

Job Recommendation System Using Natural Language Understanding

Akash Poddar
Avani Rao

Abstract

The Job Recommendation System using Natural Language Understanding aims to provide personalized job recommendations tailored to users' skills, interests, and career goals. Leveraging machine learning techniques and natural language understanding, the system analyzes job descriptions and user preferences via resumes to generate accurate matches. Key factors such as job titles, salary estimates, company ratings, and locations are incorporated to enhance recommendation relevance. This project bridges the gap between job seekers and opportunities, offering a data-driven, user-centric approach to streamline the job search process and support informed career decisions.

1 Introduction

The process of searching for the ideal job can be overwhelming, given the multitude of opportunities and the variety of factors that need to be considered. Traditional job search platforms often fail to adequately address individual preferences, making it challenging for users to identify roles that align with their unique skills, interests, and career aspirations. To overcome these challenges, the **Job Recommendation System using Natural Language Understanding** leverages machine learning and natural language processing techniques to streamline the job search process.

This system starts with data scraping from sources like Indeed using Apify API, capturing critical job-related information such as titles, salary estimates, company ratings, locations, and industries. Through robust feature engineering techniques, raw data is preprocessed and transformed into a format suitable for model training. Techniques such as handling missing data, encoding categorical variables, and feature scaling ensure data quality and consistency.

A key component of this system is the utilization of the TF-IDF (Term Frequency-Inverse Document Frequency) vectorization technique to process job descriptions and user preferences

into numerical feature vectors. This enables the system to analyze textual data effectively and identify patterns in user preferences. Coupled with the Nearest Neighbors algorithm, the system delivers highly relevant job recommendations based on the proximity of user preferences to job descriptions. Furthermore, skill extraction, facilitated by the Spacy library, enriches the recommendation process by identifying specific competencies within user-provided resumes.

To enhance accessibility and usability, the system integrates with a Streamlit-based web application, allowing users to interact with the recommendation system through a simple and intuitive interface. Users can upload their resumes, and the application processes the input, applies the trained models, and presents a curated list of job recommendations.

This paper details the methodologies, and outcomes of the **Job Recommendation System using Natural Language Understanding**, demonstrating how this innovative approach bridges the gap between job seekers and suitable opportunities. By offering personalized and data-driven recommendations, the system aims to revolutionize the job search experience and support informed career decisions.

2 Methods

This section outlines the methodologies employed to develop the **Job Recommendation System using Natural Language Understanding**, detailing data collection, preprocessing, modeling, and implementation strategies.

2.1 Data Collection and Scrapping

To build a robust job recommendation system, data was scraped from Indeed using Apify, capturing essential job-related attributes such as job titles, salary estimates, job descriptions,

company ratings, locations, and industries. Apify API is a cloud platform for web scrapping and extracting data, using which 5000 job postings for 30+ technical roles were scrapped. It provides a rich dataset for understanding job roles and user preferences. The scraped data formed the foundation for subsequent preprocessing and modeling steps, ensuring the system's ability to deliver relevant and accurate job recommendations in the real world.

A	B	C	D	E
company	ROLE	description	id	jobType
ACORD Solutions Group	Business Analyst	Business Process Analyst	40.0	Full-time
FICO	Business Analyst	Architectural Service Management	126000.0	Full-time
ation of The City University of New York	Business Analyst	IT Business		Full-time
Adobe	Business Analyst	Software		Full-time
Hendrick Autoguard	Business Analyst	Software		Full-time
ACORD Solutions Group	Business Analyst	Software		Full-time
ACORD Solutions Group	Business Analyst	Software		Full-time

F	G	H	I	J
location	positionName	rating	reviewsCount	salary
Cambridge, MA	Business Process Analyst			\$40 per class
Remote	Architectural Service Management - Business Analyst	3.6	201	\$98,000 - \$154,000 a
New York, NY	IT Business Analyst	4.2	135	\$95,000 - \$115,000 a
San Jose, CA 95121	Business System Analyst, Software	4.3	840	\$61,600 - \$128,800 a
San Francisco Blvd, CA	Business & Systems Analyst (Autoguard)	3.7	840	
Cambridge, MA	Senior Business Systems Analyst			\$40 per class
Philadelphia, PA	Business Analyst			\$51,824 - \$90,000 a
United States	Business Operations Analyst			

Figure 1: Scrapped Data from Indeed

2.2 Feature Engineering

The raw data underwent preprocessing through feature engineering to prepare it for machine learning applications. Missing values were handled by either imputation or removal, ensuring data completeness. All text data was converted to lowercase to ensure consistency and eliminate case sensitivity during analysis. Irrelevant characters and numerical values were removed to clean the textual data, focusing only on meaningful words. Sentences were broken down into individual words or tokens, enabling fine-grained text analysis. Commonly used words (e.g., "the," "and") that do not contribute to the meaning of the text were filtered out to enhance data relevance. Words were reduced to their root forms (e.g., "running" to "run"), ensuring consistency and reducing redundancy in the text data. Numerical features, such as salary estimates and company ratings, were examined for outliers. Techniques like capping or transformation were applied to minimize their influence on the model, ensuring stable and unbiased performance.

By integrating these steps, the feature engineering process optimized the dataset in the

csv format for machine learning, enabling accurate and personalized job recommendations.

COMPANY	ROLE	POSITIONNAME	SALARY_NUMERIC	RATING	REVIEWSCOUNT	CLEANED DESCRIPTION
ACORD Solutions Group	Business Analyst	Business Process Analyst	40.0			detail posted
FICO	Business Analyst	Architectural Service Management	126000.0			
Research Foundation of The City University of New York	Business Analyst	IT Business				
Adobe	Business Analyst	Software				

Figure 2: Cleaned Descriptions

2.3 Machine Learning Model Development

The system employs TF-IDF (Term Frequency-Inverse Document Frequency) vectorization to transform job descriptions and user preferences into numerical feature vectors. This representation assigns higher values to words that are more significant in a document while downplaying commonly occurring words across all documents, enhancing the precision of text analysis. It enables a nuanced understanding of the relevance of terms within a document, rather than merely assigning binary values of 1 or 0 based on word presence.

Using these TF-IDF vectors, the Nearest Neighbors algorithm identifies the closest matches between user inputs (e.g., resume content) and job descriptions. This approach generates a ranked list of personalized job recommendations by calculating the similarity between the user's skills and the job requirements. The modeling pipeline prioritizes accuracy and relevance, ensuring that the recommendations align effectively with user expectations and career goals.

```
Top 10 descriptions with least distance from KNN are:
[0.77619845 0.80297736 0.80445242 0.80616067 0.81160674 0.81410045
 0.81613072 0.81726931 0.81730875 0.81918032]

Top 10 matching descriptions are:
1115 analytica seeking junior data scientist remote...
1575 softrams one fastest growing digital service f...
2625 looking problem solver innovator dreamer searc...
621 year professional military experience year sql...
1141 location hybrid work model van wert oh data sc...
1597 detail posted sep location altamonte spring fl...
1137 requirement data scientist hawaii hawaii ons...
1121 analytica seeking senior data scientist remote...
620 amentum seeking mid data engineer rdte support...
608 job description amentum seeking mid data engin...
Name: CLEANED DESCRIPTION, dtype: object
```

Figure 3: Distance between Resume and Job Descriptions vectors using KNN algorithm.

2.4 Skills Extraction

To facilitate skill extraction, the system employs the SpaCy library along with a

predefined skill set loaded from a CSV file. A custom function integrates SpaCy's **Matcher**, which uses pattern dictionaries to identify relevant skills in the text. Patterns are dynamically created from the skill list, enabling the matcher to efficiently detect matches in a given document. The text extraction process involves parsing PDF resumes using PyPDF2, extracting content from all pages into a unified text representation. The extracted text is processed through SpaCy's natural language processing pipeline to detect and extract skills. This approach ensures accurate identification of both technical and soft skills, which are integral for generating personalized job recommendations. The implementation emphasizes flexibility, allowing seamless updates to the skill list for evolving requirements.



	A
1	technical skills
2	python
3	django
4	algorithms
5	visualization
6	machine learning
7	opencv
8	natural language processing
9	tensorflow
10	flask
11	pandas
12	implementation
13	deploy
14	programming

Figure 4: Extracted Skills file.

2.5 Streamlit Application

To provide a user-friendly interface, a Streamlit-based web application was developed as a front-end for the job recommendation system. Users can upload their resumes in PDF format, specify the number of job recommendations they wish to receive, and download the results in CSV format for further analysis. Upon uploading a resume, the application processes the file to extract relevant skills using a skills extraction module. These skills are vectorized using the TF-IDF (Term Frequency-Inverse Document Frequency) technique to create feature representations.

The system employs a Nearest Neighbors algorithm to identify the closest matches between the extracted resume skills and a preprocessed job description dataset. The recommendations are ranked based on their cosine similarity scores, ensuring alignment

between the user's profile and job requirements. The interactive application also provides progress updates during the analysis, enhancing user experience. By integrating these functionalities, the Streamlit application makes the job recommendation process intuitive, accessible, and tailored to user preferences.

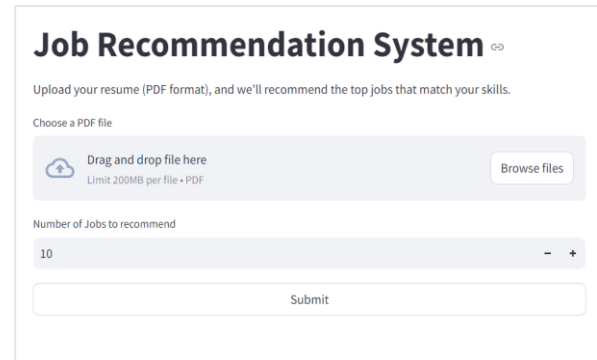


Figure 5: Streamlit application for the user.

2.6 Deployment on Google Cloud Platform

After successfully implementing the Streamlit application, the next phase involved hosting it on Google Cloud Platform (GCP) to enable global accessibility. The deployment process included packaging the application in a Docker container, uploading it to the Google Container Registry, and deploying it on GCP. An external IP address was configured to ensure seamless online access for users.

User can access the deployed application to get the best matched jobs based on their inputted resumes from the following GCP link: <https://streamlit-app-961572577028.us-east1.run.app/>

3 Results

The job recommendation system analyzed the uploaded resume ("Resume.pdf") and generated a ranked list of the top 10 recommended positions tailored to the user's skills and experience. The recommendations were based on match scores computed by the system, alongside additional criteria such as job ratings and available salary information. The system identified roles primarily in data science, analytics, and engineering, highlighting a strong alignment with the user's expertise. Positions with high match scores (ranging from 0.78 to

0.83) included roles such as Senior Data Scientist, Senior Data Analyst, and Mid Data Engineer, reflecting versatility and a diverse skill set. Notably, the highest match score achieved was 0.83, corresponding to positions like “Senior Data Scientist, Clinical Analytics” and “Mid Data Engineer.” Salary data, available for select roles, provided an additional perspective for prioritization, with the highest listed salary being \$160,050 for the “Sr. Business Intelligence Engineer” position. The system performed efficiently, generating recommendations in real time without errors, and offered users the option to download results as a PDF file. The user-friendly interface further facilitated seamless interaction, enabling quick analysis and easy retrieval of outputs from anywhere around the world using GCP was successfully deployed.

✓ Analysis complete!

Recommended Top 10 Jobs:

	COMPANY	ROLE	POSITIONNAME
1	Analytica	Data Scientist	Data Scientist
2	Alteryx, Inc.	Quality Engineer	Sr. Business Analyst (Data Engineering & Quality Oper
3	ACORD Solutions Group	Data Analyst	Health Plan Data Analyst
4	Softrams	Data Analyst	Data Analyst - Data Analytics
5	Amazon.com Services LLC	Data Engineer	Sr. Business Intelligence Engineer, Measurement, Ad T
6	Central Mutual Insurance Company	Data Scientist	Senior Data Analyst
7	Amentum	Data Engineer	Mid Data Engineer
8	Analytica	Data Scientist	Senior Data Scientist
9	Allergan Aesthetics	Data Scientist	Senior Data Scientist, Clinical Analytics
10	Amentum	Data Engineer	Data Engineer- Mid

✓ Analysis complete!

Recommended Top 10 Jobs:

		POSITIONNAME	SALARY_NUMERIC	RATING	match
1	ntist	Data Scientist	None	3.8	0.78
2	ngineer	Sr. Business Analyst (Data Engineering & Quality Operations)	136,200	3.7	0.8
3	yst	Health Plan Data Analyst	None	None	0.81
4	yst	Data Analyst - Data Analytics	None	3.8	0.81
5	neer	Sr. Business Intelligence Engineer, Measurement, Ad Tech, and	160,050	3.5	0.81
6	ntist	Senior Data Analyst	None	3.6	0.81
7	neer	Mid Data Engineer	None	3.8	0.82
8	ntist	Senior Data Scientist	None	3.8	0.82
9	ntist	Senior Data Scientist, Clinical Analytics	None	3.8	0.82
10	neer	Data Engineer- Mid	None	3.8	0.82

Figure 6: Recommended Jobs with Match Score

4 Discussion and Analysis

In this study, we developed a job recommendation system designed to identify the best-matching positions for users based on their resumes. The system employed a TF-IDF-based approach to extract and match keywords

from resumes and job descriptions. This methodology provided a robust framework for performing exact term matching, ensuring that the recommended jobs were tailored to the specific terms and skills explicitly mentioned in the resume.

One of the primary strengths of the TF-IDF model lies in its ability to prioritize unique and significant terms by assigning higher weights to rare yet relevant words. This approach ensures that the system can distinguish between closely related terms, such as "Python" and "Power BI," or "visualization" and "Seaborn," treating each as distinct entities. This exact matching capability allowed the system to make highly targeted recommendations, directly reflecting the user's explicitly stated skills and experience.

However, this approach also comes with limitations. TF-IDF does not inherently understand semantic relationships or contextual similarities between terms. For instance, it cannot infer that "Python" and "programming" are conceptually related or that "visualization" tools like "Seaborn" and "Power BI" serve similar functions in different contexts. Consequently, any variation in terminology between the resume and job descriptions could lead to mismatches or overlooked opportunities.

In contrast, employing advanced embedding techniques, such as Word2Vec, neural network-based embeddings, or large language models (LLMs), would offer a more nuanced understanding of the textual data. These methods capture semantic relationships and contextual meanings, enabling the system to recognize that terms like "Python" and "programming" or "visualization" and "Seaborn/Power BI" are related. While this would make the system more flexible and capable of generalizing across variations in phrasing, it may reduce the precision of exact matches, potentially introducing noise into the recommendations.

By focusing on a TF-IDF-based approach, our system ensures that only explicitly stated skills are considered for job matching, which is particularly valuable for scenarios where users wish to emphasize specific expertise or certifications. Future iterations could combine

the strengths of TF-IDF and embedding-based techniques, leveraging exact matching for critical terms while incorporating semantic understanding for broader contextual relevance. This hybrid approach could strike a balance between precision and flexibility, enhancing the overall effectiveness of the recommendation system.

5 Conclusion and Future work

This research focused on the development and application of a job recommendation system that utilizes advanced algorithms to match candidate profiles with suitable job opportunities. By analyzing the loaded sample resume of a user ("Resume.pdf") having data science experience, the system successfully identified and ranked the top 10 recommended positions based on match scores, job ratings, and salary information. The recommendations showcased strong alignment with the user's skills, education, and professional experience, suggesting the system's effectiveness in recognizing nuanced details in the resume and correlating them with relevant job descriptions. The system achieved a high degree of precision, with match scores ranging from 0.78 to 0.83, highlighting its ability to cater to diverse job roles, such as Data Scientist, Senior Data Analyst, and Data Engineer. Furthermore, the system demonstrated operational efficiency, completing the analysis in real time without errors and providing users with a seamless experience through an intuitive interface.

The findings suggest that the system can significantly streamline the job search process by providing tailored recommendations, empowering candidates to focus on roles that best align with their qualifications and aspirations. Additionally, the incorporation of salary and job rating information enhances decision-making by offering a holistic view of potential opportunities.

While the job recommendation system has demonstrated strong results, future enhancements can further improve its performance and usability. Integrating large language models (LLMs) like GPT-based

systems can enhance skill extraction, semantic analysis of job descriptions, and personalized recommendations. Incorporating real-time labor market trends and user feedback loops will refine match accuracy and responsiveness. Expanding the dataset to cover diverse industries, adding multilingual support, and enabling career progression analysis will make the system more inclusive and versatile. Finally, features like skill gap analysis and tailored learning resource suggestions can position the system as a comprehensive career advisory tool.

Ethical Statement

Current job recommendation systems often rely on context-based models or use broad categorizations that match job seekers with roles based on overall skill sets or visualized data such as job descriptions, company attributes, and other contextual information. These systems may give higher priority to generalized skills or keywords, sometimes overlooking specific skills listed on resumes. For example, a system that emphasizes "programming" might match candidates with roles requiring coding experience, but it may not differentiate between specific programming languages like Python, Java, or C++.

In contrast, our Job Recommendation System utilizes the **TF-IDF (Term Frequency-Inverse Document Frequency)** technique, which focuses on the exact words and their frequency within both the job descriptions and resumes. Instead of relying on contextual matching or visualizations, it directly compares the occurrence of words, such as "Python," and matches them with job descriptions that explicitly mention Python, ensuring the job recommendations are more tailored to the candidate's precise skills. This method helps both job seekers and recruiters by ensuring a higher degree of accuracy in matching specific competencies (e.g., "Python") with job requirements, rather than simply using broad, generalized keywords. It eliminates the ambiguity in matching by giving importance to the exact frequency of words, allowing for more precise and relevant matches based on the actual content of resumes and job descriptions.

Acknowledgments

The authors, Akash Poddar and Avani Rao, extend their sincere gratitude to Professor Kasia Hitczenko and the SEAS Department at The George Washington University for their invaluable support and for providing access to the experimental research facilities that made this work possible.

References

- Chandraghandi S, Shilpa S, Anamika P, Kamalakkannan R, Santhoshsivan N. 2022.** [Resume Screening Using TF-IDF](#). *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 11, Issue 5, pages 720. DOI: 10.17148/IJARCCE.2022.115166.
- Modak S, Shinde P, Tiwari A, Nalamwar S. 2024.** [A Review of Resume Analysis and Job Description Matching Using Machine Learning](#). *International Journal on Recent and Innovation Trends in Computing and Communication*, Vol. 12, No. 2.
- Maheshwary S, Misra H. 2018.** [Matching Resumes to Jobs via Deep Siamese Network](#). *WWW '18: Proceedings of The Web Conference 2018*, Lyon, France.
- Li C, Fisher E, Thomas R, Pittard S, Hertzberg V, Choi JD. 2020.** [Competence-Level Prediction and Resume & Job Description Matching Using Context-Aware Transformer Models](#). *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, November 2020.