A Comprehensive Analysis of Document Structures

1. Introduction to Document Parsing

This section introduces the fundamental concepts of parsing unstructured documents like PDFs.

1.1. The Challenge of PDFs

PDFs are designed for consistent visual presentation, not for easy data extraction. This poses a significant challenge for automated systems.

1.1.1. Font and Layout Heuristics

One common approach is to use heuristics based on font size, weight, and text position to infer the document's structure. This is the core of our current task.

1.1.2. Rule-Based Systems

These systems rely on a predefined set of rules to classify text elements. They are fast but can be brittle when encountering new layouts.

2. Methodologies for Extraction

This chapter explores different methods to extract structured information.

2.1. Statistical Approaches

Statistical methods analyze the frequency and properties of text elements to build a model of the document's structure. This can be more robust than fixed rules.