

PDF Summarizer using Hugging Face Transformers

Project Overview

The **PDF Summarizer** is a web application that allows users to upload a PDF file and instantly get a summarized version of its content using **Hugging Face Transformers**.

- **Frontend + Backend (Single Streamlit App):**
Built entirely in Streamlit, this project extracts text from uploaded PDFs, summarizes the content using a pretrained AI model, and displays both the original text and the summary on the same interface.
-

Tech Stack

Component	Technology	Purpose
Web Framework	Streamlit	Provides the interactive web interface for uploading PDFs and displaying summaries.
PDF Processing	PyMuPDF (fitz)	Extracts text from PDF pages efficiently.
NLP / Summarization	Transformers (Hugging Face)	Uses pretrained AI model to summarize long text.
Model	DistilBART (sshleifer/distilbart-cnn-12-6)	Lightweight transformer model optimized for text summarization.

Machine Learning
Backend

PyTorch

Backend framework used by
Transformers to run the summarization
model.

Programming
Language

Python 3

Core language for the application.

Project Structure

```
pdf-summarizer/  
|  
├─ app.py           # Streamlit app (UI + logic)  
├─ requirements.txt # Python dependencies  
└─ README.md        # Project documentation
```

requirements.txt

```
streamlit  
pymupdf  
transformers  
torch
```

Setup Instructions

1. Clone the repository

```
git clone <repo-url>  
cd pdf-summarizer
```

2. Create and activate a virtual environment

```
python -m venv venv  
# Windows  
venv\Scripts\activate  
# Linux / Mac  
source venv/bin/activate
```

3. Install dependencies

```
pip install -r requirements.txt
```

4. Run the Streamlit app

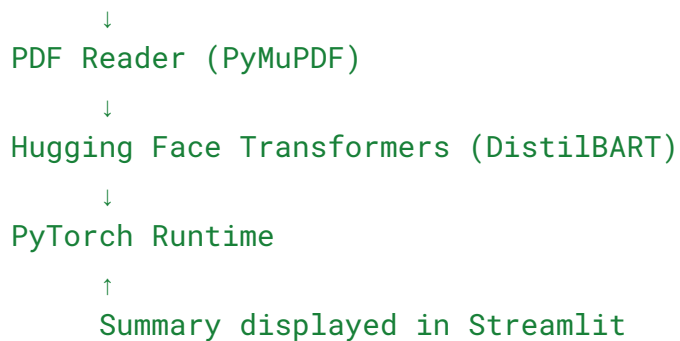
```
streamlit run app.py
```

- Opens a browser window with the **PDF Summarizer interface**.
- Upload a PDF and click “🔍 **Summarize**” to generate a summary.

Architecture


High-Level Overview:

User → Streamlit Web Interface



- **Streamlit** provides both the UI and the execution environment.
 - **PyMuPDF** extracts text from uploaded PDFs.
 - **Hugging Face Transformers** processes the text and generates a concise summary.
 - **PyTorch** runs the summarization model under the hood.
-

Flow

1. User uploads a PDF file through the Streamlit interface.
2. **PyMuPDF** extracts the text from all pages.
3. Extracted text is displayed in a **text area** for reference.
4. When the user clicks “ **Summarize**”,
 - The app loads the **DistilBART** summarization model.
 - Only the first 1024 characters are processed (due to token limits).
5. The **summary** is displayed under “ Summary”.

Summary:

The **PDF Summarizer** is an all-in-one Streamlit app that combines **AI summarization** and **PDF text extraction** to help users quickly understand lengthy PDF documents without manually reading them.