Data Collection and Preprocessing Phase

Date	27 May 2025
Team ID	SWUID20240006489
Project Title	Gemini Decode: Multilanguage Document Extraction by Gemini Pro
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification:

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:

Section	Description
Project Overview	The GeminiDecode project aims to develop a General Artificial Intelligence (GenAI) model named Gemini Pro, designed to accurately extract and interpret information from documents written in multiple languages. The primary objective is to enhance document processing efficiency and accessibility in multilingual environments, thus facilitating seamless information retrieval and decision-making across different sectors.

Data Collection Plan	1. **Academic Databases:** Online repositories containing research papers, theses, and dissertations in multiple languages. 2. **Online Libraries:** Digital libraries offering a wide range of books, articles, and periodicals in various languages. 3. **Government Portals:** Official websites providing public documents, reports, and records. 4. **Corporate Archives:** Internal document repositories from partnering organizations, including reports, manuals, and memos.

	5. **Public Datasets:** Open-access datasets available on platforms like Kaggle and UCI Machine Learning Repository, containing various types of documents.
Raw Data Sources Identified	1. **Multilingual Research Papers:** A collection of research papers in multiple languages sourced from academic databases.
	2. **Online Library Collections:** Digital books and articles in various languages sourced from online libraries.
	3. **Government Documents:** Official documents and records available on government portals.
	4. **Corporate Archives:** Internal documents from partnering organizations, including reports, manuals, and memos.

Source Name	Description	Location/URL	Format	Size	Access Permissions
Dataset 1	Multilingual Research Papers	https://tinyurl.com/49fbcrvh	DOCX	30 GB	Public

	5. Public Datasets: Open-access datasets available on platforms like Kaggle and UCI Machine Learning Repository, containing various types of documents.

Raw Data Sources:

Dataset 2	Online Library Collections	https://tinyurl.com/49fbcrvh	DOCX	20 GB	Public
Dataset 3	Government Documents	https://tinyurl.com/49fbcrvh	DOCX	15 GB	Public
Dataset 4	Corporate Archives	https://tinyurl.com/49fbcrvh	DOCX	10 GB	Private (with access)
Dataset 5	Public Datasets	https://tinyurl.com/49fbcrvh	DOCX	25 GB	Public