_worked  1348454 non-null  bool
5      job_title      1348454 non-null  object
6      company        1348443 non-null  object
7      job_location    1348435 non-null  object
8      first_seen     1348454 non-null  object
9      search_city     1348454 non-null  object
10     search_country  1348454 non-null  object
11     search_position  1348454 non-null  object
12     job_level       1348454 non-null  object
13     job_type        1348454 non-null  object
dtypes: bool(3), object(11)
```

Data Sources and Collection

Challenges in Data Collection

Ensuring Data Quality: Addressing issues such as incomplete or inaccurate data, and standardizing formats for consistency.

Managing Large Dataset: Dealing with the volume of data requires efficient storage and processing solutions.

Legal and Ethical Considerations: Ensuring compliance with LinkedIn's terms of service and data protection regulations.

Data Preprocessing

Cleaning the raw data to remove duplicates, handle missing values, and format data for analysis.
Transforming the data into a structured format suitable for analysis using big data platforms.



Tools and Technologies



PySpark:

Purpose: Apache Spark is a fast and general-purpose cluster computing system for big data processing.

Usage: Spark will be used for processing the large dataset of job listings, enabling efficient data manipulation and analysis. PySpark is also used for implementation of Job Recommendation System

Python (Jupyter IDE):

Purpose: Python is a versatile programming language widely used for data cleaning, analysis, and visualization.

Usage: Python libraries such as Pandas and NumPy will be used for data cleaning and manipulation tasks. It is also used for Exploratory Data Analysis using Matplotlib and seaborn libraries.



Tools and Technologies



mongoDB®



Power BI

MongoDB:

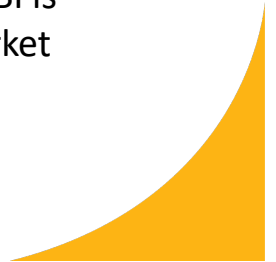
Purpose: The purpose of using MongoDB in the project is to serve as a data store for hosting the job-related data.

Usage: MongoDB is a NoSQL database that offers flexibility and scalability, making it suitable for handling large volumes of unstructured or semi-structured data, such as job postings and skill data.

PowerBI:

Purpose: Power BI is used in the project for data visualization and reporting purposes.

Usage: After analyzing and processing the data using Python in Jupyter IDE and Spark, Power BI is used to create interactive visualizations and dashboards that provide insights into the job market trends, company hiring patterns, skill mapping, and job recommendations.





Tools and Technologies

Reasons for Choosing Tools:

Scalability:

PySpark provides scalability for processing large datasets efficiently.

Versatility:

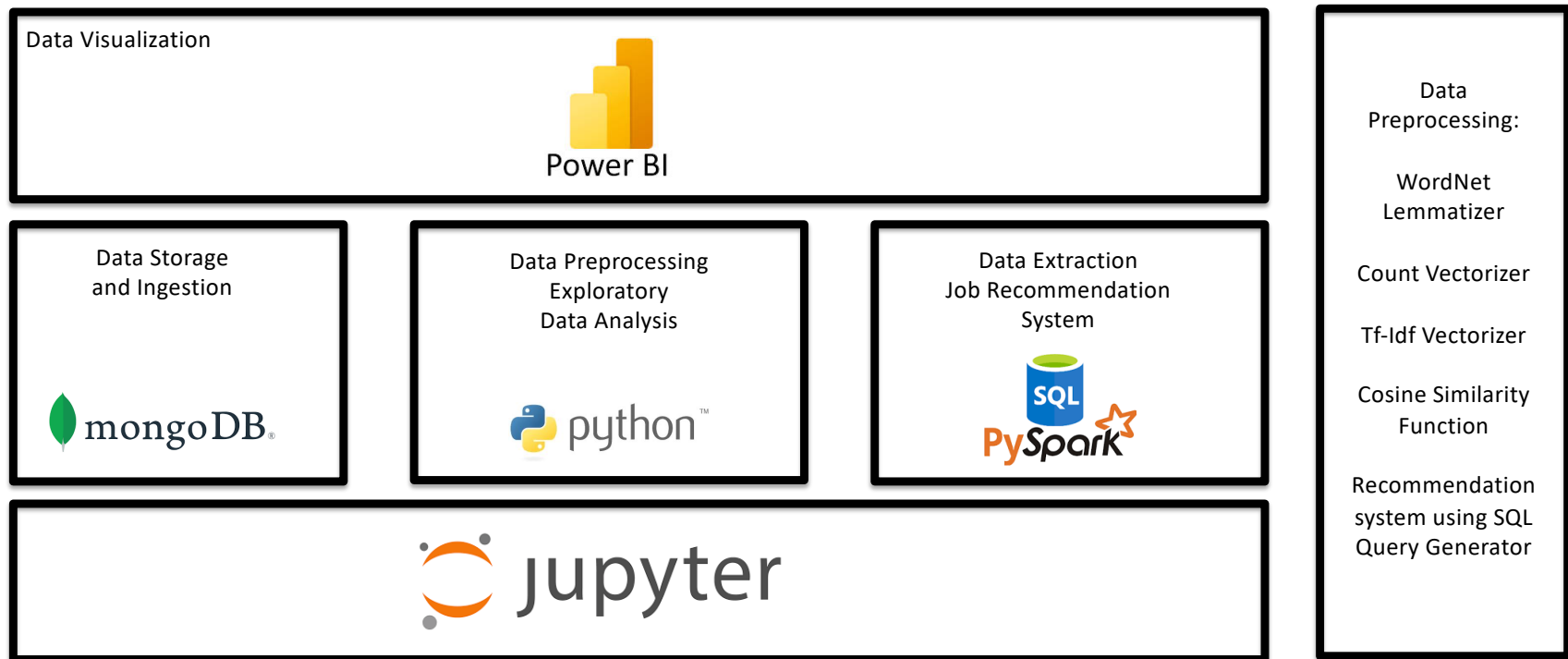
Python's versatility and rich ecosystem of libraries make it ideal for data analysis tasks.

Familiarity:

Jupyter Notebooks and SQL are widely used tools in the data science community, ensuring ease of use and familiarity for team members.



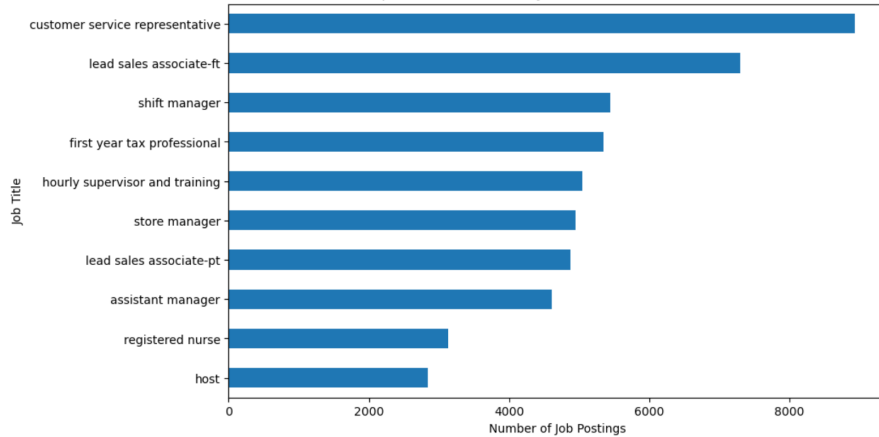
Data Stack Diagram:



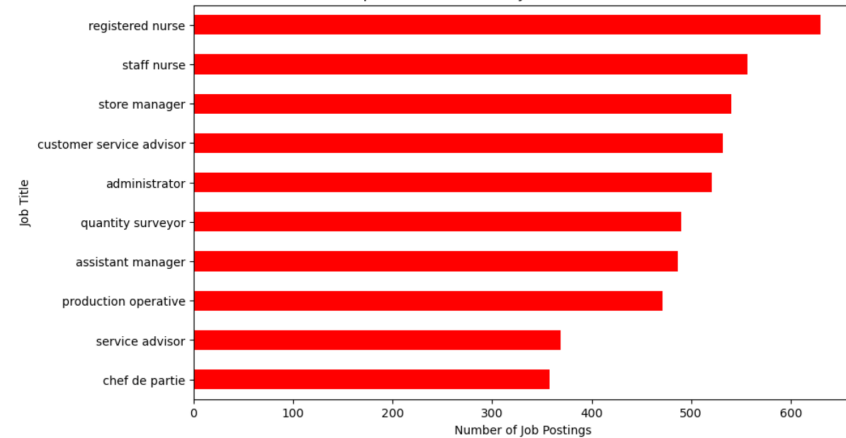


Job Demand In Different Countries:

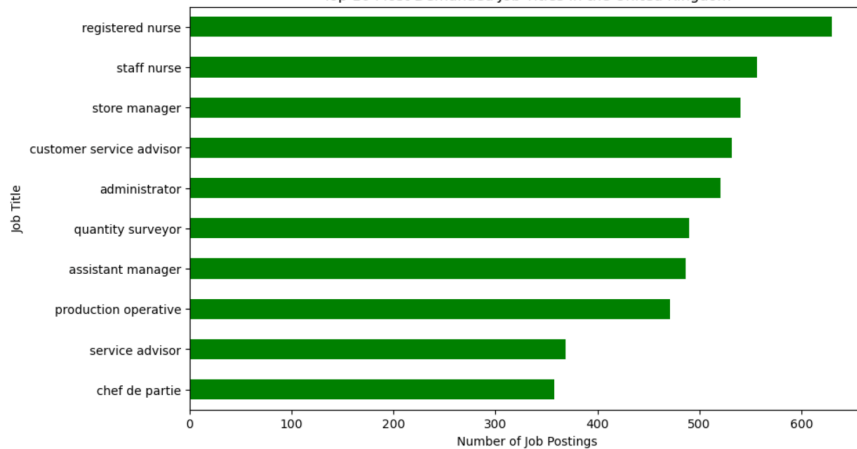
Top 10 Most Demanded Job Titles in the United States



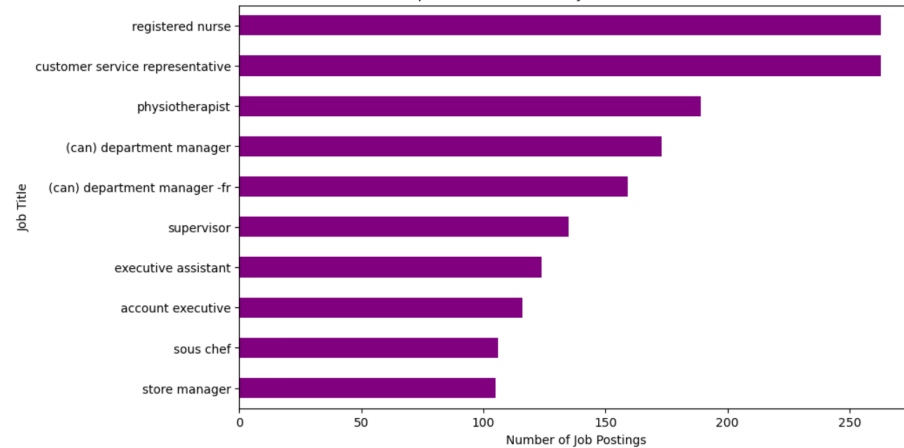
Top 10 Most Demanded Job Titles in the Australia



Top 10 Most Demanded Job Titles in the United Kingdom

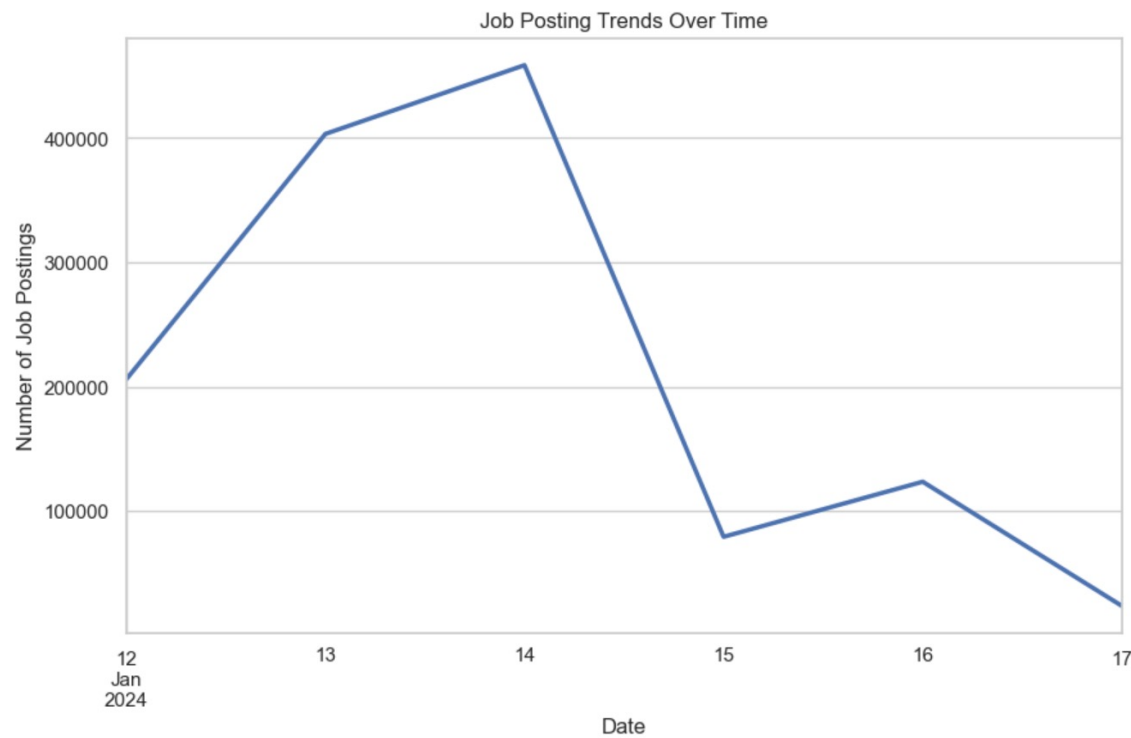


Top 10 Most Demanded Job Titles in the Canada



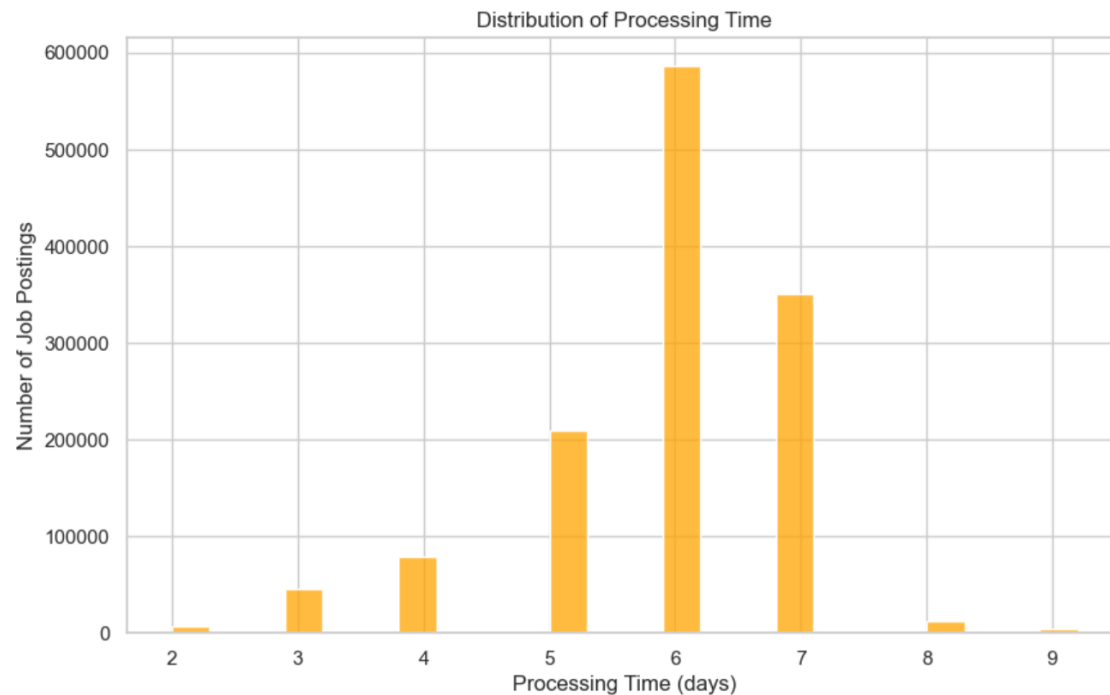


Job Posting Trends Over Time



The graph displays the number of job postings over a six-day period from **January 12, 2024** to **January 17, 2024**.

This trend suggests that job seekers would have the most opportunities if they were to search for jobs around **January 14**. However, it's important to note that these trends can vary and may not be the same in the future.



The provided chart is a bar graph that represents the distribution of processing time for job postings.

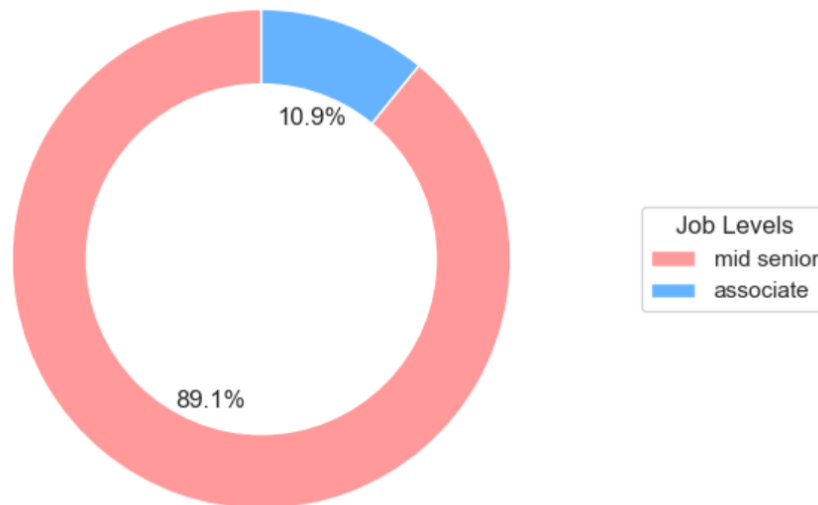
This chart provides valuable insights into the distribution of processing times for job postings.

It can help in understanding the efficiency of the job posting process and identifying potential areas for improvement.

For instance, strategies could be developed to evenly distribute the processing load across all days to avoid significant fluctuations.



Distribution of Job Levels in LinkedIn Postings

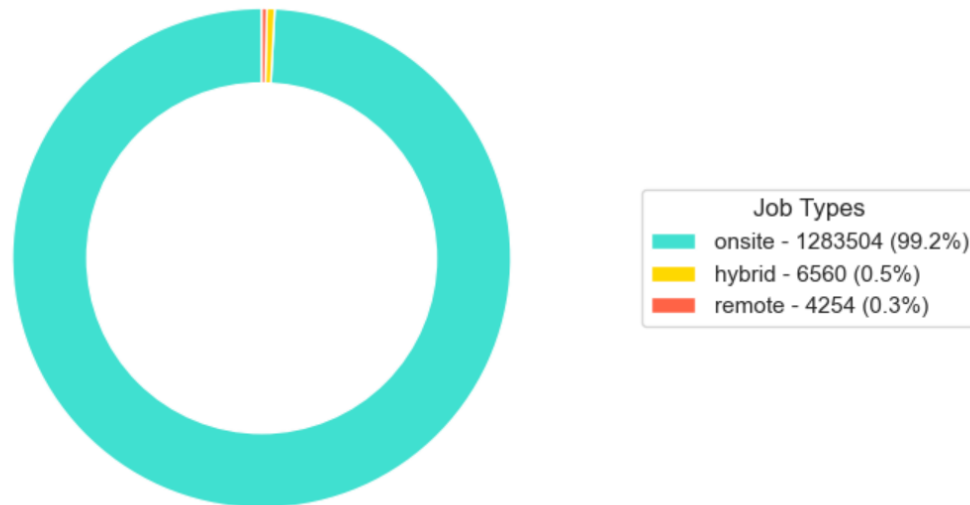


The chart shows two job levels: mid-senior and associate. The **mid-senior level** jobs are depicted in pink and **constitute the majority at 89.1%**. On the other hand, **associate level** jobs are depicted in blue and **make up 10.9% of the total**.

This chart provides valuable insights into the job market dynamics on LinkedIn, particularly regarding the distribution of job levels. It can be useful for job seekers to understand where they might have the best chances of finding job postings that match their level of experience.



Distribution of Job Types in LinkedIn Postings



The chart shows three job types: onsite, hybrid, and remote. The **onsite jobs** are depicted in cyan and constitute the majority at **99.2%**. **Hybrid jobs** are depicted in golden and make up **0.5%** of the total, while **remote jobs** are depicted in red and account for **0.3%**.

This chart provides valuable insights into the job market dynamics on LinkedIn, particularly regarding the distribution of job types. It **can be useful for job seekers to understand where they might have the best chances of finding job postings that match their preferred work arrangement.**





Most sought-after skills in different jobs

	job_title	Most Common Skill	Occurrences
116170	customer service representative	customer service	8873
229090	lead sales associate-ft	customer service	7015
163226	first year tax professional	customer service	4788
229091	lead sales associate-pt	customer service	4655
491745	store manager	customer service	4505
197225	hourly supervisor and training	inventory management	4468
463984	shift manager	training	4444
33974	assistant manager	customer service	3713
522040	travel allied health professional - ct technol...	ct technologist	2570
360218	registered nurse	nursing	2311

By finding the most common skill for each job title, you can get an idea of what skills are most in demand for each type of job. This can be useful for job seekers, recruiters, or anyone doing labor market research.



Job Recommendation System:

- In an age marked by job cuts and economic uncertainty, the job market has become increasingly challenging for both experienced professionals and recent graduates. The aftermath of such disruptions has led to heightened competition and a pervasive sense of uncertainty for job seekers. One major challenge in this scenario is the mismatch between applicants' skills and the requirements of the positions they seek.
- **Objective:** To create a job recommender system that accurately matches job seekers' skills with the most suitable job postings on LinkedIn, providing them with practical information and strategic advice.
- Content-based recommendation is an approach that tailors suggestions to users based on how closely the attributes or content of item align with their preferences. In this project, this method will focus on recommending job titles based on the similarity between the skills of the job seeker and those listed in the job postings. It will also consider other user preferences such as job location, level, and type.
 - Retrieving data from MongoDB;
 - Running Spark Session;
 - Changing the data into Spark Dataframe and performing the following functions





localhost:27017



My Queries

Performance

Databases



Search

BigData

Linkedin_job_postin...

admin

config

local

My Queries

Linkedin_job_posting_skills



localhost:27017 > BigData > Linkedin_job_posting_skills

Documents

1.3M

Aggregations

Schema

Indexes 1

Validation



Type a query: { field: 'value' } or [Generate query](#)

Explain

Reset

Find



Options

ADD DATA

EXPORT DATA

UPDATE

DELETE

1 - 20 of 1292937



```
_id: ObjectId('663c305c64fc8170e300d5d5')
sn: 0
job_link_cleaned: 3802078767
last_processed_time: "2024-01-21 07:12:29.00256+00"
got_summary: true
got_ner: true
is_being_worked: false
job_title: "Account Executive - Dispensing (NorCal/Northern Nevada) - Becton Dicki..."
company: "BD"
job_location: "San Diego, CA"
first_seen: 2024-01-15T00:00:00.000+00:00
search_city: "Coronado"
search_country: "United States"
search_position: "Color Maker"
job_level: "Mid senior"
job_type: "Onsite"
job_skills: "Medical equipment sales, Key competitors, Terminology, Technology, Tre..."
```

```
_id: ObjectId('663c305c64fc8170e300d5d6')
sn: 1
job_link_cleaned: 3803386312
last_processed_time: "2024-01-21 07:39:58.88137+00"
got_summary: true
got_ner: true
is_being_worked: false
job_title: "Registered Nurse - RN Care Manager"
company: "Trinity Health MI"
job_location: "Norton Shores, MI"
first_seen: 2024-01-14T00:00:00.000+00:00
search_city: "Grand Haven"
search_country: "United States"
search_position: "Director Nursing Service"
job_level: "Mid senior"
job_type: "Onsite"
job_skills: "Nursing, Bachelor of Science in Nursing, Masters Degree in Nursing, Ca..."
```

>_MONGOSH





Job Recommendation System:

Pre-processing the "job_skills" column

Job skills were parsed and tokenized into a list as phrases like "data analysis" rather than individual words "data" and "analysis". This method enables a more precise representation and understanding of the skills needed for each job, ensuring that subsequent analyses are based on meaningful combinations of skills rather than isolated terms.

Applying CountVectorizer

This was achieved by fitting the cleaned "job skills" column using the CountVectorizer package in Pyspark. In order to control the size of the sparse vectors, we set a minimum document frequency threshold of 100, filtering out less common or potentially noisy job skill phrases in order to focus on skills that occur with sufficient frequency and relevance across the dataset.

Calculate TFIDF & Completing the Item (Job Listing) Profile

The Inverse Document Frequency (IDF) is computed to assess the importance of each term across the entire data set, more frequently appearing job skill phrases will be given a lower weightage. This IDF value is then scaled with the TF values to compute the Term Frequency - Inverse Document Frequency (TF-IDF) scores. The TF-IDF gives a score that measures how important a job phrase is, in relation to the other job phrases that appear in the LinkedIn job dataset.

Preparing the User Profile (Test Cases)

The user profile (test cases) is constructed based on the job seeker's input, consisting of the key job attributes including the job location (city and country), level, type and skillset.

The user's preference for job city, country, level and type will constitute the Spark SQL query used to filter the item profile.



Job Recommendation System:

Building the SQL Query & Filtered Item (Job Listing) Profile

In order to reduce the number of job listings that need to be compared for skills similarity as well as make the search more efficient and targeted, the PySpark dataframe containing the full item profile is filtered using a SQL query to obtain listings that match the user's desired (i) job location, (ii) level and/or (iii) skills. These three features are optional, and the user may fill them in any combination, or not at all (in which the entire item profile will be assessed). The resulting dataframe will form the 'filtered' item profile.

Define the Cosine Similarity Function

Recommendation Algorithm: Matching Most Relevant Jobs Based on User's Skills

Based on the skills TF-IDF vectors from the item (for each job listing) and user profile, the cosine similarity scores are calculated. A higher score suggests a higher degree of similarity between the job requirements and the user's skill set, making the job more relevant and suitable for the user. The final recommendation to the user includes the top 5 most relevant jobs with the highest cosine similarity scores.

The output is then exported to `output_job_skills_match.csv`





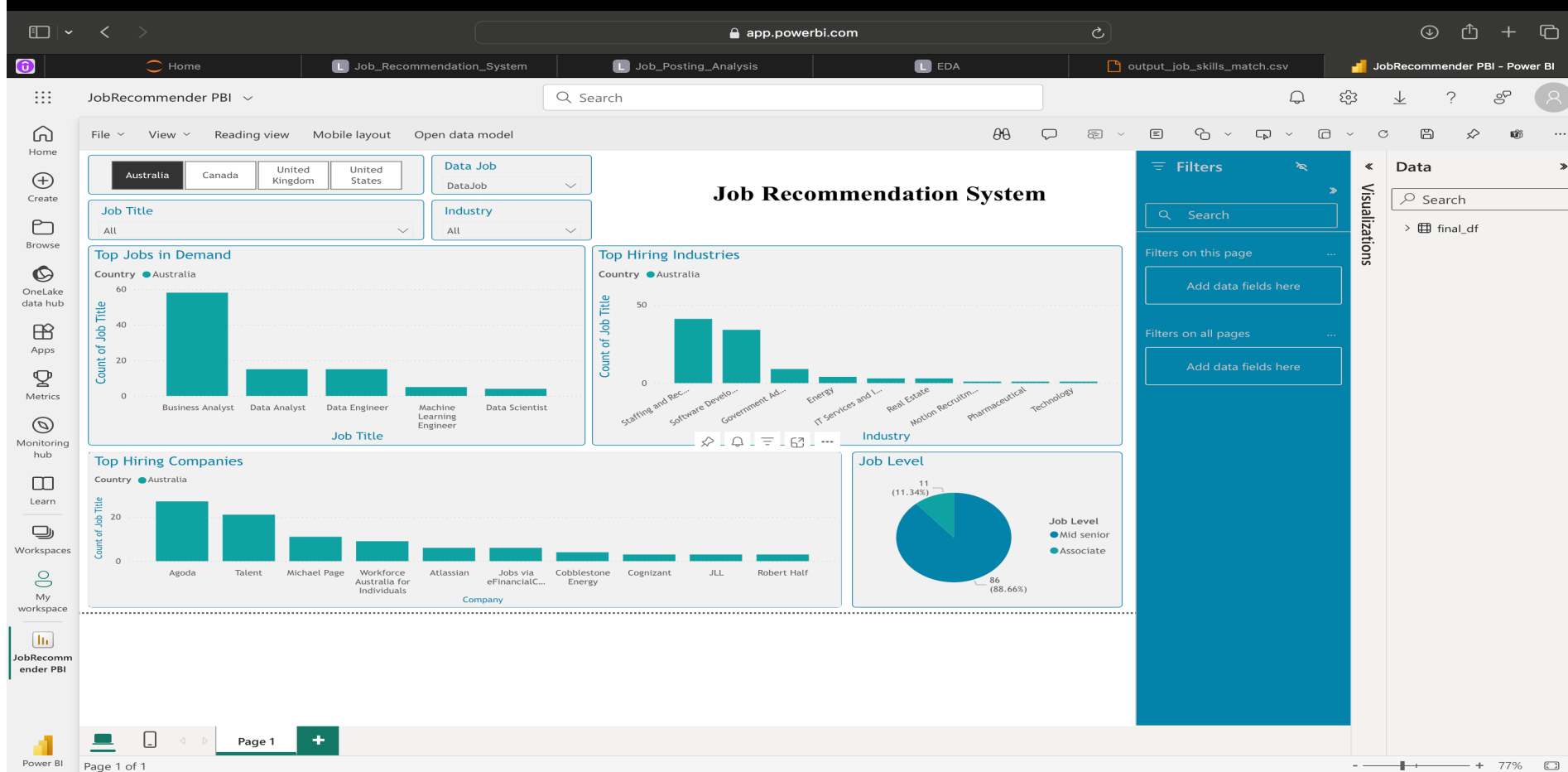
Job Recommendation System:

```
import pandas as pd
output = pd.read_csv('output_job_skills_match.csv')
output.head(5)
```

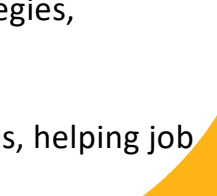
	query_sn	users_skills	job_suitability_rank	job_title	company	job_location	job_level	job_type	job_skills	similarity_score
0	1	Python, R, SQL, Machine learning, Data analysi...	1	Data Scientist II, Middle Mile Transportation	Amazon	Bellevue, WA	Mid senior	Onsite	Data Science, Statistics, Machine Learning, Op...	0.786392
1	1	Python, R, SQL, Machine learning, Data analysi...	2	Data Scientist II, Infra DCPD - PACE	Amazon Web Services (AWS)	Seattle, WA	Mid senior	Onsite	Machine learning, Data analytics, Python, SQL,...	0.561069
2	1	Python, R, SQL, Machine learning, Data analysi...	3	Data Scientist II, DSP Insurance	Amazon	Bellevue, WA	Mid senior	Onsite	Data Scientist, Data Extraction, Data Analysis...	0.371950
3	1	Python, R, SQL, Machine learning, Data analysi...	4	Applied Scientist II, Customer Behavior Analytics	Amazon	Seattle, WA	Mid senior	Onsite	Data mining, Causal inference, Machine learnin...	0.284866
4	1	Python, R, SQL, Machine learning, Data analysi...	5	Data Scientist II	The Trade Desk	Bellevue, WA	Mid senior	Onsite	Scala, Spark, Python, R, SQL, Machine learning...	0.283252



Power BI Report



Future Opportunities:

- 1.Enhanced Personalization:** Developing more advanced recommendation systems that consider a broader range of user data, such as past job experiences, education, and endorsements, to provide even more personalized job suggestions.
 - 2.Market Trend Prediction:** Using machine learning algorithms to analyze historical job posting data and predict future job market trends, helping job seekers and employers anticipate changes and adapt their strategies accordingly.
 - 3.Skill Gap Analysis:** Conducting in-depth analysis to identify skill gaps in the job market and provide recommendations for individuals and organizations to bridge these gaps through training and education programs.
 - 4.Real-time Job Matching:** Implementing real-time job matching algorithms that instantly match job seekers with open positions as soon as they are posted, increasing efficiency for both job seekers and recruiters.
 - 5.Competitive Intelligence:** Providing tools for companies to analyze their competitors' hiring trends and strategies, enabling them to stay competitive in the talent market.
 - 6.Geospatial Analysis:** Incorporating geospatial analysis to understand job market dynamics in different regions, helping job seekers identify locations with higher demand for their skills.
- 



Upcoming technology have the potential to impact the business strategy:

1.Natural Language Processing (NLP): NLP can enhance the analysis of job postings by extracting more nuanced information from job descriptions, such as required skills, qualifications, and responsibilities, leading to more accurate job matching and recommendation systems.

2.Machine Learning (ML) and Artificial Intelligence (AI): ML and AI algorithms can improve the accuracy of job recommendations by analyzing user behavior and preferences, as well as by predicting future job market trends based on historical data.

3.Blockchain: Blockchain technology can be used to verify the authenticity of job postings and candidates' credentials, reducing the risk of fraudulent job postings and improving the overall trustworthiness of the job market.

4.Big Data Analytics: Advanced big data analytics techniques can enable deeper insights into job market trends, company hiring patterns, and skill mapping, allowing for more informed decision-making by job seekers and employers.



References:

Hosain, Md & Liu, Ping. (2020). LinkedIn for Searching Better Job Opportunity: Passive Jobseekers' Perceived Experience. *The Qualitative Report*. 25. 3719-3732. 10.46743/2160-3715/2020.4449.

Utz S, Breuer J. The Relationship Between Networking, LinkedIn Use, and Retrieving Informational Benefits. *Cyberpsychol Behav Soc Netw*. 2019 Mar;22(3):180-185. doi: 10.1089/cyber.2018.0294. Epub 2019 Jan 16.

Eseryel, U. & Booij, Richard & Eseryel, Deniz. (2018). Recruitment through LinkedIn: Lessons learned from the Fortune 100 Companies.





UMBC

THANK YOU

