Data Collection and Preprocessing Phase

Date	15 October 2024
Team ID	739743
Project Title	Spooky Author Identification Using Deep Learning
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Template

Creating a data collection plan involves several key steps to ensure that the data gathered is reliable , relevant, and suitable for analysis.

Data Collection Plan Template

Section	Description

Project Overview	This project utilizes deep learning techniques to identify the author of a given text passage from three renowned spooky authors: Edgar Allan Poe, Mary Shelley, and Bram Stoker. By analyzing the unique writing styles and linguistic features of each author, the project aims to accurately distinguish between them, demonstrating the potential of deep learning in literary analysis and authorship identification.
Data Collection Plan	The data collection plan for Spooky Author Identification involves gathering 1000 text passages from Edgar Allan Poe, Mary Shelley, and Bram Stoker. Passages will be collected from literary websites, e-books, and online archives, then pre-processed and split into training, validation, and testing sets.
In Raw Data Sources Identified	**Raw Data Sources Identification:** 1. **Kaggle Datasets** - Pre-compiled datasets specific to Spooky author identification.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle Dataset	This dataset consist of information regarding Spooky author identification	https://www.kaggle.com/competitions/spooky-author-identification/data	CSV	3 GB	Public