MILESTONE 3 : OCR Text Extractor with Chunking & Chatbot

1. Introduction

The OCR Text Extractor has now evolved into a more powerful tool with the integration of text chunking and a built-in chatbot. Alongside handling multiple file formats (CSV, Excel, PDF, image, text), the system can now split extracted content into smaller chunks for efficient retrieval and allow users to ask questions about their files directly within the app. This makes the tool not only a file analyzer but also an interactive assistant.

2. Key Features

- Upload and preview files (CSV, Excel, PDF, images, TXT)
- Quick info for CSV and Excel (rows, columns, numeric/text columns)
- OCR-based PDF text extraction using PyTesseract and pdfplumber
- Extracted PDF text is copyable and downloadable as JSON
- Image preview with size and mode metadata
- Text file preview with syntax highlighting
- Chat/Q&A; feature with retrieval from relevant text chunks
- Local summary fallback if Ollama AI is unavailable
- Advanced filtering and search for structured datasets
- Download filtered or chunked data as CSV, TXT, or JSON

3. Approach

The design philosophy of Milestone 3 extends the modular, user-friendly framework of Milestone 2 by adding: 1. Chunking: Large text content is broken into overlapping segments for retrieval.
2. Retrieval-Augmented Chat: Queries are answered using the most relevant chunks.
3. Resilience: If Ollama AI is unreachable, a brief local summary of the file or metadata is provided.
This ensures consistent functionality across structured and unstructured data sources.

4. Methodology

Step-by-step process:

Setup:
- Libraries: streamlit, pandas, pillow, pdfplumber, pytesseract, openpyxl
- Tesseract OCR engine must be installed and configured
- Added Streamlit page configuration and custom CSS styling

File Upload:
- Accepts CSV, Excel, PDF, TXT, PNG, JPG, JPEG
- Auto-detects MIME type for correct pipeline

File Preview:
- CSV/Excel: Preview first 100 rows + schema metadata

- PDF: Extracts embedded text with pdfplumber, applies PyTesseract OCR for images, content is copyable and downloadable
- Images: Display with metadata (resolution, mode)
- Text: Displayed in code-styled block

Chat/Q&A;:
- Retrieval: Scores and selects top-N relevant text chunks
- Builds prompt with context + query
- If Ollama AI is running: answers using llama3.1:latest model
- If Ollama AI is unavailable: provides local summary of text or metadata

Filter & Search:
- Numeric filters with sliders
- Category filters for text
- Full-text search
- Results downloadable as CSV

5. Outcome

With Milestone 3, the application now supports not just OCR and data preview but also interactive exploration and summarization. Users can upload heterogeneous files, extract and chunk text, and query the content interactively. The local fallback ensures reliability even without AI connectivity. This makes the tool a hybrid of OCR, data analysis, and conversational document interaction.

6. Conclusion

Milestone 3 significantly enhances the OCR Text Extractor by combining OCR, chunking, and chatbot capabilities. It enables structured and unstructured data to be explored, filtered, summarized, and queried in one streamlined interface. Future improvements may include deeper NLP integration, visualization support, and cloud-based AI for large-scale processing.