Institute of Technology of Cambodia

PROJECT TITLE

BANK TRANSCRIPT SCANNER

AUTOMATING FINANCIAL DATA EXTRACTION FROM BANK STATEMENTS

Agenda

Introduction

In today's fast-paced financial environment, businesses and auditors frequently deal with large volumes of bank statements in scanned PDFs or image formats. Extracting key financial data manually is time-consuming, error-prone, and inefficient.

Our project, Bank Transcript Scanner, aims to automate the process using OCR (Optical Character Recognition), machine learning, and natural language processing (NLP). This system will extract essential transaction details such as dates, amounts, transaction IDs, and account numbers from bank statements, classify different bank formats, and store the extracted data for easy access.

By integrating this solution into a web platform, users can upload statements, retrieve structured data, and streamline financial auditing processes with greater accuracy and efficiency.

Project Objectives

1. **Automate OCR Processing** – Convert scanned bank transcripts (PDFs/images) into readable text.
2. **Classify Bank Statements** – Identify bank sources to apply specific extraction rules.
3. **Extract Key Financial Data** – Retrieve transaction details such as date, amount, transaction ID.
4. **Validate & Store Data** – Ensure accuracy and store results in a structured format.
5. **Develop a Web Platform** – Provide a user-friendly interface for uploading, viewing, and downloading extracted data.

Scope & Feature

Core Features:

- OCR-based text extraction from bank statements.
- AI-driven bank classification for format adaptation.
- Data extraction using NLP & regex techniques.
- Validation & error handling for reliable output.
- Web-based interface for document uploads & results visualization.

Methodology

Technologies Used:

- **OCR Processing**: Tesseract OCR, OpenCV
- **Bank Classification**: Machine Learning (Scikit-learn, TensorFlow)
- **Data Extractio**n: Regex, NLP (spaCy)
- **Backend**: FastAPI (Python) for API handling
- **Frontend**: Next.js (React) for user interface
- **Database**: PostgreSQL or MySQL for structured storage

Version Control & Collaboration:

- **Git & GitHub:** Track code changes, manage branches, and ensure seamless team collaboration.
- **Project Management:** Use GitHub Projects or Trello to organize tasks, milestones, and development progress.

Timeline

Finalize project proposal

Week 5

Finalize project proposal

Week 6-7

OCR pipeline development

Week 8

Bank classification model

Week 9-10

Data extraction & validation

Week 11-12

Web platform development

Week 13

Testing & debugging

Week 14

Deployment & optimization

Week 15

Final report & presentation

More detail: Detailed Project Timeline and Process

Team Members

LENG Devid

LY Chungheng

NANG Chettra

NHEN Theary

NGOUN Lyhorng

NY Chantharith

THANK YOU

HERE'S OUR DETAILED PROJECT TIMELINE