Understanding Modern Documents

Introduction

Background

PDF Structure

This section describes the internal structure of a PDF file including objects, streams, and cross-reference tables.

Document Parsing Techniques

Rule-Based Parsing

Rule-based systems rely on fixed heuristics such as font size, position, and indentation.

ML-Based Parsing

Machine learning approaches can generalize better to unseen document layouts.

Challenges

Multilingual Documents

Processing documents with multiple languages is challenging due to script and direction differences.

Inconsistent Layouts

Layout variance across PDFs requires robust generalization.