Passbook OCR Processing Pipeline (SBI Example)

1. Upload Passbook .pdf ≤ 1MB

- `def convert_pdf_to_images(pdf_file)`
- `/upload → return image`

2. Store in Bucket

- `def pdf_to_s3(bucket_name, pdf_file)`
- `→ return image (binary)`

3. Preprocessing (`def preprocess_image(image)`)

- **Image Quality Check** (Verify resolution & dimensions)
- **Noise Removal** (Blur, faded text, unnecessary characters)
- **Grayscale Conversion** (Simplify processing)
- **Deskewing** (Correct any tilt)

→ *Store in bucket*

4. Invoke Textract API (async) size ≤ 50MB

- `def analyze_passbook(bucket_name, document_name)`
- `AnalyzeDocument API`

5. Data Structuring + Condition Check

Field	Condition
Bank Name	Must match `State Bank of India (SBI)`
Branch Name	String (e.g., "MG Road, Indore")

Branch Code	5-digit numeric code (e.g., `01234`)
Account Holder Name	String (e.g., "Rahul Sharma")
Account Number	11-17 digit numeric (e.g., `3022114455667788`)
IFSC Code	11-character alphanumeric (e.g., `SBIN0001234`)
CIF Number	11-digit numeric (e.g., `12345678901`)

Next page of passbook

Balance Amount	Numeric with decimal (e.g., `₹50,000.75`)
Transaction Date	DD/MM/YYYY format (e.g., `25/03/2025`)
Transaction Type	Credit/Debit
Transaction Description	String (e.g., "ATM Withdrawal")
Transaction Amount	Numeric with decimal (e.g., `₹2,000.00`)

6. Parsing (Extract Key-Value Pairs)

- `def extract_fields_passbook(response)`
- `/extract → return extracted_data`

7. Validation & Error Handling

- Check if all required fields are present
- `def missing_fields(data)`

    - { if all fields missing } → prompt: upload doc again

- def prompt_missing_fields(missing_fields)
    - → op user ip for missing field

8. Store Extracted Data (DynamoDB)

- def store_passbook_data(extracted_data)
- → Output table in DynamoDB
