Data Collection and Preprocessing Phase

Date	12 July 2024
Team ID	SWTID1720161415
Project Title	JobSwift : Accelerating Careers with AI Powered Applications
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Template

Elevate your data strategy with the Data Collection plan and the Raw Data Sources report, ensuring meticulous data curation and integrity for informed decision-making in every analysis and decision-making endeavor.

Data Collection Plan Template

Section	Description
Project Overview	This machine learning project aims to develop an AI-powered job application platform using Google's PaLM text-bison-001 model. The objectives are to streamline the job application process by generating resumes, cover letters, and interview questions tailored to the user's profile. The project will leverage generative AI to provide personalized and efficient career advancement tools.
Data Collection Plan	Data for this project will be collected from various sources to ensure comprehensive coverage of job-related information, including resumes, cover letters, job descriptions, and user profiles. The sources include publicly available datasets, web scraping from job portals, user-generated content, and synthetic data generated through controlled prompts.
Raw Data Sources Identified	The following raw data sources have been identified for this project:

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Dataset 1	This dataset includes a variety of resumes which can be used for training and evaluating models focused on resume parsing and classification.	https://www.kaggle.com/datasets/jillanisofttech/updated-resume-dataset	CSV	Variable	Public
Dataset 2	This dataset is designed for Named Entity Recognition (NER) in resumes, containing annotated entities within resume texts.	https://www.kaggle.com/datasets/dataturks/resume-entities-for-ner	CSV	Variable	Public