EXTERN + OUTAMATION
EXTERNSHIP

ADVANCED AI-POWERED

DOCUMENT INSIGHT AND DATA EXTRACTION PROJECT

SPRINT 3 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

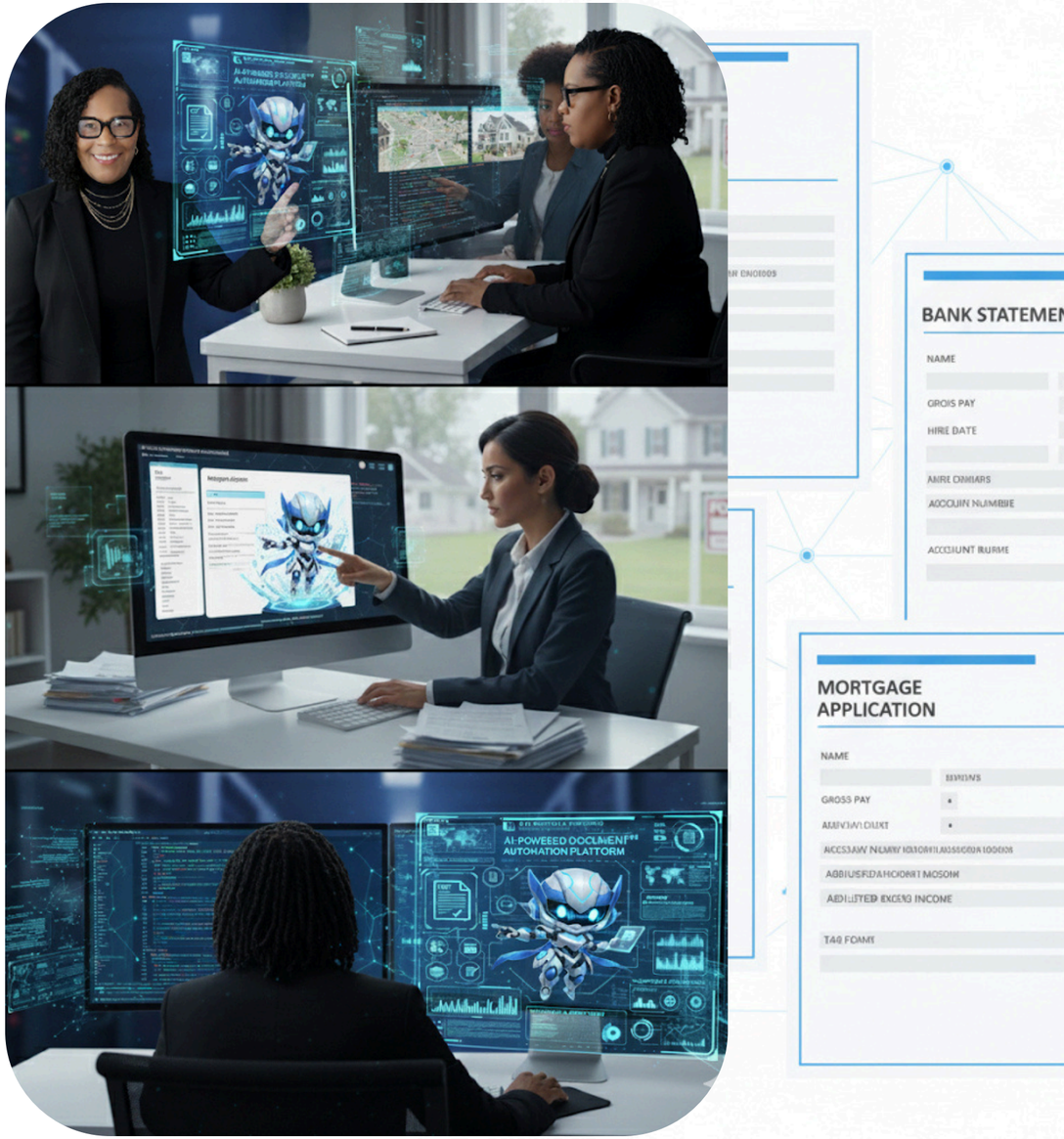
10 WEEKS

Python Data Extraction

PyMuPDF, Regex, Anchor Phrases, Structured Parsing

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)

These foundational skills, mastered in Sprint 3, directly enable the 4-Steps Document Intelligence (AI) Pipeline designed to solve our core business challenge.

SPRINT 3 FOCUS

Step 2: Data Extraction & Preparation

Step 2: Data Extraction & Preparation

- Clean, validate, and standardize the messy OCR output
- Handle inconsistencies, missing values, and process complex JSON structures using Python/Pandas

Overall Status: Sprint 3 Complete

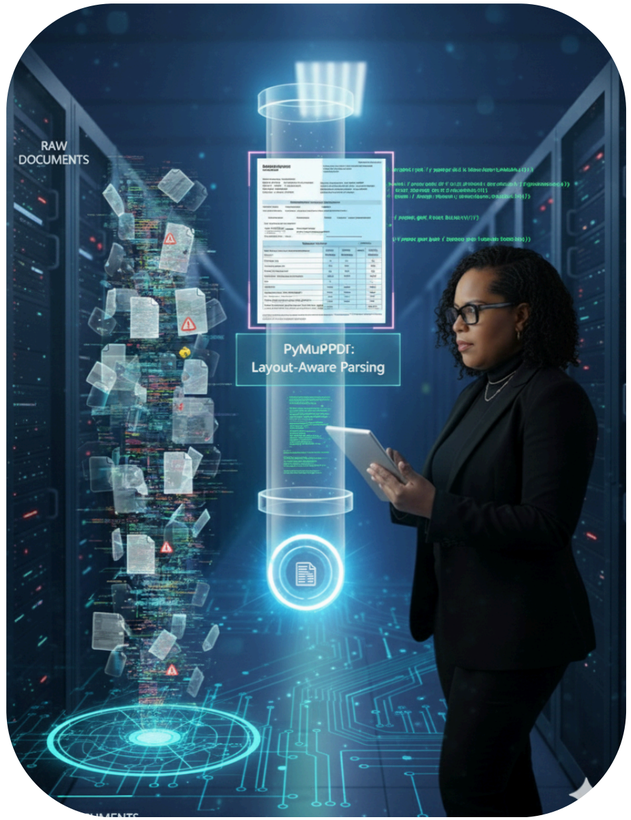
I've successfully concluded Sprint 3, focused on mastering Python Data Extraction techniques crucial for the AI-Powered Document Automation Platform.

- The primary goal was to acquire hands-on coding skills in tackling raw, messy, unstructured data from complex mortgage documents, transforming it into structured information and spatial data.

- This work lays the groundwork for subsequent AI modeling phases.

- This sprint significantly advanced my ability to transform this raw, unstructured PDF content into clean, AI-ready data, directly tackling the challenge of automating manual document processing.

Overall, Sprint 3 successfully established a robust and adaptable foundation for intelligent data extraction, marking a critical step towards our AI automation goals.



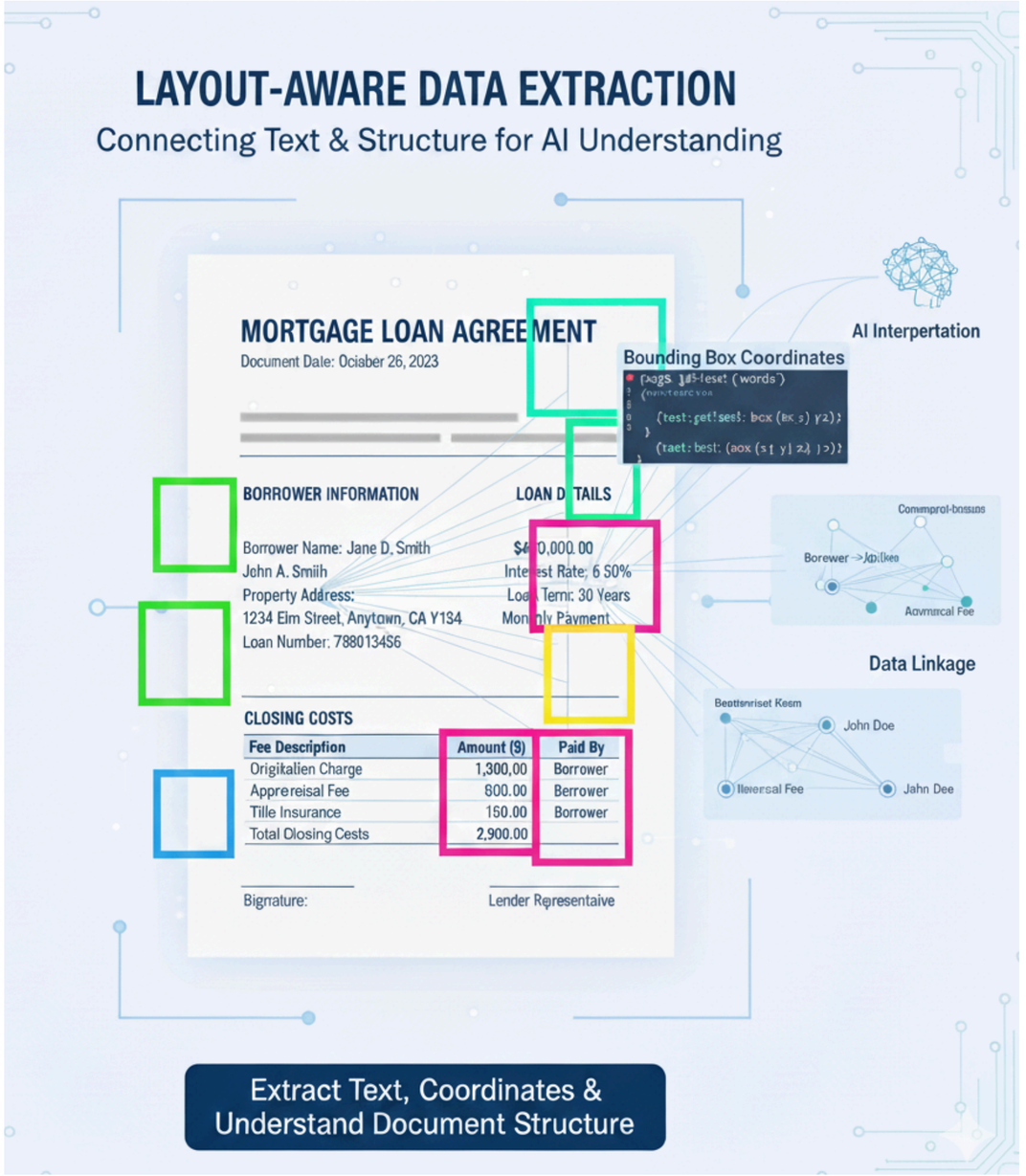
Key Accomplishments - Technical & Platform Development

My progress seamlessly moved from understanding PDF structures to applying advanced extraction methods across two crucial areas:

- Layout-Aware Data Extraction
- Hybrid Extraction Pipeline Development
- Tool Evaluation

Layout-Aware Data Extraction

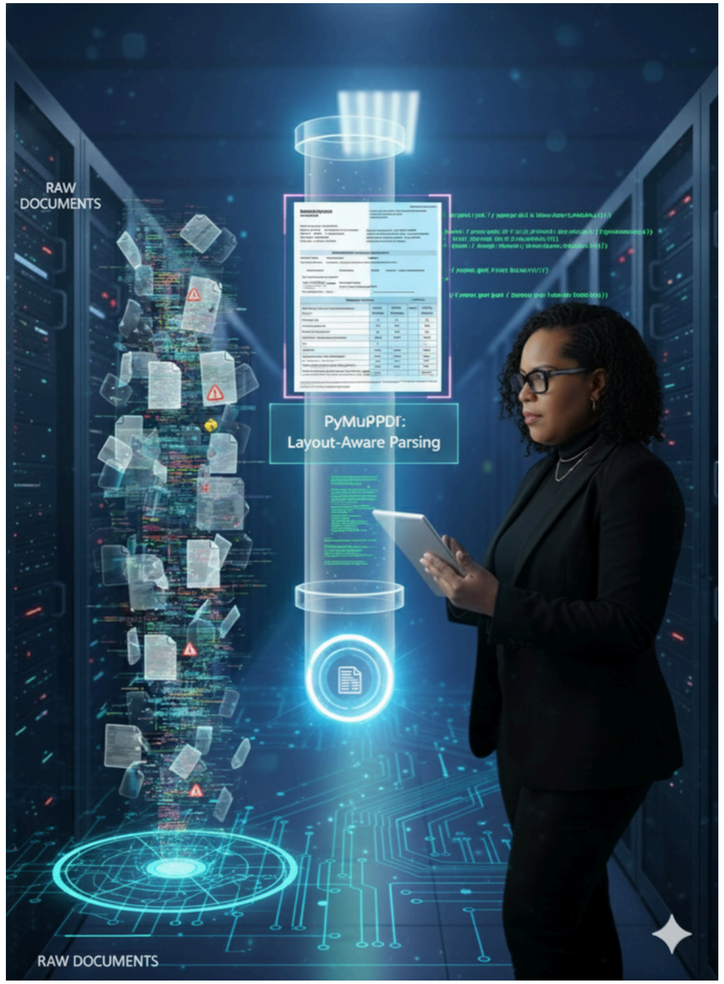
- Implemented effective techniques to extract text, values, and their precise bounding box coordinates.
- This spatial information is fundamental for understanding document structure and context, enabling more accurate data linkage and interpretation in later AI stages.
- Image of Mortgage Loan Agreement
  - Illustrates our effective technique for Layout-Aware Data Extraction.
  - By identifying text and its exact location via bounding boxes, we gain critical spatial context.
  - This structured information is fundamental for AI models to accurately link data, interpret document context, and unlock deeper insights.



Hybrid Extraction Pipeline Development

Designed and tested a rule-based pipeline that combines multiple techniques to achieve high-precision data extraction from unstructured mortgage documents:

- **Regular Expressions** (**Regex**): Used for extracting field values based on pattern matching.
- **Anchor Phrases:** Leveraged text labels (e.g., "Loan Amount:") to pinpoint and pull the correct values.
- **Structured Table Parsing:** Implemented logic to cleanly extract table data, such as itemized fees from sections like "Origination Charges".



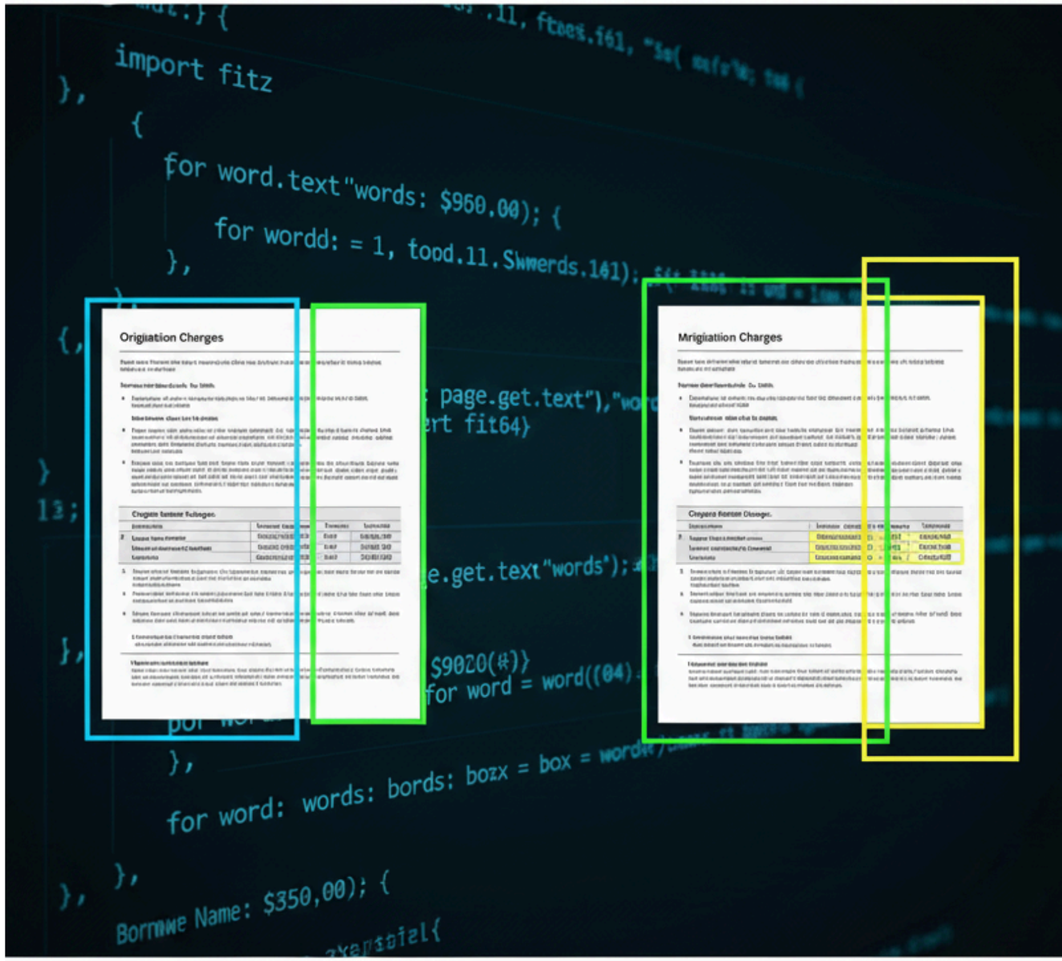
Tool Evaluation

- Performed a comparative analysis of key Python PDF libraries.
- Identified PyMuPDF as the most versatile for layout-aware extraction and pdfplumber as strong for table parsing.

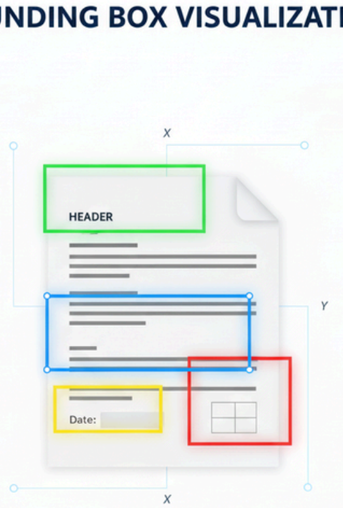
Key Accomplishments - Key Learning

- **Mastered Bounding Boxes**
  - Understand the critical role of bounding boxes in linking labels to values and handling multi-column layouts, overcoming the challenge of text losing its spatial context during extraction.

- **Debugged "Clumping" Issues**
  - Identified and mitigated the "clumping" problem common in scanned PDFs, where multiple fields are grouped into one large bounding box.
  - The solution involved using a hybrid approach.
  - Relying on specific label-searching within the large text blocks.



Understanding Bounding Boxes in Document Processing



- **What they are:** Rectangular coordinates precisely defining the location of text, images, or fields on a page.
- **Why they matter:** Add crucial spatial intelligence to extracted data, revealing document layout.
- **Key benefits for AI:**
  - Enable layout analysis (columns, sections).
  - Facilitate accurate key-value pairing (linking labels to values).
  - Aid in table structure reconstruction.
  - Provide visual context for AI models, mimicking human understanding.
- **Result:** Transforms raw text into layout-aware, structured data for deeper insights.

Key Metrics & Progress

In this sprint, I worked on tasks:

- **Extract & Structure Data from Mortgage PDF**
  - Extracted text and bounding boxes from a multi-page mortgage document, structuring the output (text + coordinates) into a clean format (e.g., JSON) for AI consumption.
    - View Github:
      - https://github.com/LashawnFofung/Python-Document-Preparation-and-Extraction/blob/main/src/Task_Extract_and_Structure_Data_from_Mortgage_PDFs.ipynb

- **Extract Key Fields from the Loan Worksheet**
  - Built the full mini-pipeline to extract both single, basic fields (e.g., Loan Amount, Applicant Name) and complex table-based fields (e.g., fees) from a Loan Worksheet PDF.
    - View Github:
      - https://github.com/LashawnFofung/Python-Document-Preparation-and-Extraction/blob/main/src/Task_Extract_Key_Fields_from_the_Loan_Worksheet.ipynb

THANK YOU

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