Textro.ai

PDF-to-Markdown Conversion Pipeline

Automated extraction and structuring of PDF content

Company:	Textro AI
Assigned to:	Esai Keshav
Start Date:	16/09/2025
End Date:	22/09/2025

Focus on your research, analysis, and approach this task is to understand your perspective, not just results.

Contents

1 Introduction

This project aims to develop a system that automatically converts any PDF file into a well-structured Markdown file (`.md`). The output must preserve headings, paragraphs, lists, figures, tables, and document hierarchy.

2 Objective

- Build a pipeline that extracts structured content from PDFs.
- Ensure Markdown output preserves:
  - Headings (# to ######)
  - Paragraphs
  - Lists
  - Figures with captions
  - Tables
  - Document hierarchy

3 Task Scope

- The system must be built using classical programming, parsing, and layout analysis techniques using **deep learning**.
- **No LLM usage is allowed** for text extraction, structuring, or Markdown generation.
- All outputs should be derived from direct PDF parsing, OCR, or rule-based layout analysis methods.

4 Automated Pipeline

1. **PDF Preprocessing:** Render each page to image for layout analysis. Extract embedded text if available; otherwise apply OCR.
2. **Layout Analysis:** Detect structural elements (titles, headings, paragraphs, lists, figures, tables, captions).
3. **Text Grouping & Recognition:** Merge text lines into blocks, handle multi-column layouts, remove headers/footers.
4. **Hierarchy Derivation:** Assign heading levels H1-H5 based on font size, boldness, indentation, or structural cues.
5. **Markdown Conversion:** Map layout classes into Markdown tags. Insert figures/tables with captions and preserve indentation.
6. **Output Assembly:** Concatenate outputs page by page and save as a clean `.md` file.

5 Deliverables

Deliverable	Description	Due Date
System implementation	Full pipeline code for PDF-to-Markdown conversion.	22/09/2025
Documentation	Report covering architecture, pipeline stages, and examples.	22/09/2025
Sample outputs	Example PDFs converted to Markdown (with images/tables).	22/09/2025
GitHub repository	Code with README, requirements, and usage instructions.	22/09/2025

Table 1: Deliverables and Timeline

6 Contact / Questions

Project owner: **Naveen V**.

For clarifications, reach out to the project owner.

*** This document is meant to guide you, but we encourage you to explore, experiment, and share your unique perspective. Your creativity and thoughtful approach are just as important as the outcome were excited to see how you solve this! ***