# PDF Document Structure Extractor using OCR + BERT

This project uses Doctr OCR and Transformers (DistilBERT) to extract structured information (like titles, headings) from PDF files and generate a logical outline of the document.

## Features

* • OCR using doctr with pretrained detection & recognition models
* • Converts PDF into word-level data with coordinates and confidence
* • Trains a classifier using BERT (distilbert-base-uncased) on labeled sentence data
* • Groups words into sentences and classifies each sentence into structural labels (title, H1, H2, H3, etc.)
* • Outputs an outline of the PDF in structured JSON format

## Project Structure

main.py # Contains all core functions   
1.csv # Labeled dataset for training   
file01.pdf # Input PDF to process   
bert-best-model/ # Trained model directory (created after training)   
README.docx # This file (for Word)

## Requirements

pip install python-doctr transformers pandas datasets evaluate torch

## How it Works

### 1. OCR and Word Extraction

pdf\_to\_wordlevel\_df(pdf\_path)

### 2. Sentence Grouping (Optional Preprocessing)

group\_words\_to\_sentences(df, labels\_to\_group)

### 3. BERT-based Classification

Trainer(...).train()

### 4. Inference and Structure Prediction

predict\_structure\_labels\_for\_pdf(pdf\_path, model, tokenizer)

### 5. Outline Generation

assemble\_outline\_json(predicted\_output)

## Training Dataset Format

Your training CSV (1.csv) should contain two columns:

text | label

-------------------------------------|------

Introduction to Robotics | H1

Robot Actuators and Power Systems | H2

Course Syllabus | title

## Example Usage

# Train the model  
trainer.train()  
  
# Predict structure labels on a new PDF  
predicted\_labels = predict\_structure\_labels\_for\_pdf("file01.pdf", model, tokenizer)  
  
# Print predictions  
for item in predicted\_labels:  
 print(f"Page {item['page']} | Label: {item['label']} | Text: {item['text'][:50]}...")  
  
# Generate a structured outline  
outline = assemble\_outline\_json(predicted\_labels)  
print(outline)

## Sample Output

{  
 "title": "Introduction to Robotics",  
 "outline": [  
 {  
 "level": "H1",  
 "text": "Robot Architecture Overview",  
 "page": 2  
 },  
 {  
 "level": "H2",  
 "text": "Sensor Types",  
 "page": 3  
 }  
 ]  
}

## TODO (Optional Enhancements)

* • Improve sentence grouping using line/block information from OCR
* • Add visualization of bounding boxes and labels
* • Support table and multi-column layout parsing

## References

• Doctr: https://github.com/mindee/doctr

• HuggingFace Transformers: https://huggingface.co/docs/transformers