Data Collection and Preprocessing Phase

Date	7 November 2024
Team ID	739939
Project Title	Image Caption Generator
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 deep learning project aims to build a model that can automatically generate descriptive captions for input images.
Data Collection Plan	The data for this project will be collected from publicly available image captioning datasets (e.g., MS COCO, Flickr30k) and potentially curated image collections with corresponding textual descriptions.
Raw Data Sources Identified	The primary raw data sources include large-scale image captioning datasets such as the Kaggle dataset, which contains images of complex everyday scenes with multiple object annotations and captions, and the Flickr30k dataset, which provides a large collection of Flickr images with human-generated captions.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle	The Kaggle Flickr8k Dataset consists of 8000 datasets and captions.txt file.	https://www.kaggle.com/datasets/adityajn105/flickr8k	Image, txt file	3.32 MB	Public