README.md

InvoicePlaceholder

Overview

This module is a part of the GridScanner app. It takes the output of the CUTIE model and parses the output based on the confidence level.

How single values are detected: Based on the maximum confidence level, the final classified field is selected.

How Address is detected: For all the classified words, the most confident word is selected.

Based on the bounding box, the word height is calculated. After that, two lines above that word and two lines below that word are selected for the green region. All words classified in that region are taken as the Address and written based on their text ids' ascending order.

How Table is detected: All the HSN class fields are detected, and then based on their positional values, the regions of separate rows are detected. For each row, the class fields of the Table are written in the Excel file.

How IDs are corrected: Regex match is made, and it is checked that an English word is not detected.

How the Sum is calculated: All the numeric values are converted to int or float. So that Sum can be directly calculated in the XLSX file itself.

How to run

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

- Start server using `uvicorn app:app --port=8001`
- The api is exposed at http://localhost:8001/getXLSX

> Run using docker

- Build doker image using `docker build -t invoiceplaceholder .`
- Run using `docker run -it --rm -p 8001:8001 invoiceplaceholder:latest`
- The api is exposed at http://localhost:8001/getXLSX

Challanges faced

- OCR detected Multiple words into one Bounding Box
- Mathematical Model applied for Numerical Values
- Problem in Multiline fields due to skewed images