**Table Extraction from PDF Files**

**Overview**

This script is designed to extract tables from PDF files and save them in JSON format. It uses deep learning models to detect tables and Optical Character Recognition (OCR) to extract text from table cells.

**Prerequisites**

* Python 3.x
* Required Python libraries: os, gc, json, argparse, pandas, torch, tqdm, pdf2image, transformers, pytesseract, PIL, numpy

**Installation**

Install the required libraries using pip:

pip install os gc json argparse pandas torch tqdm pdf2image transformers pytesseract pillow numpy

**Usage**

To run the script, use the following command:

python Complete\_Table\_Extraction.py <input\_dir> <output\_dir> <threshold> [--batch\_size BATCH\_SIZE] [--max\_threshold MAX\_THRESHOLD]

* <input\_dir>: Directory containing the input PDF files.
* <output\_dir>: Directory to save the extracted JSON files.
* <threshold>: Threshold value for table detection (between 0 and 1).
* --batch\_size BATCH\_SIZE (optional): Number of pages to process in each batch (default is 5).
* --max\_threshold MAX\_THRESHOLD (optional): Maximum threshold value for table detection (default is 0.96).

**Example**

🡪 python Complete\_Table\_Extraction.py "./input\_pdfs" "./output\_json" 0.5 --batch\_size 10 --max\_threshold 0.95

**Key Functions**

* **pdf\_to\_img(pdf\_path):** Converts PDF to images.
* **detect\_table(image\_doc, threshold):** Detects tables in images.
* **get\_table\_bbox(results, max\_threshold):** Extracts table bounding boxes.
* **highlight\_tables(image, table\_boxes, padding):** Highlights detected tables on images.
* **get\_cropped\_images(image, table\_bbox, padding):** Crops table images from main images.
* **get\_table\_features(cropped\_image):** Extracts features from cropped table images.
* **display\_detected\_features(cropped\_image, features):** Displays features on cropped table images.
* **get\_cell\_coordinates\_by\_row(table\_data):** Retrieves cell coordinates for table rows.
* **apply\_ocr(cell\_coordinates, cropped\_image):** Applies OCR to extract text from table cells.
* **process\_batch(batch\_images, batch\_num, threshold, max\_threshold):** Processes batches of images to detect and extract tables.
* **dataframe\_to\_json(table\_data):** Converts table data to JSON format.

**Progress Tracking**

* Tracks total PDFs processed.
* Shows current PDF being processed.
* Displays batch processing progress.

**Modifying Parameters**

* Adjust <threshold> for desired table detection accuracy.
* Change --batch\_size for batch processing optimization.
* Set --max\_threshold to filter table detections.