NAME: BHAVYA TANEJA

MILESTONE-3 REPORT

BFSI- OCR OF BANK STATEMENTS

OBJECTIVE

Objective: Analyze the extracted financial data to provide insights into salary and expenses.

Overview

This Python application is a sophisticated tool for extracting and visualizing data from financial documents using Optical Character Recognition (OCR) and data visualization techniques.

Key Libraries and Technologies

1. OCR and Image Processing

- **EasyOCR**:
    - Primary library for text recognition in images
    - Supports multiple languages (used with English in this code)
    - Extracts text from preprocessed images with high accuracy
- **OpenCV (cv2)**:
    - Image preprocessing techniques
    - Denoising images
    - Adaptive thresholding
    - Improving image quality for better OCR results

2. Data Extraction and Matching

- **Fuzzy Wuzzy**:
    - Implements text similarity matching
    - Uses partial ratio to identify similar text variations
    - Helps in flexible field extraction with a configurable similarity threshold

3. Visualization and Charting

- **Matplotlib**:

- Creates bar charts and pie charts
- Customizes chart appearance
- Generates visual representations of extracted data

4. Web Interface

- **Gradio**:
  - Creates an interactive web interface
  - Allows file uploads
  - Provides buttons for extraction and visualization

5. Additional Libraries

- **Cohere**: Natural language processing capabilities
- **Collections** : Efficient data storage
- **Regular Expressions** : Text validation

Extraction Process

Image Preprocessing

Using python *Steps:*

1. *Convert to grayscale*
2. *Apply denoising*
3. *Enhance contrast*
4. *Use adaptive thresholding*

Text Extraction Strategy

1. Multiple field variations are defined for different document types
2. Uses fuzzy matching to identify relevant fields
3. Validates numeric values
4. Extracts values near matched fields

Visualization Techniques

Bar Chart

- Displays extracted values across multiple images

- Color-coded for different fields
- X-axis represents images
- Y-axis represents extracted numeric values

Pie Chart

- Shows proportional representation of extracted values
- Percentage-based visualization
- Helps understand relative magnitudes

Technical Highlights

Flexible Document Processing

- Supports multiple document types:
    - Balance Sheets
    - Pay Slips
    - Bank Statements

Error Handling

- Robust error handling in image processing
- Graceful handling of missing or invalid data
- Fallback mechanisms for incomplete extractions

User Interface

- Gradio interface
- Customizable document type selection
- Multiple image upload
- Chart type selection

Limitations

- Relies on consistent document layouts
- May struggle with highly complex or non-standard documents
- Requires clean, clear images for best results

Conclusion

This tool demonstrates a powerful combination of image processing, OCR, and data visualization techniques, providing an innovative solution for automated financial document analysis.