Smart Internz

SMARTBRIDGE
Let's Bridge the Gap
a Veranda Enterprise

Data Collection and Preprocessing Phase

Date	June 2024
Team ID	740103
Project Title	The Language Of Youtube: A Text Classification Approach To Video Descriptions
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Template

Section	Description

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

Project Overview	**To develop a machine learning model using natural language processing to classify YouTube video descriptions into different categories, providing insights into language and trends of YouTube to optimize advertising strategies.**
Data Collection Plan	Data will be collected by Using YouTube Data API to fetch video details and descriptions.
Raw Data Sources Identified	Gathering data from * **YouTube Data API**: The primary source for extracting YouTube video descriptions, comments, and other metadata. * **Web Scraping**: As an alternative, use web scraping techniques to gather data directly from YouTube web pages if API limits are restrictive.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Access Permissions
Dataset	In this project we have used .csv data. This data is downloaded from kaggle.com.	Youtube Videos Dataset (~3400 videos) (kaggle.com)	CSV	Public