

Sample PDF Document with Headings

1. Introduction

This is a sample PDF document that includes headings of different levels. It is intended for testing PDF heading extraction capabilities in the RAG system.

1.1 Background

This document was generated programmatically using ReportLab, a PDF generation library for Python. It contains headings with different font sizes to test the heading detection algorithm.

2. Implementation

The implementation section describes how the RAG system processes PDF documents and extracts structural information from them.

2.1 Architecture

The PDF processing architecture consists of several components: PDF parsing, font analysis, heading detection, and metadata extraction.

2.2 Technical Details

The system uses pdfminer.six to analyze the PDF structure and extract text. Font size statistics are calculated to identify headings.

3. Conclusion

This sample PDF document demonstrates different heading levels that should be detected by the PDF heading extraction algorithm.