Building a Chatbot Dataset for UN General Debate Statements

Introduction:

Your task is to create an end-to-end solution for building a chatbot dataset based on the text of each country's statement from the general debate at the United Nations (UN). The dataset includes statements separated by country, session, and year, and has been tagged accordingly. The original text was extracted from PDFs of UN general session transcripts using OCR (Optical Character Recognition). Your objective is to pre-process the data, design a suitable architecture for the chatbot dataset, and organize the data for downstream natural language processing (NLP) tasks.

Dataset Description:

The description and details is documented over at - https://www.kaggle.com/datasets/unitednations/un-general-debates

Task Overview:

Your task is to design end-to-end architecture with efficient processing of the dataset and create a chatbot dataset that can be used for training and evaluating NLP models. The chatbot dataset should facilitate the development of a chatbot capable of answering questions or providing information based on the UN general debate statements.

Project Requirements (to be considered):

- Data Pre-processing: Pre-process the dataset to remove any artifacts from OCR scans, handle missing values (if any), and perform any necessary text cleaning and normalization.

- Chatbot Dataset Design: Design the chatbot dataset in a format suitable for training and evaluating an NLP-based chatbot model. Decide on the structure of the dataset and how to organize it for different NLP tasks.

- Data Splitting: Split the dataset into training, validation, and testing sets to allow for model training and evaluation. Consider the distribution of statements across sessions and years to ensure representative splits.

- Architecture and Framework Selection: Choose an appropriate NLP framework and architecture for building the chatbot. Justify your choice based on factors such as model performance, ease of implementation, and available resources.

- Model Training and Evaluation: Train the chatbot model on the prepared dataset and evaluate its performance using relevant NLP metrics. Discuss the model's strengths and potential limitations.

- Chatbot Deployment Considerations: Discuss the challenges and considerations involved in deploying the trained chatbot in a real-world scenario, including integration with existing systems and user interaction design.

- Future Enhancements: Propose potential future enhancements or features that could be added to improve the chatbot's functionality and performance.

Deliverables:

- A detailed report explaining your approach, design decisions, and the results of

the chatbot model evaluation.

- A presentation summarizing your methodology, key findings, cloud infrastructure and cost comparison (AWS vs. Azure) and suggestions for future enhancements

Note to the Candidate:

- You can use popular NLP frameworks such as TensorFlow, PyTorch, or Hugging Face's Transformers.
- Ensure you handle the dataset with care, as it may contain sensitive information related to statements made by different countries in the UN general debate. Anonymize and respect privacy throughout your work.
- The focus of the case study is not solely on achieving state-of-the-art performance but also on your ability to design a robust, scalable, and efficient end-to-end solution for building a chatbot dataset from the given UN general debate statements.