Data Collection and Preprocessing Phase

Date	Nov 2024
Team ID	Team-739663
Project Title	AI-Enabled Candidate Resume Screening using NLP
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Template

Elevate your talent acquisition strategy with the Resume Data Collection Plan and the Raw Resume & Job Description Sources Report, ensuring meticulous data curation and integrity for accurate candidate screening, efficient recruitment, and informed hiring decisions.

Data Collection Plan Template

Section	Description
Project Overview	To develop an AI-powered resume screening system that leverages Natural Language Processing (NLP) to automatically parse, analyze, and evaluate candidate resumes against job descriptions. The system aims to streamline and enhance the recruitment process by providing accurate and efficient candidate shortlisting.
Data Collection Plan	To effectively train and build an NLP-based resume screening system, a comprehensive data collection strategy is essential. This includes gathering resumes and job descriptions from diverse sources to cover a wide range of formats, industries, and skillsets, ensuring the model's robustness and generalizability.
Raw Data Sources Identified	For resume screening, raw data sources include publicly available resume datasets, job description data from job portals (e.g., Indeed, LinkedIn), internal HR databases, and user-uploaded resumes. These sources provide the necessary input for training the NLP models to extract and match candidate qualifications effectively.

Raw Data Sources Template

Source Name	Description	Location	Format	Access Permissions
Resume Dataset	Contains a collection of candidate resumes used for training and testing NLP-based screening models.	NA	PDF, DOCX, TXT	Public / Internal
Job Description Dataset	Includes job descriptions across various roles to be matched with candidate profiles.	NA	CSV / Text	Public / Internal
Parsed Resume JSON	Output from resume parsers in structured JSON format for model processing.	NA	JSON	Internal
Labeled Training Data	Annotated data used for supervised learning to train entity recognition (e.g., skills, education, experience).	NA	CSV / JSON	Internal