Data Collection and Preprocessing Phase

Date	Nov 2024
Team ID	Team-739662
Project Title	Chatbot based on Data Science Enquiry using NLP
Maximum Marks	2 Marks

Data Quality Report Template

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

Data Source: Customer Order Database

Data Source	Data Quality Issue	Severity	Resolution Plan
Research Papers and Articles	Inconsistent Formatting, Lack of Standardization	Medium	Develop a standardized data extraction pipeline to normalize formats and extract relevant information.
Online Forums and Q&A Platforms	Noisy Data, Subjectivity, and Bias	High	Implement robust text cleaning techniques (e.g., stop word removal, stemming, lemmatization) and sentiment analysis to filter out irrelevant and biased information.
Textbooks and Tutorials	Outdated Information, Lack of Practical Examples	Medium	Regularly update the knowledge base and incorporate real-world examples to improve the chatbot's relevance.

Data Source:

Data Source	Data Quality Issue	Severity	Resolution Plan
Internal Company Data	Incomplete Data, Missing Values, Inconsistent Labelling	High	Develop data imputation techniques (e.g., mean, median, mode imputation) and data cleaning pipelines to handle missing and inconsistent data.
Text Files	Unstructured data	High	Implement Natural Language Processing (NLP) techniques such as text preprocessing, tokenization, and named entity recognition to extract relevant information from unstructured data.
Database	Outdated/inconsistent data	Low	Schedule regular data updates and perform data normalization to ensure consistency. Implement data validation rules to detect and correct inconsistencies.