DAY – 93

15 December 2025

Preparation of PDF and Dataset for Chatbot System

1. PDF Preparation

The PDF document used in this chatbot system acts as an **authoritative knowledge source**. In this project, the official PRSC document is used to ensure that the chatbot provides **accurate and domain-specific answers**.

Steps involved in PDF preparation are as follows:

1. **Selection of PDF Document**
   A reliable and official PDF document related to Punjab Remote Sensing Centre (PRSC) is selected. The document contains information such as services, workflows, procedures, and user guidelines.
2. **PDF Cleaning and Filtering**
   The raw PDF often contains unnecessary content such as:
   - Page numbers
   - Headers and footers
   - Copyright text
   - Figure numbers and captions
   - Repeated titles

   These elements are removed programmatically to avoid irrelevant text during chatbot response generation.

3. **Text Extraction from PDF**
   The PDF is converted into machine-readable text using the **PyPDF2 library**. Each page is processed line-by-line to extract meaningful content.
4. **Noise Removal and Normalization**
   Special characters, invisible Unicode symbols, extra spaces, and incomplete figure references are removed. Only clean, readable sentences are retained.

5. **Automatic Question–Answer Identification**
   A rule-based question detection technique is applied to identify possible questions using:
   - Question words (what, how, why, explain, steps, etc.)
   - Question marks
   - Numbered question formats

   The text following each detected question is treated as its corresponding answer.

6. **Integration with Knowledge Base**
   The extracted question–answer pairs are added to the chatbot's internal knowledge base and merged with existing dataset information. This allows the chatbot to answer both predefined and document-based queries.

2. Dataset (dataset.txt) Preparation

The dataset file acts as a **primary structured knowledge source** for the chatbot. It contains manually verified questions and answers for faster and more accurate responses.

Steps involved in dataset preparation are as follows:

1. **Dataset File Structure**
   The dataset is stored in a plain text file named dataset.txt. The file is organized language-wise to support bilingual interaction.

   Format:

   [english]
   What is PRSC? = Punjab Remote Sensing Centre is an autonomous organization...
   What services does PRSC provide? = PRSC provides GIS, GPS, and remote sensing services...

[punjabi]

PRSC ਕੀ ਹੈ? = ਪੰਜਾਬ ਰਿਮੋਟ ਸੈਂਸਿੰਗ ਸੈਂਟਰ ਇੱਕ ਸੁਤੰਤਰ ਸੰਸਥਾ ਹੈ...

2. **Manual Question Selection**

   Important and frequently asked questions are manually selected from:

   - Official PRSC documents
   - User manuals
   - Common user queries

   This ensures relevance and correctness.

3. **Answer Writing and Validation**

   Answers are written in **simple, clear, and factual language**. Each answer is reviewed to avoid ambiguity and ensure consistency.

4. **Language Separation**

   English and Punjabi questions are stored under separate tags to allow proper language detection and response generation.

5. **Data Normalization**

   Questions are stored in lowercase and without unnecessary symbols. This improves matching accuracy when user queries are compared using machine learning models.

6. **Dataset Expansion**

   The dataset can be easily extended by adding new question–answer pairs without modifying the system code, making the chatbot scalable.

Conclusion

Both **PDF and dataset preparation** play a crucial role in the chatbot's performance. The dataset provides **fast and structured responses**, while the PDF enables **dynamic knowledge extraction and advanced query handling**. Together, they ensure that the chatbot delivers accurate, reliable, and bilingual responses.