Adobe India Hackathon 2025 – Connecting the Dots

# Round 1A – PDF Outline Extraction

In this round, we were asked to extract a structured outline from a given PDF document. This included capturing all the headings (H1, H2, H3) along with their corresponding page numbers.

To solve this, we used a Python script built on top of PyMuPDF (fitz) to traverse the document’s outline tree. The goal was to collect all the hierarchical headings and return them in a clean JSON format.

Here’s what we did:

* • Loaded the PDF using PyMuPDF
* • Parsed the outline tree recursively
* • Mapped headings to their page numbers
* • Formatted the final result as a structured JSON

The script was containerized and designed to run inside a CPU-only Docker container within 60 seconds.

Below is a snapshot of the terminal output:

