**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| **Date** | 11 March 2025 |
| **Team ID** | 740052 |
| **Project Title** | AI-Based Intelligent Insight Extractor |
| **Maximum Marks** | 6 Marks |

**Data Exploration and Preprocessing Template:**

Data exploration and preprocessing for a text summarization project begin with thorough exploratory data analysis (EDA) to understand the structure, distribution, and relationships within the textual data. Handling missing or incomplete text entries is crucial, often addressed through techniques like data imputation or removal. Text data is cleaned and normalized through processes such as lowercasing, removing stop words, and stemming or lemmatization. Tokenization and vectorization (e.g., TF-IDF or word embeddings) convert text into numerical formats suitable for machine learning models. These steps ensure the dataset is well-prepared for training accurate summarization models.

|  |  |
| --- | --- |
| **Section** | **Description** |
| Data Overview |  |
| Word Frequencies |  |
| Word Tokenization |  |
| Normalization |  |
| Sentence Tokenization |  |
| **Data Preprocessing Code Screenshots** | |
| Loading Data |  |
| Load Spacy Pipeline |  |
| Text Transformation |  |
| Feature Engineering | Attached the codes in final submission. |
| Save Processed Data | - |