Pathway Pro

Avi Mish 208470575
Netanel Shalev 206897969
Amir Menshar 325938645

Introduction and Chosen Path

In today's competitive job market, **staying ahead** in your career is more important than ever. To help professionals grow, we propose creating a LinkedIn-**integrated 2 features add-on** to LinkedIn that will help users who land a new job or users that wants to get promoted. According to the chosen path it will analyze user profiles and their job description and suggests resources like courses, insights and connections with key people in their field directed for everyday end-users. Specifically, it would focus on in-company data if available. Imagine to yourself that you are starting a new position as Data Scientist role in IBM. For a strong base to success, the tool would scan your IBM colleagues' profiles, identify the top-performing Data Scientists, and see what specific skills, courses, or certifications they have in common – and provide you with insights and recommendations, a feature that LinkedIn lacks.

Our main question is: **What sets successful professionals apart from those who struggle?** Specifically, we want to understand the qualities and habits of employees who thrive in their roles. Based on this, we aim to provide users with guidance to improve their job performance, career stability, and advancement.

Benefits of the New Solution:

This tool will help LinkedIn users take charge of their careers by offering:

- Job Stability: Offering tips and tools to align with the top employees in the user's company.
- Tailor made Networking: Suggesting relevant networking opportunities by top employees in the user's company.
- Personalized Learning: Recommending courses and training programs tailored to each user's needs and custom by his job description and current company.

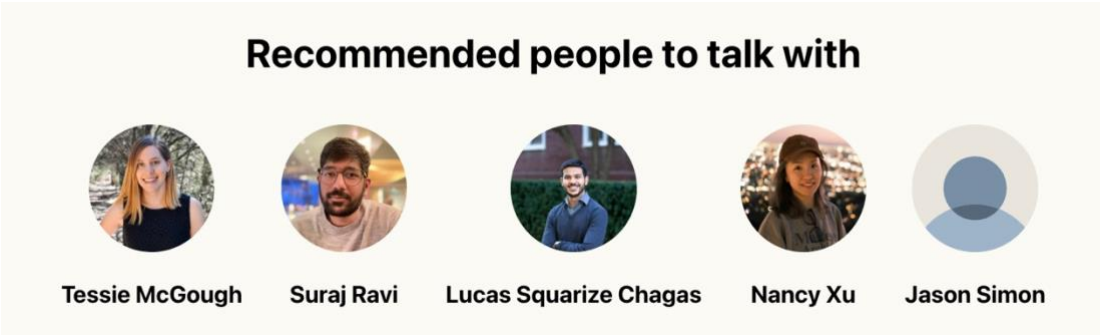
Data Collection and Integration:

- Data used from the original data tables: We used the Profiles data table extensively, mainly the columns: [Id, Experience, Education, Courses, Certifications, Current Company, Position].
- We used Scraping to enrich our data and results, in two cases:
  - Shanghai Ranking - We calculated for each employee his education score (higher ranked universities got a higher score and so on), which in turn helped us calculate an employee chance to advance in the company.
  - Layoffs scraped data - We recognized good employees that survived layoffs in the company of the end-user, helping the user connect with better employees that will provide better insights on how to be invaluable for the company.
- For the shanghai universities ranking we collect 1000 records where each record contains rank and a matched university name.
- For the Layoffs data – we scraped 500 records where each record stands for an event of layoff at some company. E.g. Amazon, 9000 employes, January 2023.

Data Analysis:

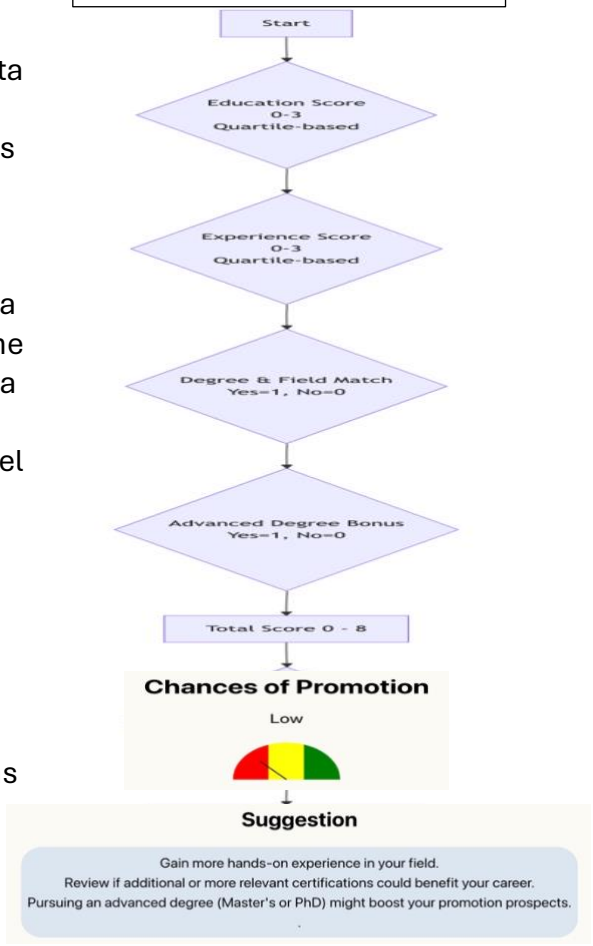
- **Feature 1** focuses on suggesting relevant courses/certifications for the user. First, it does so by looking at the user's company employees which are from the same field as the user, and then it merges similar courses/certifications names. The feature leverages *spaCy*, a powerful natural language processing (NLP) library, to process and normalize course and certification titles. Using named entity recognition (*NER*) and text similarity techniques, *spaCy* helps identify variations of the same course or certification, ensuring consistency across records. This enables a more structured and searchable database, reducing redundancy and improving the accuracy of course and certification recommendations. Also, the feature gives general idea of industry-wide trends for the user's position, using *fuzzywuzzy* string matching to find and merge again the most relevant courses and certifications.

-**Feature 2** deals with finding the most similar employees to consult with based on job titles in the new company. First it uses extensive filtering. We filter by field, e.g. for engineer focusing only on other engineers in the company. We search for big gaps in the resume of all the company employees. Big gaps are over 3 months between jobs. We filter out people with gap percentage of above 50% of their total jobs. Then it filters out all the employees we have been cut off from a layoff, assuming they had lower performance. Then It utilizes SentenceTransformer "*all-miniLM-L6-v2*", a compact yet efficient transformer model, to generate embeddings for job titles. By converting job titles into dense vector representations, the model captures semantic similarities between different roles. The system then applies cosine similarity to compare these embeddings and identify employees with the closest matching expertise. This enables quick and effective knowledge-sharing by recommending the most relevant colleagues for consultation.



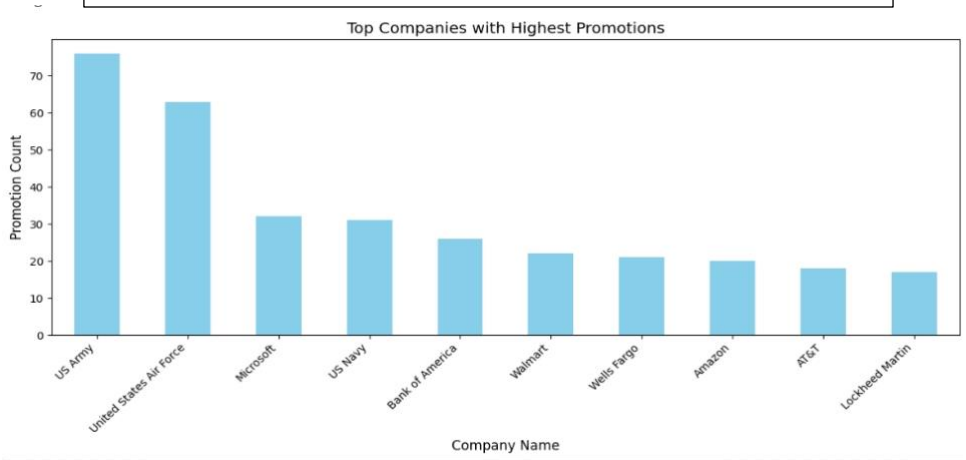
- **Feature 3** attempts to provide the user with a general idea of how he will fare in promoting his position in the workplace. The feature gathers data from employees in the rank desired by the user and provides insights to the user. The prediction is given as Low/Medium/High chance of promotion and is calculated by using a weight function and decision boundaries. It is done in the following way: First it compares the user education-score, a metric we created to calculate how prestigious the end-user education is, which is calculated using a weight function (for example, ivy-league universities receive a multiplier of 20, and the level of the degree also has a different multiplier, for example bachelor is 3) and is compared to the quartiles of the education scores in the desired rank. Then, it does a very similar thing with experience, which is measured in months of experience. Again, based on the user's quartile experience it gets a score. Then it checks if the user has a relevant degree type, or whether he has master's/PhD, and gives a score in accordance. At the end, a prediction is made by a decision boundary for the corresponding accumulated score of a user.

Graph: an example of how promotion chance is calculated.



- **Feature 4** tries to better connect the end user with employees or mentors that had already been through the progression track he is trying to get into, employees that advanced through the ranks of the company, which will undoubtedly have meaningful advice and insights for the user that is trying to get ahead in his company.

Graph: companies that have the highest amount of in-company promoting



AI methodologies:

Our project applies various algorithms to analyze career growth, find similar employees, and unify courses and certifications. To process course and certification data, we used *spaCy* for text analysis and fuzzy matching to merge similar course names, ensuring consistency across records. To identify the most relevant colleagues for consultation, we used SentenceTransformer *all-miniLM-L6-v2* to generate job title embeddings, followed by cosine similarity to rank and finding the closest matches. Promotion chances were estimated by analyzing employee career paths using quartile-based ranking, where we considered factors like education, experience, and other factors. For career growth tracking, we used historical employee data to detect patterns in promotions and applied network analysis to suggest employees for mentorship. Evaluation methods included *cosine similarity* scoring to verify the accuracy of employee matching, *mean absolute error (MAE)* to assess the reliability of promotion predictions, and precision-recall analysis to measure the effectiveness of course unification. By combining these techniques, our project provides structured career insights and helps employees make data-driven professional decisions.

Evaluation and Results:

- The evaluation process focused on measuring the accuracy and effectiveness of the selected features. For course and certification unification, we analyzed how well different course names and certifications were grouped under a single standardized format. This was done by manually reviewing the results to ensure that similar courses were correctly merged. In the employee similarity search, job titles were compared based on their textual and contextual similarities, and the effectiveness of the matching process was evaluated by checking how relevant the suggested colleagues were to the user's role. For promotion probability estimation, historical career data was analyzed to identify key factors influencing career advancement, such as education, experience, and past promotions. The accuracy of the predictions was assessed by comparing the estimated likelihood of promotion with actual career progression trends. Finally, career growth tracking was evaluated by analyzing past employee transitions within a company and verifying that the system correctly identified employees who successfully advanced in their roles.

- The evaluation led to several key findings that demonstrated the value of our approach. The course and certification unification successfully grouped similar learning programs, ensuring that employees could easily find relevant certifications without redundant listings. The employee similarity search provided meaningful recommendations, as most of the suggested colleagues held positions closely related to the user's role, making it easier to identify experts for consultation. Career growth tracking highlighted potential mentors who have followed similar career trajectories. These findings confirm that the project provides useful career insights that can support employees in making informed professional decisions.

Limitations and reflections:

One of the main challenges we faced during the project was data limitations, as we relied on publicly available information and had no direct access to LinkedIn's internal data, which restricted the depth of our analysis. Scraping skills data was not possible, as LinkedIn constrains didn't allow us to scrape this data. We wanted to gain insights from those skills, like we did in the courses and certifications. Standardizing course and certification names also required manual adjustments due to inconsistencies across providers. Matching employees by job title posed difficulties because roles were written differently across profiles, making it harder to identify truly similar positions. Time constraints limited our ability to fine-tune and test multiple models for promotion prediction, meaning our results were based on general trends rather than highly optimized algorithms. Additionally, computational resources were a constraint, as processing large datasets and running similarity comparisons required significant time and memory, sometimes slowing down development. Other challenges we faced was inconsistency in how information was written across different sources. Similar terms appeared in different variations, like "Senior Data Scientist" and "Senior DS," making it harder to group them correctly. This issue also applied to skills, courses, and certifications. To solve this, we had to normalize the data, unifying different variations under a single term to improve accuracy and consistency. Despite these challenges, the project successfully demonstrated how structured data analysis and career insights can assist professionals, though further refinements could improve accuracy and scalability. Moreover, we weren't able to compare our prediction of chance to promote, because the amount of employees in the data that actually promoted inside their company was relatively low and didn't provide meaningful conclusions.

Conclusion:

In conclusion, our project successfully developed a system that helps employees gain career insights by analyzing job titles, courses, certifications, and promotion trends. We managed to unify course and certification data, making it easier for users to identify relevant learning opportunities. Our approach to finding similar employees provided useful recommendations for consultation and mentorship. By analyzing career progression trends, we identified key factors influencing promotions, allowing users to estimate their chances of advancement. Despite challenges in data scraping, standardization, and resource limitations, we built a functional tool that highlights career growth patterns and helps professionals make informed decisions. This project demonstrates the potential of structured data analysis in improving career planning, with opportunities for further refinement and expansion.

Appendix

A. Data Sources and Collection Methods

A.1 Original Data Tables

- **Profiles Data Table:** Includes key columns such as:
  - ID
  - Experience
  - Education
  - Courses
  - Certifications
  - Current Company
  - Position

A.2 Additional Data Sources

- **Shanghai University Ranking Data:**
  - Collected **1,000 records**, each containing a university name and rank.
  - Used to calculate the "education score" for employees, which helps estimate their likelihood of career progression.
- **Layoffs Data:**
  - Scraped **500 records**, each representing a layoff event (e.g., company name, number of employees laid off, date).
  - Used to identify employees who successfully survived layoffs, providing insights into job stability.

B. Methodologies Used

B.1 Text Processing for Course and Certification Unification

- **spaCy NLP Library**:
  - Applied **Named Entity Recognition (NER)** and **text similarity techniques** to detect variations in course/certification names.
  - Standardized different names referring to the same course, improving data consistency.
- **Fuzzy String Matching (fuzzywuzzy library)**:
  - Merged closely related course names based on similarity scores.

B.2 Employee Similarity Matching

- **SentenceTransformer ("all-miniLM-L6-v2")**:
  - Transformed job titles into dense vector representations.

    - Applied **cosine similarity** to match employees with similar expertise.

B.3 Promotion Probability Estimation

- Used a **quartile-based ranking system**:
    - Compared the end-user's **experience and education** against those in the desired rank.
    - Assigned a **Low/Medium/High** promotion likelihood based on:
        - Position tenure
        - Education prestige (education score)
        - Degree type (Bachelor's, Master's, PhD)
    - Applied **decision boundaries** to make final predictions.

B.4 Career Growth Tracking

- **Network Analysis**:
    - Traced promotion patterns within a company.
    - Identified mentors who followed similar career paths.

C. Evaluation Metrics

C.1 Accuracy and Effectiveness Measures

- **Course and Certification Unification**:
    - Manually reviewed groupings to verify correct merging of course names.
- **Employee Similarity Search**:
    - Evaluated based on **relevance of suggested colleagues** to the user's role.
- **Promotion Probability Estimation**:
    - Assessed by comparing predicted promotion likelihoods with real-world career progression data.
- **Career Growth Tracking**:
    - Verified if identified mentors had meaningful career trajectories for the end-user.

C.2 Model Performance Metrics

- **Cosine Similarity Score**: Used for employee matching accuracy.
- **Mean Absolute Error (MAE)**: Measured reliability of promotion predictions.
- **Precision-Recall Analysis**: Evaluated effectiveness of course/certification unification.

D. Limitations and Future Work

D.1 Data Constraints

- **Limited access to LinkedIn's internal data** restricted our analysis depth.
- **Scraping constraints** prevented us from gathering skill-based insights.

D.2 Challenges in Standardization

- Job titles and courses had **inconsistent naming conventions** across sources.
- Required **manual data normalization** to improve accuracy.

D.3 Computational Limitations

- Processing **large datasets** and running **similarity comparisons** required substantial time and memory.
- Further **optimization and scaling** can enhance performance.

D.4 Future Improvements

- **Expand data collection** by integrating company-specific HR data (if available).
- **Improve NLP models** to refine course unification and job similarity matching.
- **Enhance prediction models** with more in-depth employee career trajectory insights.

E. Tools and Technologies Used

- **Programming Language**: Python
- **Libraries and Frameworks**:
  - **spaCy** (NLP processing)
  - **FuzzyWuzzy** (string matching)
  - **SentenceTransformer** ("all-miniLM-L6-v2" for embeddings)
  - **Scikit-learn** (cosine similarity, classification models)
  - **Pandas & NumPy** (data manipulation and processing)
  - **NetworkX** (career growth analysis)
- **Scraping Tools**:
  - **BeautifulSoup & Selenium** (for Shanghai ranking and layoffs data)
- **Data Storage**:
  - CSV files and SQL database for structured storage