yavar

Yavar Internship Selection – May 2025

Assessment Problem Statement - 1

Title: Invoice Data Extraction & Verification from Scanned PDFs

Design and implement a solution that reads scanned invoice PDFs (non-readable image-based documents), extracts structured information using **open-source OCR models**, and performs **verifiability checks** to ensure data integrity. The final output should be available in both JSON and Excel formats.

Background

Businesses frequently receive invoices as scanned PDFs that are not text searchable. Automating the extraction and validation of key information from these documents is crucial for ERP integration workflows.

Important: Expected Input & Output structure

- **Input:** Scanned invoice PDFs placed in current_directory/input/
- **Output:** Structured outputs in current_directory/output/ containing:

  (a) extracted_data.json: Parsed invoice data in JSON format

  (b) extracted_data.xlsx: Same data in Excel format

  (c) verifiability_report.json: Confidence scores and field-level checks

  (d) seal and signature of vendor (cropped image)

Fields to Extract

General Information

- invoice_number
- invoice_date
- supplier_gst_number
- bill_to_gst_number
- po_number
- shipping_address
- seal_and_sign_present (boolean – was a seal or signature detected and save the picture as well in output folder)

Table Contents (Multiple Rows)

- no_items (total number of line items)
- description
- hsn_sac
- quantity
- unit_price
- total_amount (per row)
- serial_number (if available)

Verifiability Requirements

For each extracted field and the overall invoice:

1. **Confidence Scores**
    - Return a confidence score (range: 0–1) for each extracted field.
2. **Line Item Validation**
    - Ensure that for each row: unit_price × quantity = total_amount.
3. **Total Calculation Checks**
    - Validate:
        - subtotal = sum(line_total)

},

        - final_total = subtotal - discount + GST

4. **Field Flags**
    - For each key field and formula check, include a verified: true/false flag.
5. **Optional**
    - Log or highlight fields with failed validation or low confidence for manual review.

Technical Constraints

- Solution should be robust across different invoice formats and image qualities.
- Image preprocessing (e.g., denoising, binarization) is encouraged.
- Code should be modular and well-documented.

Evaluation Criteria

- Accuracy of field extraction and structure mapping
- Clean documentation of approach adopted
- Reliability of confidence scores and verifiability logic
- Clarity and maintainability of code.
- Solution approach and generalisability
- Use of only open-source components

verifiability_report.json

```
{

 "field_verification": {

   "invoice_number": {

     "confidence": 0.94,

     "present": true

   },

   "invoice_date": {

     "confidence": 0.91,

     "present": true
```

```
    },
    "supplier_gst_number": {
      "confidence": 0.89,
      "present": true
    },
    "bill_to_gst_number": {
      "confidence": 0.88,
      "present": true
    },
    "po_number": {
      "confidence": 0.86,
      "present": true
    },
    "shipping_address": {
      "confidence": 0.90,
      "present": true
    },
    "seal_and_sign_present": {
      "confidence": 0.80,
      "present": true
    }
  },
  "line_items_verification": [
    {
      "row": 1,
      "description_confidence": 0.93,
      "hsn_sac_confidence": 0.85,
      "quantity_confidence": 0.96,
      "unit_price_confidence": 0.95,
      "total_amount_confidence": 0.94,
      "serial_number_confidence": 0.87,
      "line_total_check": {
        "calculated_value": 300,
        "extracted_value": 300,
        "check_passed": true
      }
    }
  ],
  "total_calculations_verification": {
    "subtotal_check": {
      "calculated_value": 300,
      "extracted_value": 300,
      "check_passed": true
    },
    "discount_check": {
      "calculated_value": 0,
      "extracted_value": 0,
      "check_passed": true
    },
    "gst_check": {
      "calculated_value": 54,
      "extracted_value": 54,
      "check_passed": true
    },
    "final_total_check": {
      "calculated_value": 354,
      "extracted_value": 354,
      "check_passed": true
    }
  },
  "summary": {
    "all_fields_confident": true,
    "all_line_items_verified": true,
    "totals_verified": true,
    "issues": []
  }
}
```

Timeline for completing assessment:

Time of expected completion: EoD: 30-May-2025

(a) A writeup explaining crisply the approach adopted, models used, fine tuning, preprocessing done etc

(b) Solution to be uploaded in Github and the link to be provided to the Yavar team with appropriate permissions.