Data Collection and Preprocessing Phase

Date	16 July 2024
Team ID	
Project Title	CodeXchange: An AI-Powered Code Translator Tool using Palm's chat-bison-001
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	CodeXchange leverages machine learning (ML) to facilitate the translation of code snippets across different programming languages. It utilizes natural language processing (NLP) and deep learning techniques to understand and convert code from one language to another seamlessly.
Data Collection Plan	1. Gather data that represents a wide variety of code snippets in different programming languages to train the translation models effectively. 2. Collect data on user interactions with the platform to improve user experience and provide personalized recommendations.
Raw Data Sources Identified	The raw data sources include datasets obtained from reputable repositories and platforms focused on programming languages and code examples. These datasets represent a subset of the collected information, encompassing variables such as programming language, code structure, syntax, and usage patterns. This data is crucial for training the AI model to accurately translate and interpret code across different languages and frameworks. .

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle dataset	The dataset encompasses a wide range of languages such as Python, Java, JavaScript, C++, and more, along with associated metadata like usage contexts, frameworks, and comments.	https://www.kaggle.com/datasets/sreejithravindran7/codexchange	CSV	654 KB	Public

Dataset 1	Description of the data in this source.	Link of Dataset 1	CSV	XX GB	Public
Dataset 2	Description of the data in this source.	Link of Dataset 2	Image	YY GB	Private (with access)
…	…	…	…	…	…