DAY – 85, 86, 87

3, 4, 5 December 2025

PDF Processing in My Project

Why PDF Processing Is Needed

In my project, most of the important information is present in the **PISMS User Manual (PDF)**.
This PDF is very long, so users cannot easily search or read it.

Therefore, the chatbot is designed to:

- **Read the PDF automatically**
- **Extract useful information**
- **Answer user questions directly from the PDF**

Overall Idea of PDF Processing

The PDF processing works in **four main stages**:

1. PDF loading
2. Text extraction
3. Text cleaning and filtering
4. Question–answer mapping

1. PDF Loading

First, the system loads the PDF file from the backend.

- The PDF file is stored in the project folder.
- Python library **PyPDF2** is used to open the PDF.

- Each page of the PDF is accessed one by one.

This allows the chatbot to read the entire document programmatically.

2. Text Extraction from PDF

After loading the PDF:

- Text is extracted page by page using PyPDF2.
- Each page's content is converted into **plain text**.
- The extracted text is split into **lines**.

At this stage:

- The text still contains unwanted content like:
  - Page numbers
  - Headers and footers
  - Figure references
  - Extra symbols

So cleaning is required.

3. Text Cleaning and Filtering (Important Part)

This is the **most important step**.

The system removes:

- Page headers like *"Punjab Remote Sensing Centre"*
- Page numbers
- Copyright lines
- Figure references such as *Fig. 1, Fig. 2*
- Extra spaces and useless symbols

Only **meaningful lines** are kept.

This ensures that the chatbot learns only **useful information**, not garbage text.

4. Question Detection in PDF

After cleaning the text, the system tries to **identify questions**.

A line is treated as a question if:

- It ends with a question mark (?)
- It starts with words like:
    - What
    - How
    - Steps
    - Procedure
    - Explain
    - Define

Example:

What is Chakbandi?
Steps to create a new request

These lines are marked as **questions**.

5. Answer Extraction

Once a question is detected:

- The lines **below the question** are treated as its answer
- The answer continues until:
    - Another question is found
    - Or a new section starts

So the structure becomes:

Question → Answer paragraph

These question–answer pairs are stored internally.

6. Storing Extracted Data

All extracted **question–answer pairs** from the PDF are stored in memory as:

- A dictionary or list
- Format:
  Question → Answer

This data is later combined with:

- The FAQ dataset file

7. Matching User Question with PDF Content

When a user asks a question:

1. User input is cleaned using NLP
2. The chatbot compares the user query with:
   - Questions extracted from the PDF
3. Similarity matching is performed using:
   - TF-IDF
   - Cosine similarity
4. The most relevant PDF answer is selected

8. Showing Answer to User

Finally:

- The selected answer is formatted
- Displayed in the chatbot window
- Shown in real time

How This Helps the User

- User does NOT read the full PDF
- User gets **direct answers**
- Saves time
- Reduces confusion
- Improves usability

Explanation of dataset.txt

What is dataset.txt in My Project?

dataset.txt is a **predefined knowledge base file** used by the chatbot.
It contains **frequently asked questions (FAQs) and their answers** related to irrigation services and land records.

This file helps the chatbot give **instant and accurate responses** without searching the PDF every time.

Structure of dataset.txt

The dataset is written in a **question–answer format** and organized by language.

Example structure from the file:

[english]
what is your name = I am your assistant chatbot.
hello = Hello! How can I help you today?
What is Jamabandi = Jamabandi is an official Record of Rights document...
What is Chakbandi = Chakbandi refers to land consolidation...

- The **left side** is the user's question
- The **right side** is the chatbot's answer
- = is used as a separator
- [english] indicates language section

dataset

Why dataset.txt Is Used

The dataset is used for:

- Common and repetitive questions
- Fixed definitions (Jamabandi, Chakbandi, A-Form, Naksha Nakal, etc.)
- Greeting messages (hello, bye, thank you)

This makes the chatbot **fast and efficient**.

How dataset.txt is loaded in the Project

1. The file is loaded when the chatbot server starts
2. Each line is read one by one
3. The question and answer are separated using =
4. Data is stored in a dictionary format:
5. Question → Answer
6. The chatbot keeps this data in memory for quick access

How the Chatbot Uses dataset.txt

When a user asks a question:

1. The user input is cleaned using NLP
2. The chatbot compares the input with questions in dataset.txt
3. If a match is found:
    - The answer is returned immediately
4. If no match is found:
    - The chatbot searches the **PDF-extracted data**
    - Or asks the user to rephrase

Priority Logic (Important for Viva)

✔ **Dataset is checked first**

✔ PDF is checked second

Reason:

Dataset answers are:

- More accurate
- Short
- Pre-verified

Advantages of using dataset.txt

- Faster response time
- No heavy processing
- Easy to update
- No database required
- Works even with low resources

Limitations of dataset.txt

- Cannot answer questions outside stored data
- Manual updating is required
- Limited to predefined knowledge