Skill-Based Job Recommendation: A Machine Learning Approach on LinkedIn Job Listing

GOH DAI YONG ADISON
NISIN SAJ

JIAO HUIMING
RUSSELL QUAH LIANG WEI

1.0. INTRODUCTION

In an era marked by layoffs and economic uncertainty, the landscape of job seeking has become increasingly daunting for both seasoned professionals and recent graduates alike. The aftermath of such upheavals has ushered in intensified competition within the job market (Times, 2023), accompanied by a pervasive sense of uncertainty looming over every application submitted. Amidst this challenging backdrop, a significant hurdle emerges – the glaring mismatch between the skill sets possessed by applicants and the requirements stipulated by the positions they aspire to secure.

This disjuncture between candidates' proficiencies and job specifications often arises from a multifaceted array of factors. One prominent contributor is the pervasive practice of candidates casting a wide net, indiscriminately applying to numerous positions in a scattergun approach. While this tactic may stem from a sense of urgency or a desire to maximize opportunities, it frequently leads to a squandering of valuable time and effort, both for the applicants themselves and the employers tasked with evaluating their suitability.

Recognizing the detrimental impact of this misalignment, there arises an imperative need to devise innovative solutions that streamline the job seeking process, mitigate frustration, and alleviate financial strain for all parties involved.

By harnessing the power of advanced algorithms and data analytics, the job recommender systems function as a beacon of guidance amidst the sea of job postings, offering tailored recommendations tailored to the unique skill sets, experiences, and career aspirations of each individual user. Through sophisticated matching algorithms, this system sifts through vast troves of job listings, pinpointing those opportunities that best align with the candidate's profile, thereby minimizing the likelihood of wasted efforts on ill-suited positions.

Moreover, beyond merely serving as a facilitator for job discovery, the job recommender system embodies a promise of efficiency and efficacy in the recruitment process. By presenting candidates with targeted recommendations, it empowers them to channel their energies more purposefully, directing their applications towards roles where they stand a higher probability of success. Simultaneously, employers' benefit from a more refined pool of applicants, sparing them the arduous task of sifting through scores of irrelevant resumes.

1.1. Problem Statement

Our project aims to develop a job recommender system that accurately matches job seekers' skills with the most appropriate LinkedIn job postings.

Additionally, we aim to derive valuable patterns and insights from job postings to provide job seekers with actionable information and strategic guidance.

2.0. DATA PREPARATION

2.1. Data set

For this project, we've selected a dataset available on the Kaggle platform titled "1.3M LinkedIn Jobs & Skills (2024)" (Kaggle, 2024). This dataset comprises three distinct CSV files: Job_skills.csv, jobs_posting.csv, and Job_summary.csv. Our focus lies primarily on leveraging the job_skills and job_posting datasets, as the job_summary data primarily consists of job descriptions, which are not essential for our task of recommending job titles based on skills. Below is the breakdown of the dataset.

Table 1: Data Summary

File	Features	Dataset Size
Job skills (job_skills.csv)	job_link (PK/FK), job_skills	672MB, 1296381 rows, 2 columns
Job posting details (linkedin_job_postings.csv)	job_link (PK/FK), last_processed_time, got_summary, got_ner, is_being_worked, job_title, company, job_location, first_seen, search_city, search_country, search_position, job_level, job_type	415MB, 1348454 rows, 14 columns

2.2. Data Cleaning

The data had to go through a thorough cleaning process. The cleaning was performed in 3 different places, some of the details were cleaned in local python, some were in Google cloud python and using google cloud pyspark.

2.3. Local Python

After extracting data from the Kaggle website, we conducted preliminary cleaning procedures within our local Python environment before integrating the datasets. These procedures encompassed essential tasks like splitting URL links to generate unique identifiers. The job link columns were cleaned by removing the leading URL characters, leaving behind a numerical string that could be used as a foreign key to join both job_skills and job_posting_details table.

Before joining the two tables, duplicate job postings were removed from both tables using the newly created numerical string. An inner join was then used to merge the job_skills and job_posting_details on the numerical unique identifier. Preliminary text cleaning on removing space and cleaning up unicodes was also done at this stage.

2.4. Google Cloud Python

Subsequently, the combined table was uploaded to Google Collab which can facilitate collaborative work between team members. Cleaning done at this stage was achieved using a combination of regular expressions, manual cleaning and referencing the actual company on LinkedIn. The top 10 job titles and top 200 company names were the focus of this stage of data cleaning, facilitating a comprehensive industry-based analysis.

2.5. Google Cloud Pyspark

The last stage of the data cleaning process was again done on Google Collab but with PySpark Dataframes this time to effectively handle the large volume of the data set. Specifically, the focus lies on cleaning "Job Skill," a crucial step in extracting valuable insights from textual data and in preparation for similarity computation.

Job skills were parsed and tokenized into a list as phrases like “data analysis” and not individual words “data” and “analysis”. This approach facilitates more accurate representation and interpretation of the skills required for each position, ensuring that subsequent analyses are based on meaningful skill combinations rather than isolated terms.

Lemmatizing, an integral part of the cleaning pipeline, involves reducing words to their base or root form, thereby standardizing variations, and enhancing the consistency of the dataset. This step is particularly valuable in maintaining semantic coherence and improving the accuracy of downstream analyses.

Furthermore, the removal of stop words plays a pivotal role in streamlining the dataset by eliminating common words that carry little semantic meaning. By filtering out stopwords, the focus shifts to meaningful content, enhancing the quality and relevance of the processed data.

2.6. Data Preprocessing

Once the data has been appropriately cleaned, the next step is to process it to be able to be fed into the recommender system. Once again, we continued the use of Pyspark Dataframes on Google Collab to handle the volume of the data.

The first step of the pre-processing is to calculate the Term Frequency (TF) of the job skills. This is a measure of how frequently a job skill phrase appears in the dataset of job postings. This was achieved by fitting the “job skills” column, representing various skills across all job listings, 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. Doing so led to a sparse vector size of 21,083. Subsequently, we transformed the cleaned job skill phrases for each job listing into a sparse vector of token counts each.

Next, 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. Instead of utilizing a sparse matrix representation, each job row is associated with a sparse vector identity, encapsulating the TF-IDF information. This sparse vector identity serves as the item profile crucial for the Recommender System in section 4, facilitating personalized recommendations based on the content similarity between job postings.

3.0. DATA ANALYSIS

3.1. Exploratory Data Analysis

Understanding trends in job demand extends beyond more job searching. It influences career trajectories and informs governmental policies. Awareness of these trends is pivotal for strategic career planning and maintaining competitiveness in the professional landscape.

Upon thorough examination of the job market across various countries, our dataset revealed a predominant concentration of job postings in the United States. Consequently, our exploratory data analysis (EDA) has centered on this geographical region, with a specific focus on roles pertaining to data, an area currently experiencing heightened activity.

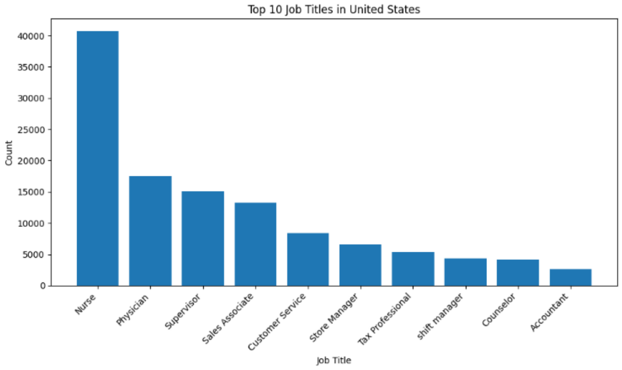

Figure 1: Top 10 Job titles in the United States

Our findings illuminate that nursing emerges as the most sought-after job title across diverse sectors followed by the Physician. The healthcare sector in the US is one of the largest and most dynamic industries, driven by factors such as an aging population, advancements in medical technology, and increasing healthcare needs. (University, 2023) As a result, there is a continuous demand for skilled healthcare professionals to meet the diverse needs of patients across various settings which shows how nursing and physician same into the top listed jobs.

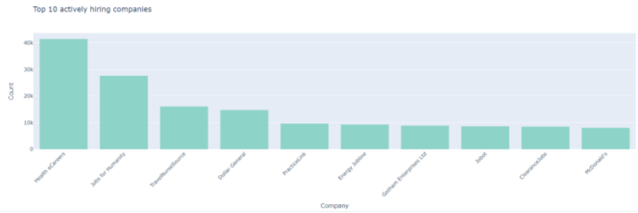

Figure 2: Top 10 Actively hiring companies

Upon analyzing the actively hiring companies, we observed that staffing and recruiting firms stand out as the top contenders on the list. Notably, they predominantly focus on hiring for the healthcare sector. This trend is unsurprising, considering that a significant portion of healthcare institutions rely on third-party recruiters, especially when seeking frontline healthcare workers, and particularly if they aim to recruit talent from diverse geographical locations.

Analysis of the top 10 job titles underscores the prominence of industries such as healthcare, retail, finance and services. Consequently, the significance of skills such as leadership, customer services and finance management in today's workforce cannot be overstated. Within data-related roles, business analysts emerge as highly coveted, closely followed by data engineers and analysts. Notably, the IT (Information Technology) services, consulting and software development sectors emerge as frontrunners in recruitment activities with a predilection for mid-senior level positions.

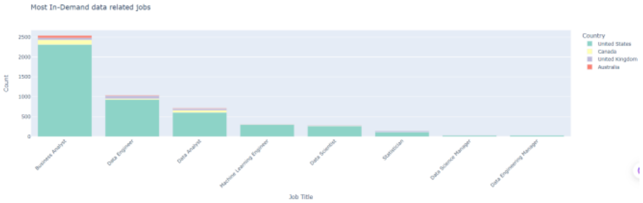

Figure 3: Most in demand data related jobs

This comprehensive understanding of job market dynamics provides invaluable insights into the evolving demands of industries, enabling stakeholders to adapt and thrive in a competitive environment. With these insightful observations, we transition to the subsequent phase of our report: the discussion of our recommender system. This tool serves to align student skills with industry demands, thereby enhancing their efficacy in navigating the job market and fostering a successful career path.

3.2. Dashboarding

A Power BI dashboard has been meticulously crafted to empower users with intuitive tools for slicing and dicing the visuals derived from exploratory data analysis. This dashboard offers users a comprehensive overview of historical trends, allowing them to tailor their analysis to their specific requirements. For instance, individuals interested in data-related roles can narrow down their search by industry, company, and job title, providing valuable insights into job availability. The user interface of the Power BI dashboard is thoughtfully designed to facilitate easy navigation, leveraging measured and DAX formulas for seamless interaction. The visuals featured include top jobs in demand, leading hiring industries, prominent hiring companies, and job levels. Additionally, the dashboard offers slicers for refining searches based on job title, country, data relevance, and industry, empowering users with precise control over their data exploration journey.

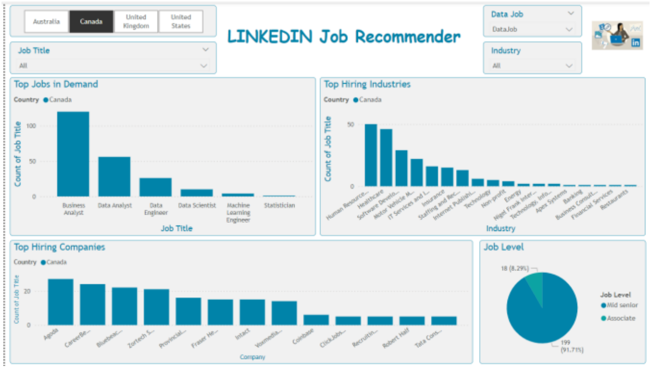

Figure 4: Power BI Dashboard

4.0. RECOMMENDER SYSTEM

Our job recommender system is a sophisticated tool designed to streamline and personalise the job search process by recommending relevant job listings based on the job seeker's preferences.

4.1. Content-based Recommendation

Content-based recommendation is a personalized recommendation approach that suggests items to users based on how well the attributes or content of those items matches the preferences of the user. In the context of our job recommendation system, this method focuses on suggesting job titles based on the similarity between the skillsets of the job seeker and those listed in the job postings. Additionally, other user preferences including job location, level and type will also be considered.

4.2. Creation and Process of Recommender System

A flow chart depicting the recommendation process is shown in Figure 5.

The item profile consists of the job listings, as well as metadata relating to the job, which includes the job location (city and country), job level (associate or mid-senior), job type (onsite, hybrid or remote) and the skillsets required for the job. The item profile was constructed as described in section 2.6, where the skillsets required for each job listing were represented as a sparse vector with a corresponding TF-IDF value for each skill.

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.

The user profile 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 described earlier. Similar to the data cleaning steps for job listing skills described in section 2.5, each input user skill phrase will be tokenized, lemmatized and stopwords removed. Subsequently, the TF-IDF sparse vector will be calculated based on the CountVectorizer & IDF fitted on the skills within job listings described in section 2.6, thereby forming the user's skills profile.

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.

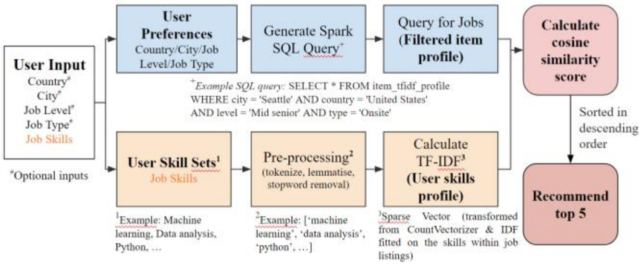
User Input
Country*
City*
Job Level*
Job Type*
Job Skills
*Optional inputs
User Preferences
Country/City/Job Level/Job Type
Generate Spark SQL Query+
Query for Jobs (Filtered item profile)
+Example SQL query: SELECT * FROM item_tfidf_profile WHERE city = 'Seattle' AND country = 'United States' AND level = 'Mid senior' AND type = 'Onsite'
User Skill Sets1
Job Skills
Pre-processing2
(tokenize, lemmatise, stopword removal)
Calculate TF-IDF3 (User skills profile)
1Example: Machine learning, Data analysis, Python, …
2Example: ['machine learning', 'data analysis', 'python', …]
3Sparse Vector (transformed from CountVectorizer & IDF fitted on the skills within job listings)
Calculate cosine similarity score
Sorted in descending order
Recommend top 5

Figure 5: Flow Chart of Recommendation Process Given a User Input

4.3. Evaluation of the Recommender System

The recommender system was evaluated on 10 test cases involving a variety of skills and intended jobs (refer to Table 2). Comparing the expected type of roles to the corresponding top matched job titles, the recommender was generally able to recommend roles that are relevant to users' input skills for all cases.

Table 2: Summarized results of test cases

No.	Expected Type of Role / Skills	Best Matched Role & Company	Similarity Score
1	Data Science / Analytics	Data Scientist II @ Amazon	0.786
2	Data Science / Analytics	Senior Data Scientist AI/ML @ Guardian Life	0.584
3	Communications	Communications Manager @ Michael Page	0.523
4	Marketing / General Management	Practice Development Consultant @ Cynosure, LLC.	0.588
5	Software engineering	Coders @ Braintrust	0.577
6	Business development / sales	Inside Sales Account Executive @ Square	0.747
7	Finance / Portfolio Management/ Audit	Senior TMT/Consumer Analyst @ Mondrian Alpha	0.375
8	Business development / sales	Account Executive @ Infobip	0.428
9	Nursing	Staff Nurse @ Health Recruit Network	0.323
10	Legal	Litigation Paralegal @ AMS Staffing Inc.	0.473

The full list of inputs & outputs, including the specific skillsets of each user and job, can be found in the supplementary CSV files: "test_cases.csv" and "output_job_skills_match.csv" respectively.

However, the similarity scores for several of the test cases, such as case 7 and 9 were relatively low at 0.375 and 0.323 respectively. Observing Table 3, which compares the user input skills and job skills required, some of the skills required by the role instead appear to be broad descriptions of the job (e.g. TMT (Technology, Media & Telecommunications), consumer names, hourly pay, paid breaks, recognition). Furthermore, there are skills that are similar in meaning, such as "Asset Management" and "Portfolio Management". However, they would be represented as different tokens and dimensions in the TF-IDF sparse vector. As a result, the cosine similarity calculation is not able to capture the similarity between such closely related skills.

Therefore, this highlights several limitations of TF-IDF and cosine similarity, including not being able to capture semantic meaning or closely related phrases, as well as being sensitive to noise (i.e. irrelevant descriptions instead of actual job skills).

Table 3: User input skills & job skills required for cases 7 & 9

No.	User Input Skills	Job Skills Required
7	financial analysis, budgeting, forecasting, auditing, investment analysis, risk assessment, financial modeling, tax planning, compliance, asset management	Investment Analysis, Portfolio Management, Hedge Fund, Long/Short Equity Investing, Private Equity Style Approach, TMT, Consumer Names, AUM, Long Term Time Horizon
9	patient care, patient assessment, clinical assessment, medication administration, wound care, patient education, compassion, empathy, communication, teamwork, nursing, record keeping, care planning	nursing, medication administration, patient assessment, record keeping, care planning, communication, teamwork, NMC registration, care home experience, professional development, hourly pay, paid breaks, training, supportive work environment, cuttingedge facilities, recognition, career advancement

Overall, applying a content-based framework for our job recommender system offers personalized job recommendations by considering both the unique combination of skillsets possessed by job seekers and that required by employers. Additionally, the solution is transparent and able to identify features or skills that cause the user to be matched to a job. In addition, there is no cold start issue as users and recruiters alike only need to key in the skillset possessed or desired to initiate the matching process, making their experience seamless.

5.0. CONCLUSION AND FUTURE WORK

The successful implementation of this system underscores its potential to revolutionize the job search process, enhance efficiency, and streamline the recruitment journey for both job seekers and employers.

Looking ahead, there are several avenues for future exploration and enhancement of our job recommender system.

First, a more robust cleaning approach for the job skill will improve the recommender system performance. The current methodology for job skills cleaning still misses out skills that are equivalent but put in a different way. For example, “analysis” vs “analytical skills” vs “analytics” generally are referring to the same thing. But they are currently processed as individual job skills.

Secondly, the model could be deployed as a webservice that is connected to the LinkedIn job portal through APIs. Having a user interface layered on top of this would allow for users to input their job skills and run the recommender model in real time. Users would be able to get their job recommendations in real time.

Thirdly, expanding the scope of the system to incorporate additional data sources, such as economic climate, job market trends and industry insights, could provide more comprehensive and contextually relevant recommendations.

Fourth, application of knowledge graphs as part of a hybrid recommender system, or word embeddings such as Word2Vec or GloVe can be used to capture semantic meaning and relationships between different job skills. Thereby, potentially improving on the performance of the recommender system and its ability to provide contextualized responses.

Through ongoing refinement and innovation, our job recommender system can continue to evolve as a valuable tool for empowering individuals in their career journeys.

5.1. SUPPLEMENTARY MATERIAL

The supplementary materials related to this project, including

datasets, code and the Power BI dashboard file, are available in the following Git repository:

https://github.com/sajnisin/Job-Recommender-System

References

Kaggle. (2024). *1.3M Linkedin Jobs & Skills (2024)*. Retrieved from Kaggle: https://www.kaggle.com/datasets/asaniczka/1-3m-linkedin-jobs-and-skills-2024

Times, T. B. (2023, 6 18). *Layoffs and AI are changing tech’s once-invincible job market*. Retrieved from The Business Times: https://www.businesstimes.com.sg/working-life/layoffs-and-ai-are-changing-techs-once-invincible-job-market

University, M. (2023, 06 01). *Why Are Nurses in Demand? The Nursing Shortage, Explained*. Retrieved from Marquette University College of Nursing: https://mastersnursing.marquette.edu/blog/why-are-nurses-in-demand/

Tables and Figures