

PDF Documentation Extractor + Chunking + Ollama Integration

1. Objective

The goal of this project is to:

- Upload a PDF document.
 - Extract text from all pages.
 - Clean the text and remove empty lines.
 - Split (chunk) the text into smaller parts for easier handling.
 - Convert each chunk into JSON and save it as separate files.
 - Integrate Ollama so the user can ask custom questions about each chunk and get intelligent responses.
-

2. Technologies Used

- Python – Programming language.
 - Streamlit – Web frontend for user interaction.
 - pdfplumber – Extract text from PDF pages.
 - json – Convert extracted chunks into JSON format.
 - subprocess – Connect Streamlit app with Ollama CLI.
 - Ollama – Local AI model used for answering queries.
-

3. Working Procedure


1. User uploads a PDF document (e.g., document.pdf).
2. Text from all pages is extracted using pdfplumber.
3. The extracted text is cleaned to remove empty lines.
4. The full text is split into smaller chunks of lines (user defines chunk size).
5. Each chunk is displayed in Streamlit for review.
6. Each chunk is saved as a separate JSON file (output_chunk_X.json).
7. User can enter custom prompts for each chunk.

8. The app sends the chunk + prompt to Ollama and displays the AI's response.


4. Screenshots

PDF Table Extractor + Chunking + Ollama

Upload a Dataset PDF

 Drag and drop file here
Limit 200MB per file • PDF

Browse files

 dataset.pdf 56.0KB ×

Enter chunk size (rows per chunk)

5 − +

Chunk 1

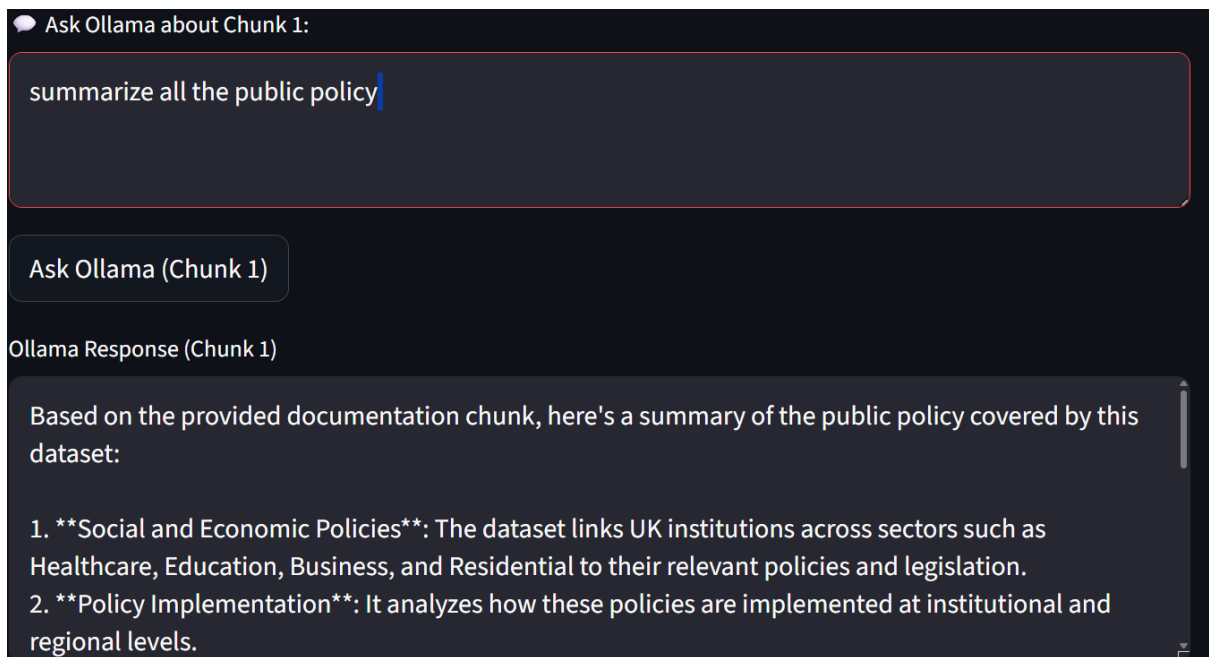
Chunk 1 Text

Documentation: UK Institutions & Policy-Linked Projects Dataset

1. Overview

Public policy provides the framework that guides the decisions and actions of UK government institutions, agencies, and officials. It defines the principles, strategies, and legal frameworks the government follows to address social, economic, and political issues. This dataset links UK institutions and projects across sectors such as Healthcare, Education, Business, and Residential to their relevant policies and legislation, with details like location,

```
{  
  "chunk_text": "Documentation: UK Institutions & Policy-Linked Projects Dataset"  
}
```



5. Applications

- Extracting structured datasets from PDFs.
 - Splitting large datasets for efficient analysis.
 - Converting PDF tables to machine-readable JSON.
 - Asking AI-powered queries (summaries, filtering, analysis) directly on dataset chunks.
-

6. Conclusion

This project integrates **PDF extraction, chunking, and AI-based analysis** into a single Streamlit application.

It enables users to handle large datasets more effectively and leverage for custom insights, summaries, and Q&A.