EXTERN + OUTAMATION
EXTERNSHIP

ADVANCED AI-POWERED

DOCUMENT INSIGHT AND DATA EXTRACTION PROJECT

SPRINT 4 STATUS UPDATE

by Lashawn Fofung, MBA, PMP, CSM, CSPO

L F

SKILLS

- Python
- AI/ML
- Data Scrapping
- Prompt Engineering
- LLMs
- Text Extraction
- Document Intelligence
- Python Automation
- Optical Character Recognition (OCR)
- Retrieval-Augumented Generation (RAG)

Helping businesses improve, automate, and architect critical tasks (such as document handling, customer service, and data management) by saving thousands of hours and boosting accuracy.

Mortgage Documents

SPRINTS

2025

01 AI Document Intelligence
02 Python & Google Colab
03 Python Data Extraction
04 Optimizing OCR
05 Implementing RAG
06 Optimizing RAG Pipelines
07 Blob Processing & Classification
08 Interactive Chatbot
09 AI-Powered Document Automation Platform

PROJECT TIMELINE

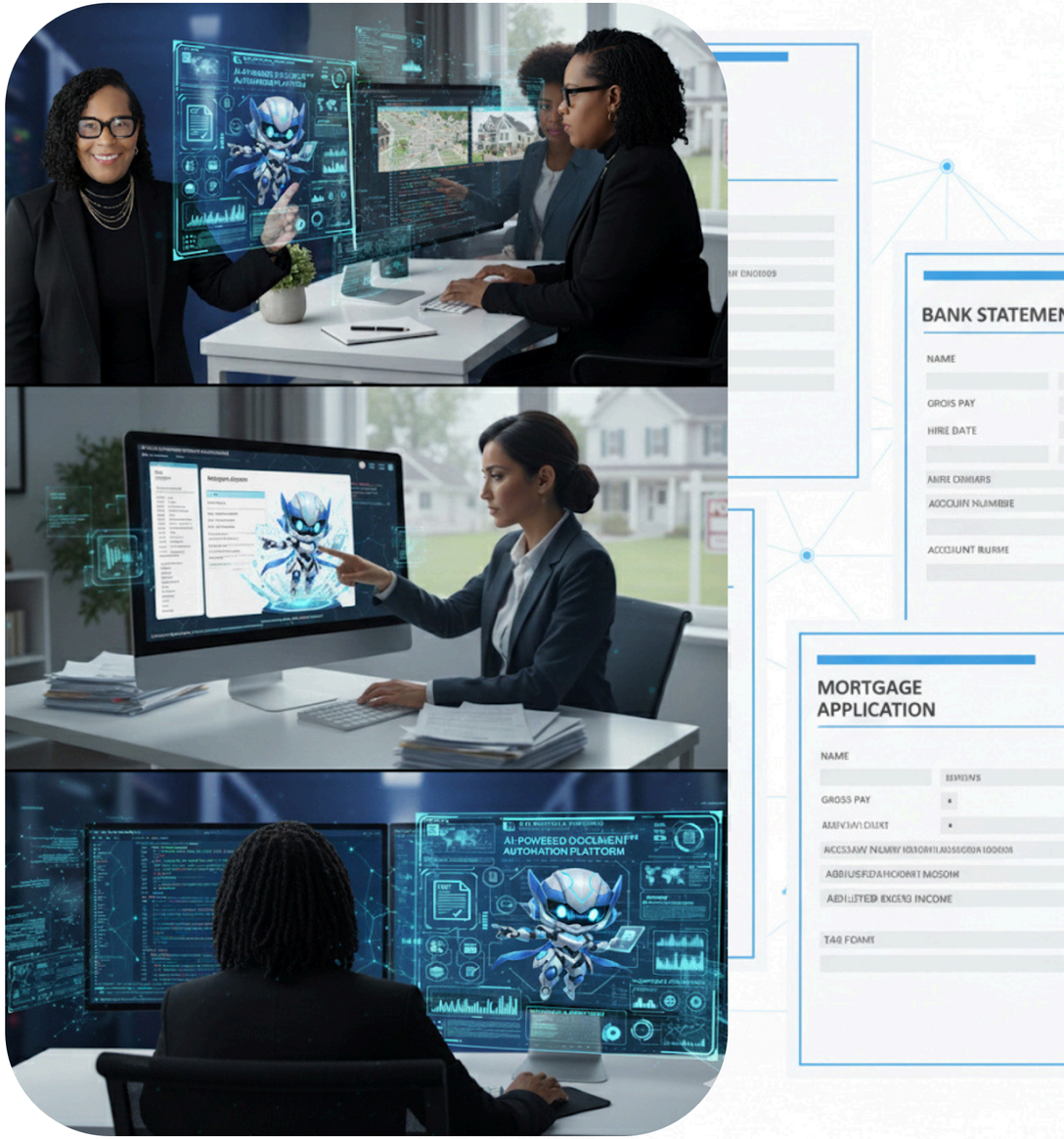
10 WEEKS

Optimizing OCR

Business Automation

Background: 10-Week Project

The Outamation Advanced AI-Powered Document Insights and Data Extraction Externship is a comprehensive 10-week program. Its core goal and purpose are to master the required skills to create an AI-Powered Document Automation Platform. The program focuses on bridging foundational technologies (ML, LLMs, NLP, Computer Vision) with high-value industry applications, such as solving complex, document-heavy challenges in mortgage automation, ultimately transforming slow, manual review processes into automated, efficient workflows.



Challenge & Solution: Building The Document Intelligence (AI) Pipeline

- **The Business Challenge:**
  - Automating the highly manual and time-consuming process of mortgage underwriting by handling 200-300 pages of unstructured documents.

- **The Solution:**
  - Designing and implementing a robust, AI-Powered Document Automation Platform to transform raw files into structured, intelligent insights.

- **The 4-Step Document Intelligence Workflow:**
  - Step 1. Document Acquisition & Pre-Processing (The Input)
  - Step 2. Data Extraction & Preparation
  - Step 3. Knowledge Base & Modeling (The AI Engine)
  - Step 4. Output, Review, & Integration (The Results)

This sprint mastered the critical front-end of the AI pipeline: acquiring unstructured documents, pre-processing images for clarity, extracting text via OCR, and preparing clean, structured data for analysis.

SPRINT 4 FOCUS

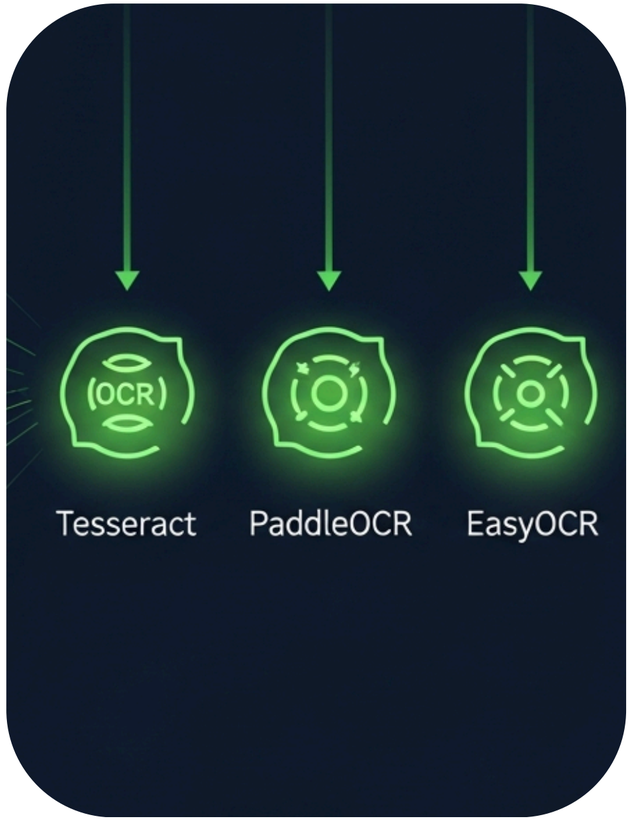
Steps 1 & 2: End-to-End OCR

(Acquisition, Pre-processing, Extraction, & Preparation)

Overall Status: Sprint 4 Complete

I've successfully concluded Sprint 4, focused on mastering Advanced OCR Comparison and Layout-Aware Extraction for the AI-Powered Document Automation Platform.

- **Comparative Engine Analysis:** Completed rigorous, head-to-head benchmarking of Tesseract, PaddleOCR, and EasyOCR to select the optimal, high-accuracy engine.
- **Pipeline Phase Complete:** Finalized the Document Acquisition & Pre-Processing phase, establishing a high-fidelity input source for downstream AI/LLM modeling.
- **Structured Data Output:** Advanced ability to transform complex, multi-page scans into clean, layout-aware JSON outputs, directly solving the input quality challenge.
- **Reliability Foundation:** Successfully established a robust OCR foundation critical for ensuring the overall accuracy and reliability of the AI-Powered Document Automation Platform.



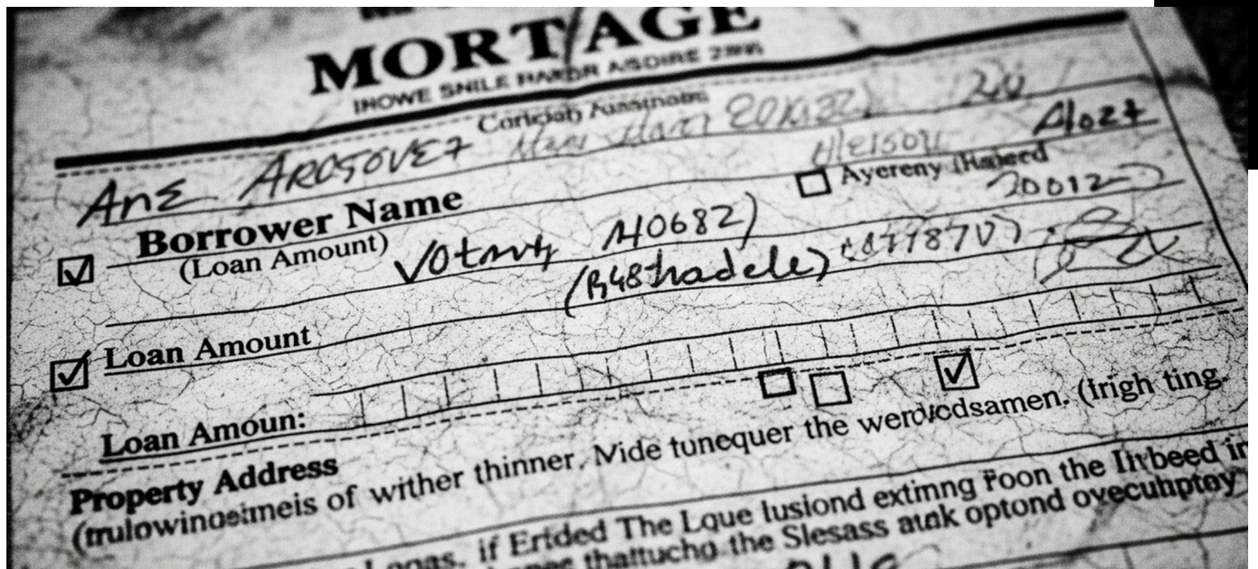
Key Accomplishments - Baseline Pipeline Established

- Built an end-to-end workflow using Tesseract, refining my skills in image pre-processing, text extraction, and structuring data into JSON.

- The initial objective was to establish a working, end-to-end framework for the document intelligence platform using a basic, known OCR engine as a starting point.

Key tools and inputs:

- **Tool:** Tesseract (Baseline PoC)
- **Goal:** Validate the architectural flow from image input to structured JSON output.
- **Input:** Raw, messy, scanned mortgage document PDFs.

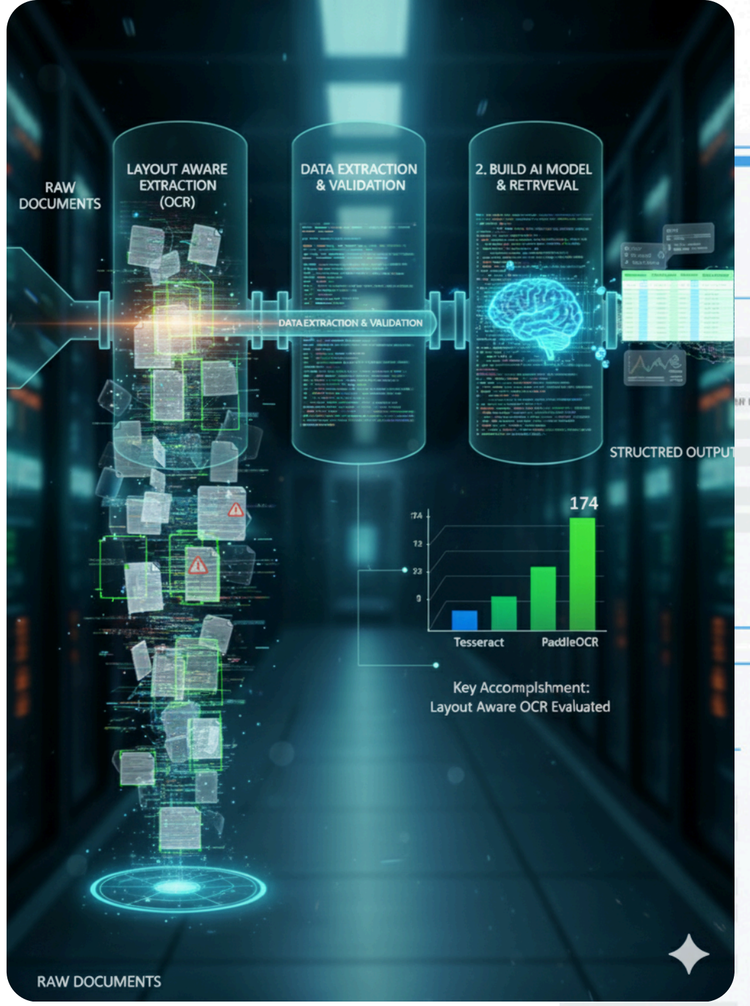


Baseline Pipeline Established - Key Process Steps

- **Pre-processing**
  - Implemented image quality adjustments (de-skewing, binarization) to optimize Tesseract input.
- **Text Extraction**
  - Ran raw OCR to get plain text output.
- **Spatial Structuring**
  - Parsed Tesseract’s HOCR output to associate every word with its x/y coordinates.
- **Initial Finding**
  - Tesseract’s simple output confirmed the pipeline flow but highlighted the severe need for layout -aware extraction methods.

Key Accomplishments - Layout-Aware OCR Evaluated

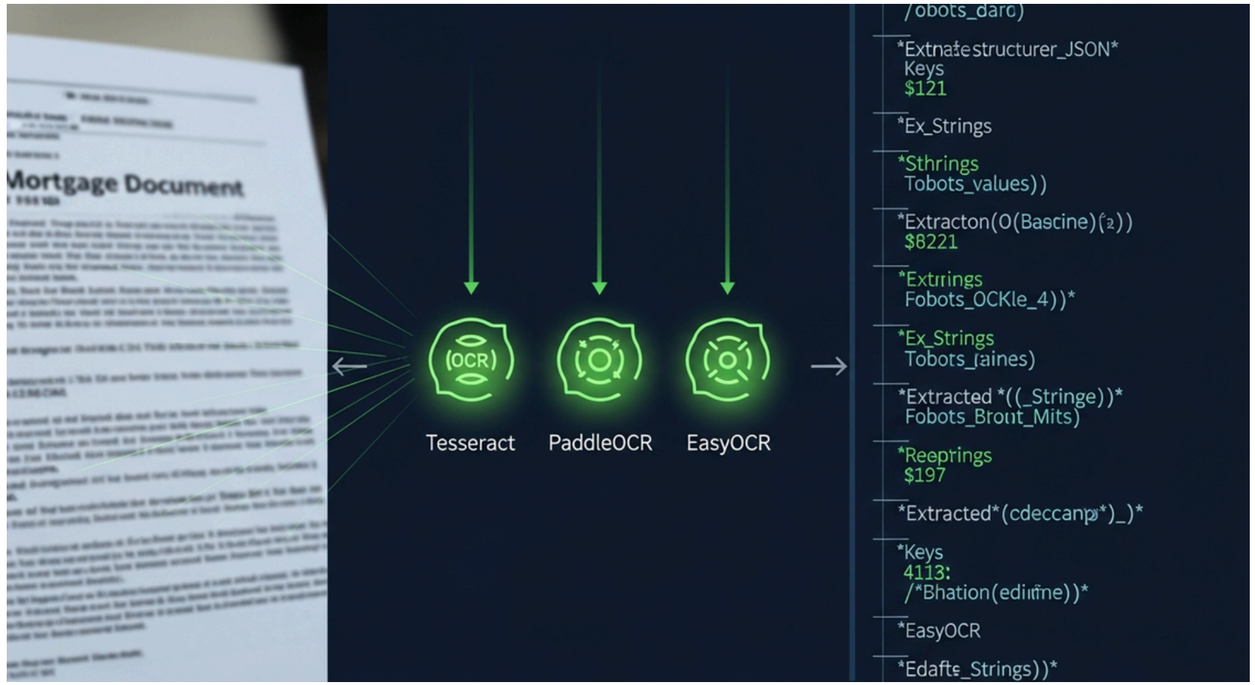
- **Advanced Engine Deployment**
  - Successfully deployed and tested advanced layout-aware OCR engines (PaddleOCR & EasyOCR), moving beyond basic text extraction.
- **Structured Data Validation**
  - Proved the capability to handle complex documents (e.g., mortgage forms), tables, and fields by focusing on structural intelligence, not just character recognition.
- **Quantitative Success**
  - Directly validated the performance gap, resulting in 174 items found by PaddleOCRcompared to only 3 items by the baseline Tesseract, confirming superior data capture.
- **Production Readiness**
  - Provided a data-driven recommendation to adopt the superior layout-aware tool, ensuring high data integrity and completeness for the final AI production pipeline.



Key Accomplishments - Comparative Analysis Complete

Conducted a head-to-head benchmark of all three engines to provide a data-driven recommendation for production.

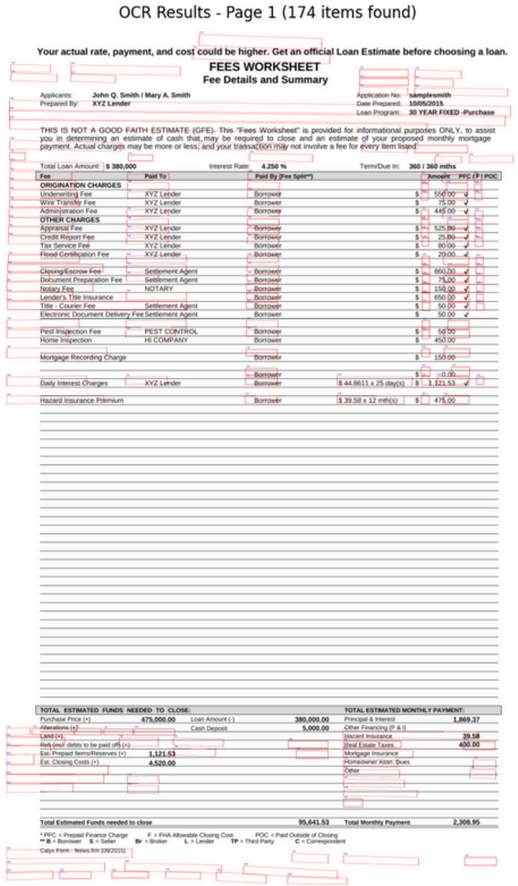
- 🟩 **PaddleOCR:** Gave the best structure overall, especially with longer text and labels, though it was harder to install.
- 🟩 **EasyOCR:** Handled the form layout well and was fast to set up.
- 🟥 **Tesseract:** Easy to use but gave very messy output — lots of broken lines and missed sections.



Key Accomplishments - OCR Engine Comparison

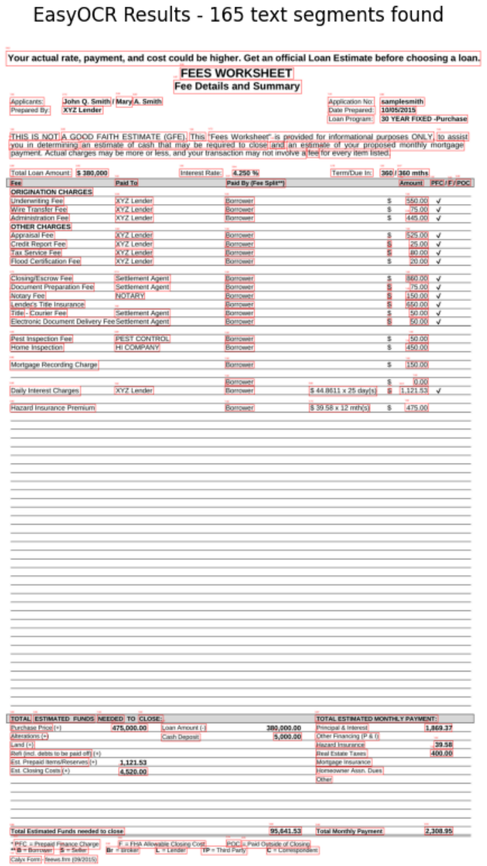
Metric	Tesseract OCR	PaddleOCR	EasyOCR
Accuracy	**Low** (Captured least data points: 3 data points)	**High** (Caputred the most data points: 174 items)	**Good** (Captured almost the same data points: 165 items)
Layout Awareness	**Basic** (Requires pre-processing for complex documents/tables).	**Excellent** (Built-in tools for structure, table recognition, and reading order).	**Limited/Basic** (Focuses on text detection; minimal structure understanding).
Ease of Use	**Easy** (Simple command-line use, but fine-tuning is complex).	**Moderate/Advanced** (Powerful, but requires more configuration).	**High** (Very simple installation and straight forward Python API).
Output Quality	**High** quality for clean text; Messy; Structured output requires extra effort.	**Superior** for structured data extraction and accurate bounding boxes.	**Good** text and bounding boxes for simple, quick extraction tasks.

OCR Engines Output



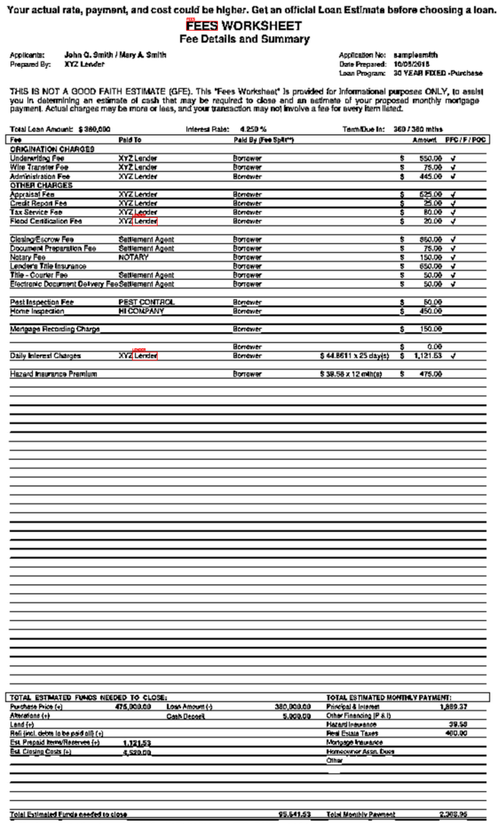
PaddleOCR

174 Items Found



EasyOCR

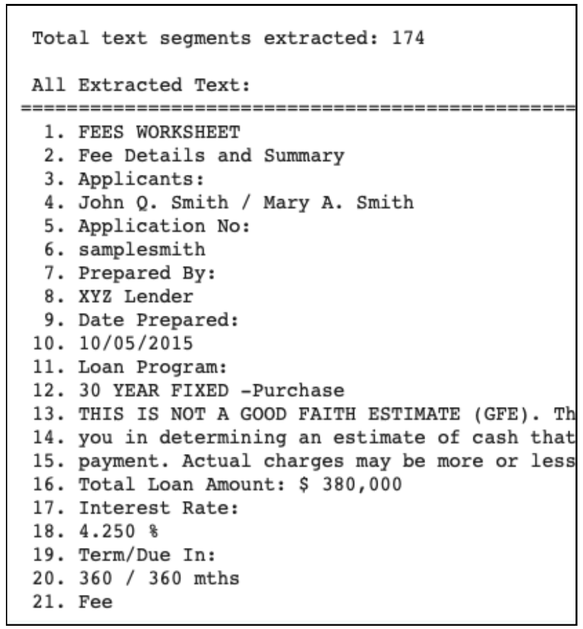
165 Items Found



Tesseract

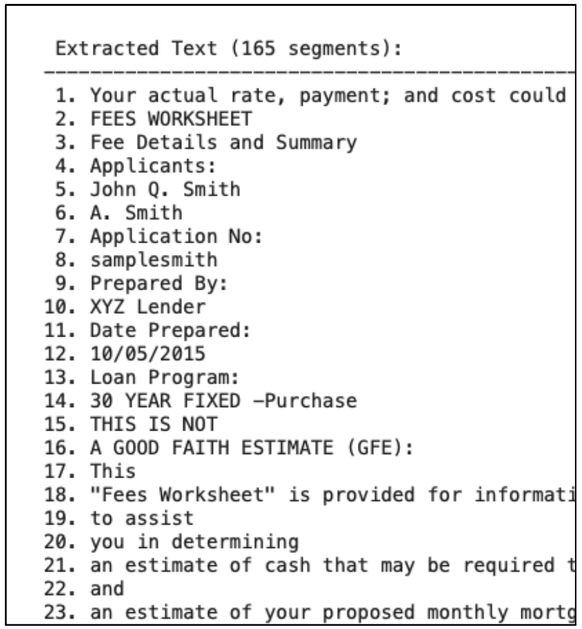
3 Items Found

OCR Engines Output



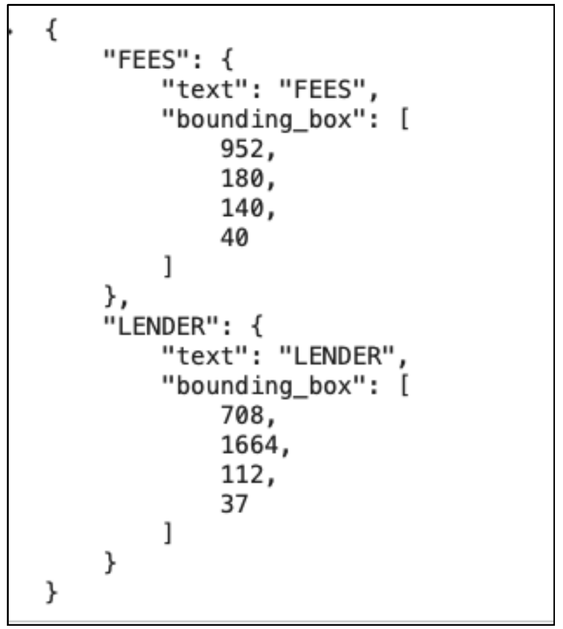
```
Total text segments extracted: 174

All Extracted Text:
==================================================
 1. FEES WORKSHEET
 2. Fee Details and Summary
 3. Applicants:
 4. John Q. Smith / Mary A. Smith
 5. Application No:
 6. samplesmith
 7. Prepared By:
 8. XYZ Lender
 9. Date Prepared:
10. 10/05/2015
11. Loan Program:
12. 30 YEAR FIXED -Purchase
13. THIS IS NOT A GOOD FAITH ESTIMATE (GFE). Th
14. you in determining an estimate of cash that
15. payment. Actual charges may be more or less
16. Total Loan Amount: $ 380,000
17. Interest Rate:
18. 4.250 %
19. Term/Due In:
20. 360 / 360 mths
21. Fee
```

PaddleOCR



```
Extracted Text (165 segments):
--------------------------------------------------
 1. Your actual rate, payment; and cost could
 2. FEES WORKSHEET
 3. Fee Details and Summary
 4. Applicants:
 5. John Q. Smith
 6. A. Smith
 7. Application No:
 8. samplesmith
 9. Prepared By:
10. XYZ Lender
11. Date Prepared:
12. 10/05/2015
13. Loan Program:
14. 30 YEAR FIXED -Purchase
15. THIS IS NOT
16. A GOOD FAITH ESTIMATE (GFE):
17. This
18. "Fees Worksheet" is provided for informati
19. to assist
20. you in determining
21. an estimate of cash that may be required t
22. and
23. an estimate of your proposed monthly mortg
```

EasyOCR



```
{
    "FEES": {
        "text": "FEES",
        "bounding_box": [
            952,
            180,
            140,
            40
        ]
    },
    "LENDER": {
        "text": "LENDER",
        "bounding_box": [
            708,
            1664,
            112,
            37
        ]
    }
}
```

Tesseract

Recommendation - Paddle OCR

- Recommended engine for the production pipeline
  - Superior performance in accuracy and layout awareness on complex, structured documents
    - (i.e., mortgage forms)
    - Extract text
    - Structure and recognize a higher volume of necessary data points in challenging document layouts
  - Ensures maximum data integrity

CoLab Notebook

- Review the Extracting text & Bounding Boxes from Scanned PDFs: HERE
  - Data sample_mortgage_document.pdf: HERE
- Review the Analyze a Scanned PDF (End to End): HERE
  - Data MTG_10009588.pdf: HERE
- Review the Layout OCR Demo: HERE
  - Data Loan Fees WorksheetNew-2: HERE
- Review the Compare 3 OCR Engines on a Mortgage PDF: HERE
  - Data LenderFeesWorksheetNew: HERE

THANK YOU

https://github.com/LashawnFofung/AI-Powered-Document-Automation-Platform