SUCCESS STORY

Nested Tables & Machine Drawing Text Extraction For An Oil & Gas Company

DOMAIN

Oil & Gas Industry

KEY HIGHLIGHTS

- 4x faster automated text extraction using teX.ai.
- The need for human intervention was reduced by over 80%.
- The quality of their process had increased by over 75%.

TECHNOLOGIES

The solution was built leveraging Python and several of its libraries.

OCR:
Tesseract, Tesserocr, OCRmyPDF, PyTesseract

Preprocessing and Post Processing Tools:
xPDF, Poppler, OpenCV, Pandas, Json

Table Detection and Extraction:
Camelot, OpenCV, LSD (line segment detection), csv, TensorFlow, FCN (Fully Convolutional Networks), CNN (Convolutional Neural Networks)

Application Deployment:
Flask, Docker

Nested Tables & Machine Drawingtext Extraction For An Oil & Gas Company

CUSTOMER BACKGROUND

The Client is one of the pioneers in the oil and gas business, with a focus on innovation to find ways to help their customers to fuel progress in agriculture, industry, medicine, science, space, technology, and transportation. The combination of engineering disciplines, computer science, geophysics, and metallurgy help create a winning formula for all stakeholders in such projects.

BUSINESS REQUIREMENTS

Given the document intensive nature of business, the client generally had to deal with numerous PDF documents dealing with complex drilling machine parts diagrams and data in nested tables and various other formats. Their requirement was to extract data and save in a format that could facilitate further analysis downstream.

CHALLENGES

- Client had hundreds of PDF documents and each of these PDF documents had pages ranging from 2 to 100 pages. In some cases, the required data was not present in all of the pages of the PDF documents.
- There were 5 different formats of documents consisting of engineering drawings, nested tables, un-demarcated tables, etc. This requires model creation for each of the document format.

OBJECTIVE

To leverage teX.ai for the automated text extraction process with an accuracy target of over 80% and requiring less than 50% of the current time taken.

SOLUTION OVERVIEW

- **Quality File Validation**
  - Extraction of chemical composition file and Converting it to a key-value pair.
  - These chemical composition type PDF are 10 pages long.
- **Survey Files**
  - Automatic identification of Survey(s) tables from multi-page documents followed by extraction.
- **Well Schematics**
  - Identify and extract the nested tables as separate entities. These documents had a combination of nested tables with complex drilling equipment's drawing.

APPROACH & IMPLEMENTATION

teX.ai was leveraged to process text for all the 3 use cases

Quality File Validation

- The Analysis table which contained the chemical composition details was identified in the document and extracted using OCR.
- The time taken to extract is just a few seconds and accuracy more than 85%.

Public Files (Surveys)

- First isolated the survey tables using the keyword search leveraging OCR.
- Survey details are then extracted using techniques such as Tabula or Camelot.

Well Schematics

- All the nested tables were extracted as separate tables and saved in CSV format.
- The nested tables are extracted in 2 stages leveraging FCN model at stage 1 and OpenCV in the next stage to detect rows in the table.

Deployment

- Once the AI models were built and the required accuracy and performance tuning complete, Indium deployed teX.ai with an admin interface built using Flask and containerization using Dockers.

BUSINESS IMPACT

- 4x faster text extraction from the source docments, by leveraging teX.ai in the automated process flow.
- The need for human intervention was reduced by over 80%.
- The quality of their process had increased by over 75%.

About Indium

Indium is a Digital Engineering Services leader and Full Spectrum Integrator that helps customers embrace and navigate the Cloud-native world with Certainty. With deep expertise across Applications, Data & Analytics, AI, DevOps, Security and Digital Assurance we "Make technology work" and accelerate business value, while adding scale and velocity to customer's digital journey on AWS.

indium
Make Technology Work

USA	INDIA	UK	SINGAPORE
Cupertino | Princeton Toll-free: +1-888-207-5969	Chennai | Bengaluru | Mumbai Toll-free: 1800-123-1191	London Ph: +44 1420 300014	Singapore Ph: +65 6812 7888

www.indiumsoftware.com

For Sales Inquiries
sales@indiumsoftware.com

For General Inquiries
info@indiumsoftware.com