Data Collection and Preprocessing Phase

Date	17th July 2024
Team ID	SWTID1720025517
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

Project Overview: This machine learning project aims to develop an AI-powered platform, CodeXchange, to translate code from one programming language to another using Google's PaLM text-bison-001 model. The objective is to streamline the process of code translation by generating equivalent code snippets in the target language, ensuring syntactical and functional correctness.

Data Collection Plan: Data for this project will be collected from various sources to ensure comprehensive coverage of different programming languages and coding styles. The sources include publicly available code repositories, datasets from coding platforms, user-generated code snippets, and synthetic code examples generated through controlled prompts.

Raw Data Sources Identified:

Source Name	Description	Location/URL	Format	Size	Access Permissions
Dataset 1	This dataset includes a variety of code snippets in multiple programming languages, which can be used for training and evaluating the code translation model.	Kaggle Code Snippets Dataset	CSV	Variable	Public
Dataset 2	This dataset contains annotated code pairs in different programming languages, useful for supervised learning and model evaluation.	Kaggle Code Translation Pairs	CSV	Variable	Public