Springboard (Milestone 3&4) - Report

❖ Introduction

The project aims to develop a comprehensive tool for extracting and analyzing data from various types of documents. The primary focus is on automating the data extraction process using OCR technology and providing visual insights into the extracted data.

❖ Objectives

- To implement OCR functionality for text extraction from images.
- To analyze the extracted text and identify key fields based on document types.
- To visualize the extracted data using graphical representations.
- To create a user-friendly interface for interaction with the tool.

❖ Technology Stack

- PaddleOCR: Utilized for Optical Character Recognition to extract text from images.
- Cohere API: Used for natural language processing to interpret and analyze the extracted text.
- Gradio: A framework for building interactive user interfaces, allowing users to upload files and view results.
- Matplotlib: Employed for generating visualizations of the extracted data.
- Cloudinary: Integrated for image storage and retrieval.

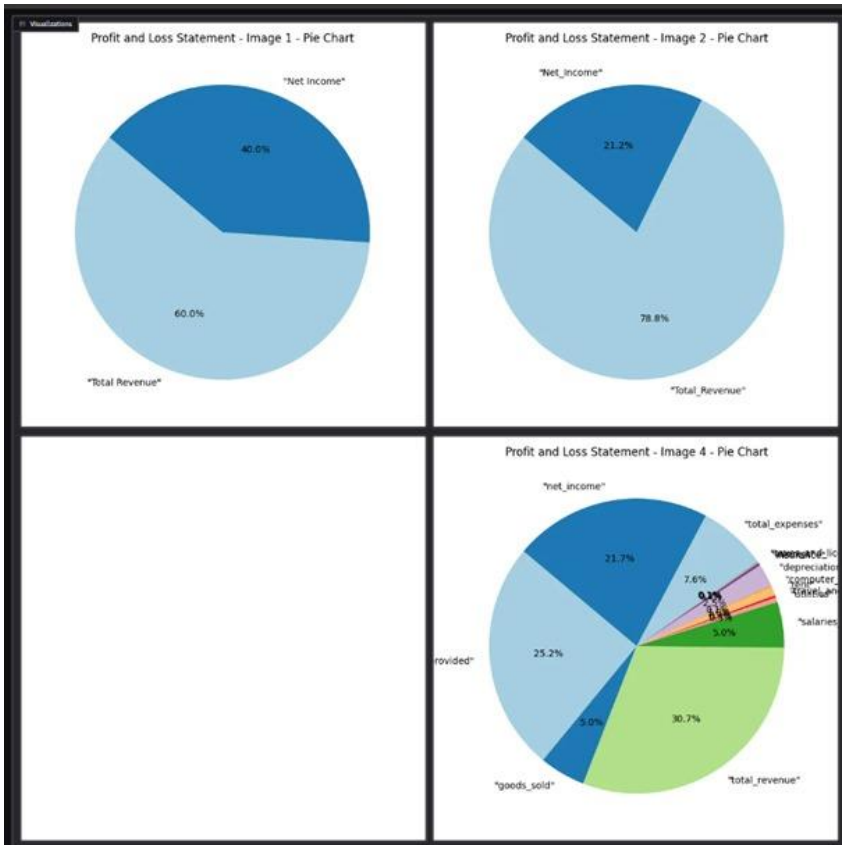
❖ Implementation Steps

1. Setup Environment: Installed necessary libraries and configured API keys.
2. Image Handling: Developed functions to download images from URLs and fetch images from Cloudinary.
3. OCR Processing: Implemented the process_documents function to handle image input, perform OCR, and extract relevant information based on document type.
4. Data Visualization: Created visual representations of extracted data using Matplotlib, allowing users to view insights in Bar Plot or Pie Chart formats.
5. User Interface: Designed an interactive UI using Gradio, enabling users to select document types, upload images, and view results seamlessly.

Saladi V S S Siva Hemanth Kumar , Hemanthsaladi2004@gmail.com

❖ Results

The tool successfully extracts and analyzes data from various document types. The visualizations provide clear insights into the extracted information, making it easier for users to interpret the data.



❖ Example Document Types

- Salary Slip: Extracts fields such as Net Salary, Gross Salary, and Basic Salary.
- Profit and Loss Statement: Identifies Total Revenue and Net Income.
- Checks: Retrieves Account Number, Amount, and Bank Name.

❖ Challenges Faced

- Ensuring accurate text extraction from images with varying quality and formats.
- Handling different document layouts and structures during the analysis phase.
- Integrating multiple APIs while managing authentication and rate limits.

❖ Future Work

- Enhance the OCR accuracy by training on a more extensive dataset.
- Expand the tool to support additional document types.
- Implement user authentication and data storage for a more robust application.

Saladi V S S Siva Hemanth Kumar , Hemanthsaladi2004@gmail.com

❖ Conclusion

The project successfully demonstrates the potential of combining OCR technology with data analysis and visualization to streamline the extraction of information from documents. The user-friendly interface and effective visualizations make it a valuable tool for users needing quick insights from their documents.

❖ Final Project:

Colab Notebook: Click Here to Access Colab Notebook