

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

Preparation Template

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: <ul style="list-style-type: none"> - 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:

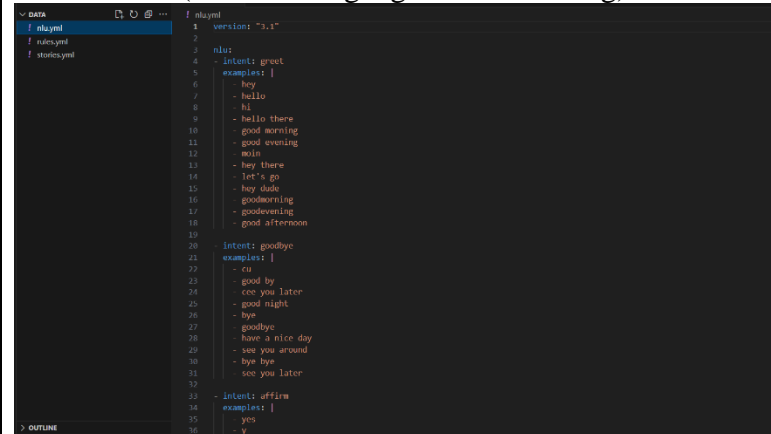
```
>>> pip install rasa
```

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) :



```
1 version: "3.1"
2
3 nlu:
4   - intent: greet
5     examples: |
6       - hey
7       - hello
8       - hi
9       - hello there
10      - good morning
11      - good evening
12      - moin
13      - hey there
14      - let's go
15      - hey dude
16      - goodmorning
17      - goodevening
18      - good afternoon
19
20   - intent: goodbye
21     examples: |
22       - cu
23       - good by
24       - see you later
25       - good night
26       - bye
27       - goodbye
28       - have a nice day
29       - see you around
30       - bye bye
31       - see you later
32
33   - intent: affirm
34     examples: |
35       - yes
36       - y
```

domain.yml

```
MAJOR | domain.yml
> data | 40 responses:
> actions | 52 utter_id_that_helps
> data | 53 - text: "Did that help you?"
> data | 54
> data | 55 utter_happy:
> data | 56 - text: "Great, carry on!"
> data | 57
> models | 58 utter_greeting:
> my_new | 59 - text: "bye"
> tests | 60
> config.yml | 61 utter_lamabot:
> credentials.yml | 62 - text: "I am a bot, powered by Rasa."
> domain.yml | 63
> endpoints.yml | 64 utter_data_science:
65 - text: Sure! Data science is a multidisciplinary field that uses statistical and computational techniques to extract meaningful insights from data.
66 - text: Sure! Data science is a combination of mathematics, statistics, programming, and domain expertise to interpret large data sets.
67 - text: Sure! Data science is the scientific study of data to gain knowledge and make informed decisions.
68 - text: Sure! More it is! Data science is the study of data to extract insights for business.
69
70 utter_primary:
71 - text: Sure! A typical data science project involves: data collection (gathering data from various sources), data cleaning (handling missing values, outliers, and inconsistent data), exploratory data analysis (EDA) (understanding patterns, trends, and relationships within the data), model training (using algorithms to train models and make predictions), model evaluation (testing the model's performance and accuracy), model deployment (integrating the model into production systems), and monitoring and maintenance (ensuring the model remains accurate over time).
72
73 text: Sure! Let's break it down into steps:
74 1. Problem Definition: Understand the problem and define project objectives.
75 2. Data Collection: Gather relevant data from various sources.
76 3. Data Cleaning: Preprocess and clean the data for accuracy and consistency.
77 4. Data Exploration: Conduct exploratory data analysis (EDA) to understand data patterns and relationships.
78 5. Feature Engineering: Create or modify features to improve model performance.
79 6. Model Selection: Choose appropriate machine learning algorithms based on the problem type.
80 7. Model Training: Train the model on the dataset.
81 8. Model Evaluation: Assess the model's performance using metrics like accuracy, precision, recall, etc.
82 9. Model Deployment: Deploy the model for real-world use.
83
84 OUTLINE | 85
86 TIMELINE | 86
```

Stories.yml

```
MAJOR | data | stories.yml
> data | 12 - story: bad path 2
> actions | 13
> data | 14 - story: enquiry path
> data | 15 steps:
> data | 16 - intent: data_science
> data | 17 - action: utter_data_science
> data | 18 - intent: primary
> data | 19 - action: utter_primary
> data | 20 - intent: relation
> data | 21 - action: utter_relation
> data | 22 - intent: applications
> data | 23 - action: utter_applications
> data | 24 - intent: tools
> data | 25 - action: utter_tools
> data | 26 - intent: libraries
> data | 27 - action: utter_libraries
> data | 28 - intent: skills
> data | 29 - action: utter_skills
> data | 30 - intent: programming
> data | 31 - action: utter_programming
> data | 32 - intent: type
> data | 33 - action: utter_type
> data | 34 - intent: rules
> data | 35 - action: utter_rules
> data | 36 - intent: analytics
> data | 37 - action: utter_analytics
> data | 38 - intent: important
> data | 39 - action: utter_important
> data | 40 - intent: scientist
> data | 41 - action: utter_scientist
> data | 42 - intent: analyst
> data | 43 - action: utter_analyst
> data | 44 - intent: advantages
> data | 45 - action: utter_advantages
> data | 46 - intent: disadvantages
> data | 47 - action: utter_disadvantages
> OUTLINE | 48
> TIMELINE | 49
```