PDF Document Summarization Tool - Documentation

1. Overview

This Streamlit-based tool enables users to upload PDF files and get a summarized version of the content using Hugging Face Transformers.

It extracts text from the PDF and generates concise summaries using pretrained models like DistilBART.

2. Technologies Used

- Streamlit: UI and interaction
- PyPDF2: PDF text extraction
- Hugging Face Transformers: Summarization model
- Torch: Deep learning backend

3. Workflow

1. User uploads a PDF via Streamlit.
2. Text is extracted from all pages using PyPDF2.
3. Text is chunked (if too long).
4. Each chunk is passed through the Hugging Face summarization pipeline.
5. Summarized results are displayed on the screen.

4. How to Run

1. Clone or download the project folder.
2. Create a virtual environment (optional).
3. Install dependencies using: pip install -r requirements.txt
4. Run the app: streamlit run app.py

5. Model Used

Model: sshleifer/distilbart-cnn-12-6

It's a distilled version of BART optimized for speed and summarization tasks.

You can switch to facebook/bart-large-cnn for better quality (but slower).

6. Sample File Structure

pdf_summarizer/

|- app.py

|- requirements.txt

|- README.md (optional)

7. Possible Improvements

- Add ability to export summary as .txt or .pdf
- Integrate local LLMs (like Ollama)
- Use LangChain and a vector DB for large PDF summarization (RAG)