

Dataset:

The Job Posting Dataset is a robust collection of diverse professional opportunities, rich in details such as Job Titles, Required Qualifications, Preferred Skills, and Salary Ranges. It spans various sectors, offering a comprehensive view of industry-specific roles and requirements. Key entities like Job ID, Job Description, and Employment Type undergo meticulous parsing and normalization, laying a structured foundation for in-depth data analyses. Employing advanced data mining and NLP techniques, the dataset is refined for accuracy, feeding into algorithms for market trend analysis, demand prediction, and skill gap identification.

The Job Seekers Dataset, as exemplified by the extensive 70k Job Applicants Data from Kaggle, is a critical component in creating a holistic knowledge graph. This dataset assembles a diverse range of candidate profiles, each a complex mix of Skills, Professional Experiences, Educational Backgrounds, Certifications, and Employability Scores. Essential entities like Candidate ID, Professional Achievements, Academic Qualifications, and Geographical Preferences are systematically extracted and standardized using advanced NLP and machine learning methodologies. This structured transformation is key to assimilating varied candidate data into the knowledge graph, allowing for a nuanced representation of the job market. The granularity of this dataset lays a fertile ground for sophisticated predictive analytics, significantly boosting the job recommendation system's precision in aligning candidate profiles with relevant jobs.

Dataset Size:

Currently, our approach involves working with a subset of the dataset, specifically focusing on a selection of 1000 candidate entries and 1000 job entries. This strategic subset allows for efficient processing and analysis, striking a balance between computational efficiency and the representation of diverse candidate and job profiles.

Files:

Job_descriptions.csv File: The "Job_descriptions.csv" file serves as a comprehensive repository of job-related information, encompassing details such as Job ID, job title, required skills, required qualifications, necessary experience, and location specifics. Each of these details is thoughtfully organised into separate columns, featuring preprocessed data that eliminates empty cells or dummy values.

Resume Processed.csv File: The "Resume Processed.csv" file is dedicated to capturing candidate information in a structured manner. It includes essential details such as Candidate ID, skills, qualifications, and the candidate's professional experience. This data is thoughtfully preprocessed, adhering to a meticulous approach that excludes empty cells or placeholder values.

Additional files contain skills information, specifically separated from the two main files, to meet the specific requirements of Neo4j. In adherence to Neo4j's structure, the skills are meticulously processed to ensure that duplicates are eliminated, and each skill is appropriately categorized under its respective node type.