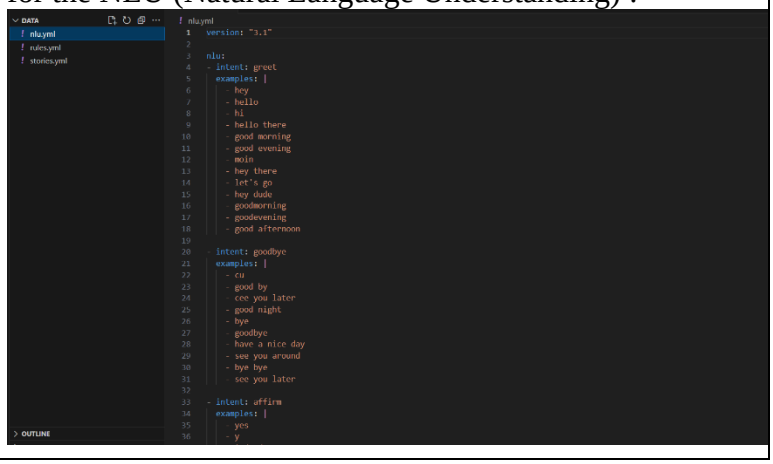
Data Collection and Preprocessing Phase

Date	Nov 2024
Team ID	Team-739662
Project Title	Chatbot based on Data Science Enquiry using NLP
Maximum Marks	6 Marks

PreparationTemplate

The data collection and preparation phase for the chatbot based on the data science enquiry project using NLP involves several steps. Firstly, text data is collected from various sources, including data science textbooks, research papers, online forums, and websites. The collected data is then processed by tokenizing text, removing stop words and punctuation, converting text to lowercase, and removing special characters and numbers.

Section	**Description**
Data Overview	Collect text data from various sources, including: - Data science textbooks and research papers - Online forums and discussion boards (e.g., Kaggle, Reddit) - Data science-related websites and blogs
Data Preparation	The preparation phase of the chatbot based on the data science enquiry project using NLP involves several key steps.
Handling missing values	To address this issue, various strategies can be employed, including data imputation using statistical methods such as mean, median, or mode imputation, or using machine learning algorithms such as regression or decision trees.

Handling Outliers in Data	Outliers can be handled using various techniques such as data trimming, data transformation, and robust regression methods. Data trimming involves removing a percentage of the data from the extremes, while data transformation involves transforming the data to reduce the effect of outliers. Robust regression methods, such as the least absolute deviation method, can also be used to reduce the impact of outliers.
Install Rasa and Dependencies and Data Preparation	
Installation of Rasa	Once the virtual environment is active, you can install Rasa and its dependencies using pip, the Python package manager. Run the following command in the terminal: >>> pip install rasa
Setting up Rasa Project	>>> rasa init
Collecting the data: nlu.yml file	The nlu.yml file in Rasa is used to define the training data for the NLU (Natural Language Understanding) : 

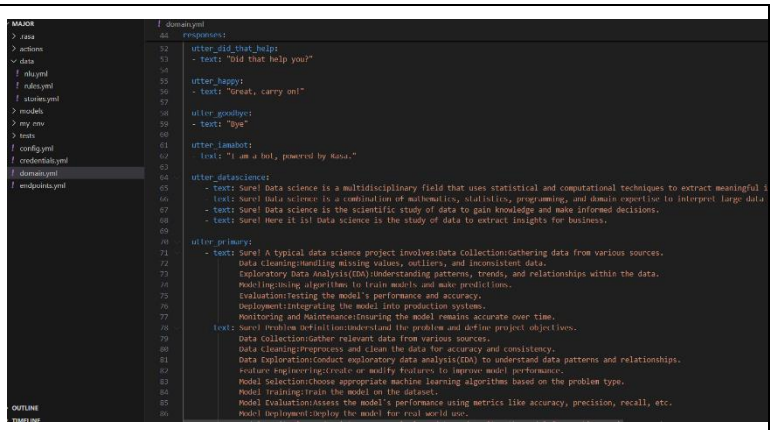
```
DATA
nlu.yml
rules.yml
stories.yml
OUTLINE
```

```yml
version: "3.1"

nlu:
- intent: greet
  examples: |
    - hey
    - hello
    - hi
    - hello there
    - good morning
    - good evening
    - moin
    - hey there
    - let's go
    - hey dude
    - goodmorning
    - goodevening
    - good afternoon

- intent: goodbye
  examples: |
    - cu
    - good by
    - cee you later
    - good night
    - bye
    - goodbye
    - have a nice day
    - see you around
    - bye bye
    - see you later

- intent: affirm
  examples: |
    - yes
    - y
```

domain.yml	
Stories.yml