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 Quality Report Template

The Data Quality Report Template will summarize data quality issues from the selected sources related to resumes and job descriptions. It includes severity levels and resolution plans, helping to systematically identify, assess, and rectify data inconsistencies for building an effective resume screening system.

Data Source: Customer Order Database

Data Source	Data Quality Issue	Severity	Resolution Plan
Resumes (PDF, DOCX, TXT)	Inconsistent formats, missing fields, unstructured data	Medium	Develop a standardized data extraction pipeline to normalize formats and extract relevant information.
Job Descriptions (Internal HR Systems)	Redundant text, vague language, inconsistent terminology	High	Implement robust text cleaning techniques (e.g., stop word removal, stemming, lemmatization) and sentiment analysis to filter out irrelevant and biased information.
Manually Collected Data	Typographical errors, data entry mistakes	Medium	Regularly update the knowledge base and incorporate real-world examples to improve the chatbot's relevance.

Data Source	Data Quality Issue	Severity	Resolution Plan
Resumes (PDF, DOCX, TXT)	Inconsistent formats, missing fields, unstructured data	Medium	Develop a standardized data extraction pipeline to normalize formats and extract relevant information.
Job Descriptions (Internal HR Systems)	Redundant text, vague language, inconsistent terminology	High	Implement robust text cleaning techniques (e.g., stop word removal, stemming, lemmatization) and sentiment analysis to filter out irrelevant and biased information.
Manually Collected Data	Typographical errors, data entry mistakes	Medium	Regularly update the knowledge base and incorporate real-world examples to improve the chatbot's relevance.
Public Resume Datasets	Outdated or irrelevant information	Medium	Filter and curate the dataset to retain only relevant and recent resumes. Remove duplicates and outdated job history entries.
Company HR Database	Inconsistent labeling, incomplete candidate profiles	High	Apply label encoding and implement imputatio field naming conventions.

Data Source:

Data Source	Data Quality Issue	Severity	Resolution Plan
Resume Files (PDF, DOCX, TXT)	Inconsistent formatting, unstructured data, missing sections	High	Build an NLP-based resume parser to standardize formats and extract relevant sections (e.g., education, skills, experience). Handle missing fields with rules.
Job Description Texts	Ambiguous language, lack of structure, overlapping job roles	Medium	Use keyword extraction, text classification, and POS tagging to extract responsibilities and required skills accurately.
Public Resume Datasets	Duplicate entries, outdated content	Medium	Perform deduplication and filter resumes based on recency and relevance. Integrate logic to prioritize updated resumes.
HR Database (Internal)	Inconsistent labeling, missing candidate metadata	High	Apply label normalization and develop imputation strategies (mean/mode) for missing fields like skills or experience.
User-uploaded Resumes	Typographical errors, varied file encodings	High	Implement encoding standardization, spell correction, and robust text cleaning using NLP preprocessing techniques.