Terra
Edgesoft

Use Case:

Data Extraction: Complex Tabular Form

Introduction

Terra Edge Soft (TES) successfully extracted data from a Digital PDF document having multiple tables (Horizontally and Vertically Arranged) for one of the leading Accounting/ Auditing / Taxation Advisory firm having multinational presence. The table comprises submission data of Import and export of consignment to Government Authorities. TES was chosen over leading standard solutions available in the market. The solution was customized to suit the client's business domain as well as specific form and format.

Background - Problem Statement:

In the era of Information Technology, the Portable Document Format (PDF) has become the most commonly used data format to share data. Documents like Financial Statements of all kinds, Engineering Bill of Materials - Parts Lists, HR Statements , Cost Sheets of complex engineering projects etc. contain complicated maze of tables. One of our clients, a reputed Accounting and Consulting company, needed to extract tabular data from the Declaration to Government Authority on material import/export in order to make further analysis. These statements are in Pure Digital PDF Format. Data Extraction using standard tools was not successful in consistent extraction of Key-Value Pair data and it further requires lots of manual efforts for post processing.

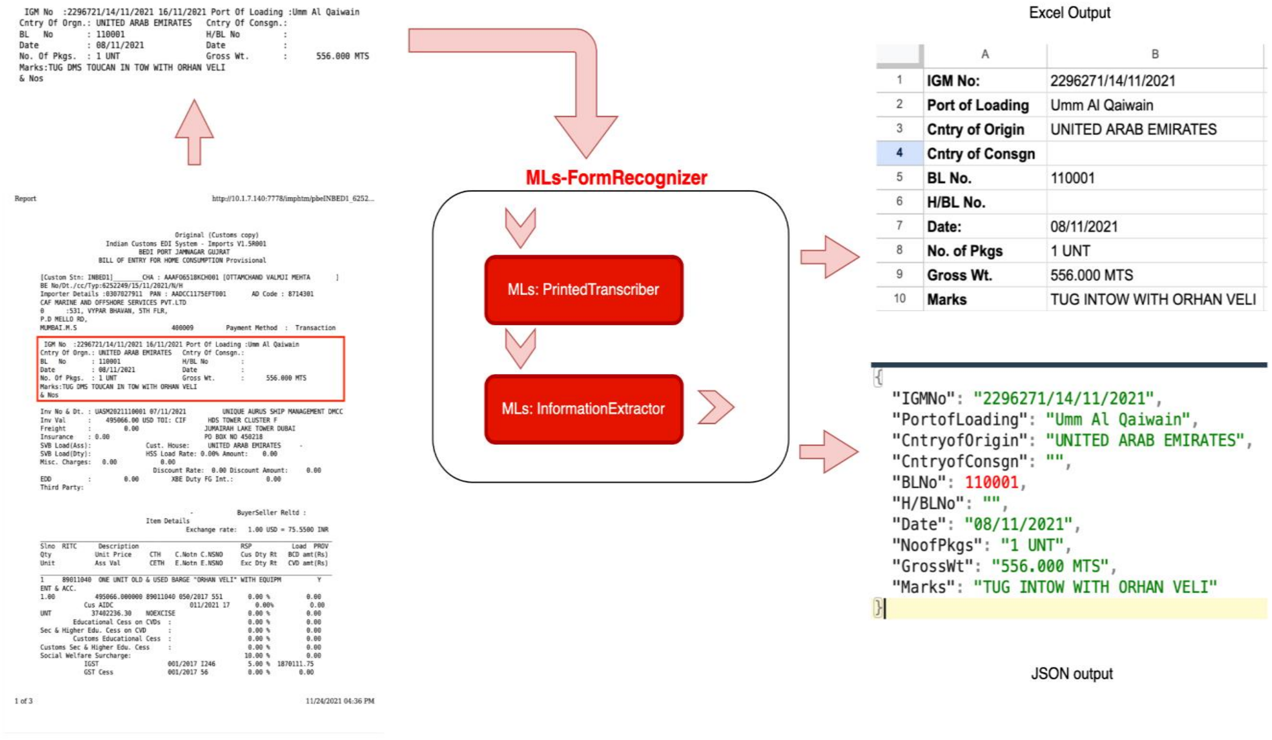
Furthermore, challenges associated with table extraction from Digital PDFs, is reduced accuracy even after using reputed PDF Table data extraction solutions.

- PDFs often contain complex layouts and formats, making them challenging to extract tabular data accurately.
- PDFs with multiple tables or nested tables can pose complications as the extraction process needs to distinguish correctly between different table structures.

Terra
Edgesoft

Approach:

To streamline data extraction and analysis workflow, Terra Edgesoft developed software: an automated application to extract tabular data assisting companies entailing less time, money and human resources making it flawless. Application extracts data from complicated tabular arrangement of Indian government documents. Our solution MLs: **FormRecognizer** consists of two primary components:



```
 IGM No  :2296721/14/11/2021 16/11/2021 Port Of Loading :Umm Al Qaiwain
Cntry Of Orgn.: UNITED ARAB EMIRATES   Cntry Of Consgn.:
BL   No        : 110001                H/BL No         :
Date           : 08/11/2021            Date            :
No. Of Pkgs.   : 1 UNT                 Gross Wt.       :      556.000 MTS
Marks:TUG DMS TOUCAN IN TOW WITH ORHAN VELI
& Nos
```

```
Report                                   http://10.1.7.140:7778/imphtm/pbeINBED1_6252...

                         Original (Customs copy)
                 Indian Customs EDI System - Imports V1.5R001
                        BEDI PORT JAMNAGAR GUJRAT
                BILL OF ENTRY FOR HOME CONSUMPTION Provisional

[Custom Stn: INBED1]________CHA : AAAFO6518KCH001 [OTTAMCHAND VALMJI MEHTA      ]
BE No/Dt./cc/Typ:6252249/15/11/2021/N/H
Importer Details :0307027911  PAN : AADCC1175EFT001      AD Code : 8714301
CAF MARINE AND OFFSHORE SERVICES PVT.LTD
0      :531, VYPAR BHAVAN, 5TH FLR,
P.D MELLO RD,
MUMBAI.M.S                        400009        Payment Method  :  Transaction
 IGM No  :2296721/14/11/2021 16/11/2021 Port Of Loading :Umm Al Qaiwain
Cntry Of Orgn.: UNITED ARAB EMIRATES   Cntry Of Consgn.:
BL   No        : 110001                H/BL No         :
Date           : 08/11/2021            Date            :
No. Of Pkgs.   : 1 UNT                 Gross Wt.       :      556.000 MTS
Marks:TUG DMS TOUCAN IN TOW WITH ORHAN VELI
& Nos
Inv No & Dt. : UASM2021110001 07/11/2021        UNIQUE AURUS SHIP MANAGEMENT DMCC
Inv Val      :    495066.00 USD TOI: CIF      HDS TOWER CLUSTER F
Freight      :          0.00                JUMAIRAH LAKE TOWER DUBAI
Insurance    : 0.00                         PO BOX NO 450218
SVB Load(Ass):               Cust. House:     UNITED ARAB EMIRATES     -
SVB Load(Dty):               HSS Load Rate: 0.00% Amount:    0.00
Misc. Charges:   0.00           0.00
                              Discount Rate:  0.00 Discount Amount:     0.00
EDD          :          0.00        XBE Duty FG Int.:         0.00
Third Party:

                                         -           BuyerSeller Reltd :
                             Item Details
                                      Exchange rate:   1.00 USD = 75.5500 INR

Slno  RITC       Description                                RSP            Load  PROV
Qty             Unit Price      CTH    C.Notn C.NSNO       Cus Dty Rt    BCD amt(Rs)
Unit            Ass Val         CETH   E.Notn E.NSNO       Exc Dty Rt    CVD amt(Rs)

1     89011040  ONE UNIT OLD & USED BARGE "ORHAN VELI" WITH EQUIPM            Y
ENT & ACC.
1.00            495066.000000 89011040 050/2017 551        0.00 %          0.00
             Cus AIDC                      011/2021 17       0.00%           0.00
UNT            37402236.30    NOEXCISE                     0.00 %          0.00
           Educational Cess on CVDs  :                      0.00 %          0.00
Sec & Higher Edu. Cess on CVD        :                      0.00 %          0.00
           Customs Educational Cess  :                      0.00 %          0.00
Customs Sec & Higher Edu. Cess       :                      0.00 %          0.00
Social Welfare Surcharge:                                  10.00 %          0.00
             IGST                  001/2017 I246            5.00 %  1870111.75
             GST Cess              001/2017 56              0.00 %        0.00

1 of 3                                                       11/24/2021 04:36 PM
```

Excel Output

	A	B
1	**IGM No:**	2296271/14/11/2021
2	**Port of Loading**	Umm Al Qaiwain
3	**Cntry of Origin**	UNITED ARAB EMIRATES
4	**Cntry of Consgn**	
5	**BL No.**	110001
6	**H/BL No.**	
7	**Date:**	08/11/2021
8	**No. of Pkgs**	1 UNT
9	**Gross Wt.**	556.000 MTS
10	**Marks**	TUG INTOW WITH ORHAN VELI

```
{
  "IGMNo": "2296271/14/11/2021",
  "PortofLoading": "Umm Al Qaiwain",
  "CntryofOrigin": "UNITED ARAB EMIRATES",
  "CntryofConsgn": "",
  "BLNo": 110001,
  "H/BLNo": "",
  "Date": "08/11/2021",
  "NoofPkgs": "1 UNT",
  "GrossWt": "556.000 MTS",
  "Marks": "TUG INTOW WITH ORHAN VELI"
}
```

JSON output

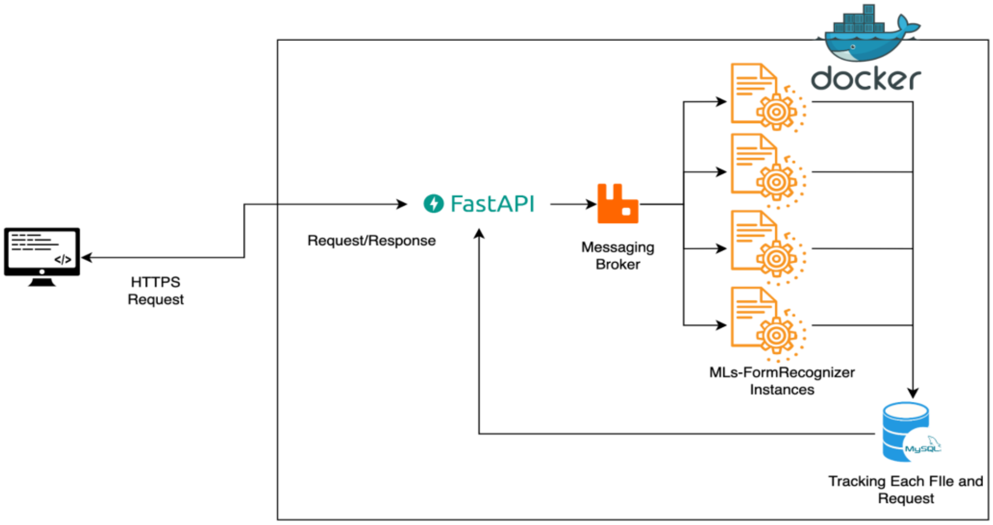
1. MLs: PrintedTranscriber:

A standalone and customizable OCR engine is used to extract text associated with the digital PDF, including metadata, and determine the positioning and arrangement of text and other elements on each page. It provides functionality to parse the internal structure of a PDF file and extract the desired content.

2. MLs: Information Extractor:

Information Extractor is designed to receive the output of the MLs: PrintedTranscriber and transform it into the desired JSON file or Excel file. This module is tailored to the specific document and table layout from which the necessary data needs to be extracted.

Solution Details:



As depicted in the diagram, the comprehensive solution is a containerized Application Docker-based API (REST) that integrates multiple modules to deliver cutting-edge technology. This solution encompasses a cohesive and robust architecture, leveraging state-of-the-art components for optimal performance and functionality. To address the user's need for secure access and PDF file uploads, a robustly crafted solution ensures data integrity and reliable processing. Given the potential for a high volume of requests, a queue-based approach is adopted to handle the workload efficiently. This approach guarantees no request is lost and enables to deliver the expected output to users within a short timeframe.

To support this system, a database is utilized to track and log individual requests. This allows to maintain a comprehensive record of all interactions and facilitates easy monitoring and troubleshooting. In the event of a system failure, a real-time mail alert is provided. This promptly notifies the relevant stakeholders, ensuring appropriate action to address any issues minimizing potential disruptions.

By leveraging this combination of queue-based processing, robust database tracking, and real-time alerts, System is designed to provide a secure and wonderful User Experience while accessing API and uploading PDF files. This comprehensive approach guarantees data integrity, minimizes downtime, ensures timely delivery of the expected results and reliability of application.

Deployment:

To provide choice of access to User following two options available.:

- Hosted on a Clients, server facility.
- Hosted by us - Accessible via API

Typical USE CASES: Table Extraction

1. Financial Analysis: Financial data from statements, Reports, and Filings, facilitating financial analysis, trend identification, and forecasting.
2. Market Research: Industry reports, Surveys, and Competitor analysis documents, providing insights into market size, growth rates, consumer behavior, and competitive landscapes for strategic decision-making.
3. Healthcare and Medical Research: Medical data from clinical trial reports, Research Papers, and Patient records, supporting epidemiological studies, medical research, and patient care decision-making.
4. Government and Public Administration: Reports, Surveys, and Public records, enabling data analysis, policy formulation, and evidence-based decision-making.
5. Manufacturing : Pdf Drawings containing Table like Bill of Material, Cross Reference List,
6. Document Digitization and Archiving: Table extraction services convert pdf documents into digital formats, creating searchable databases, automating data entry, and preserving historical records for efficient document management and retrieval.

Benefits:

- **Improved Accuracy:** Manual data extraction is prone to errors, but the application ensures a consistent and accurate extraction of tables, minimizing the error rate to less than 1%.
- **Increased Efficiency:** The table extraction application automates the process of extracting structured data from unstructured sources, resulting in a substantial time and effort savings for analysts.

- **Scalability:** The application can handle large volumes of documents, allowing analysts to process a higher number of tables and scale their data analysis operations.
- **Flexibility and Versatility:** The application supports various table formats and can be customized to cater to specific data extraction requirements.
- **Cost Efficiency:** Solution significantly reduces post-processing efforts to almost zero, resulting in substantial cost reduction while handling large and complex files. Additionally, it is highly cost-competitive, to comparative solutions in the market.

Overall, the table extraction application empowers data analysts to streamline their workflow, extract valuable insights, and make informed decisions by efficiently extracting structured data from unstructured documents.

More Information

www.terraedgesoft.com

contact@terraedgesoft.com