ionidea AI
ENGAGE EXCELLENCE

Case Study | AI, NLP, RPA, DATA ENGINEERING

Complex PDF Document Data Extraction to a Large Accounting firm

Problem

One of the largest accounting firms faced a laborious process of manually extracting data from complex PDF documents for government reporting.

These documents contained multiple tables (horizontal, vertical, matrix) with critical import/export consignment information.

Manual extraction led to slow turnaround times and the potential for errors in regulatory submissions.

Solution

Developed two distinct tailored solutions:

1. Advanced Table Extraction from PDFs

- Intelligent Table Detection: Accurately identifying and delineating diverse table types within complex PDFs.
- Data Extraction OCR technologies combined with rule-based or machine learning techniques for extracting data from identified tables.
- Output: Transformation of extracted data into structured formats (e.g., Excel, JSON).

Architecture

- **OCR Engine:** The standalone and customizable OCR engine efficiently extracts text from digital PDFs, operating independently without requiring additional customization, providing JSON output containing layout information and associated text, and offering flexibility to accommodate various input file formats through parameter-driven functionality.
- **Mat2Info:** extracts text and information from complex tabular structured PDF files into Excel, employing advanced algorithms for precise data extraction, undergoing rigorous development to ensure high confidence in accurately reading client-provided documents, thereby ensuring reliability and accuracy.
- **Template Matching** efficiently recognizes fixed sections within documents by identifying and matching predefined templates, ensuring consistent extraction of relevant data across documents despite layout variations, and assigning values to specific keys/fields.

Complex PDF Document Data Extraction to a Large Accounting firm

Solution

2. Captcha Analytics for Security Enhancement

- **Captcha Simulation:** An engine to mimic real-world AI bot attempts to break Captchas.
- **Vulnerability Detection:** Identification of weaknesses in Captcha design.
- **Security Improvement:** Data-driven recommendations to strengthen Captcha security against automated attacks.

Architecture

- **OCR Engine:** The standalone and customizable OCR engine efficiently extracts text from digital PDFs, operating independently without requiring additional customization, providing JSON output containing layout information and associated text, and offering flexibility to accommodate various input file formats through parameter-driven functionality.
- **Business Logic:** interprets input data, applies predefined business rules, manages validation, calculations, and other operations, ensuring consistency, compliance, and efficiency in business processes.

Results

- Significantly streamlined data extraction from PDFs, reducing turnaround time and minimizing errors.
- Enhanced compliance with government reporting requirements.
- Proactive strengthening of Captcha security systems to protect sensitive financial data.
- Client has processed over 5500 documents in Table Extraction Service and over 25000 Captcha Images in Captcha Analysis Service

Technology Stack

- **PDF Handling:** PyPDF2
- **OCR:** In-house build OCR Engine
- **Data Transformation:** Pandas (Python)
- **Captcha Analytics:** Image processing library to improve image quality and In-house build OCR Engine

Software Development

- **Methodology:** Iterative development for refinement, especially for Captcha analytics.
- **Focus:** Accuracy and adaptability to handle variations in PDF table structures.
- **Security:** Prioritize secure handling of sensitive financial information.

Before Metrics

Time-consuming manual data extraction from PDFs.

Potential for errors impacting regulatory compliance.

Vulnerability of Captchas to automated attacks.

After Metrics

Significant reduction in data extraction time (e.g., from days to hours).

Improved accuracy and compliance with government reporting.

Measurable increase in Captcha robustness (e.g., decrease in successful bot attacks).