PDFParser API Reference

Version: 1.0.0
Last Updated: 2025-01-15

Overview

PDFParser is a Python library for extracting structured content from PDF documents. It provides a simple, intuitive API for parsing PDFs into machine-readable formats.

Installation

Install via pip:

```
pip install pdfparser
```

Or install from source:

```
git clone https://github.com/example/pdfparser.git
cd pdfparser
pip install -e .
```

Quick Start

Here's a simple example to get you started:

```
from pdfparser import PDFParser

# Initialize parser
parser = PDFParser()

# Parse a PDF file
document = parser.parse("document.pdf")

# Access extracted content
print(f"Title: {document.metadata.title}")
```

API Reference

PDFParser

The main class for PDF parsing.

Constructor

```
PDFParser(config: Optional[PDFConfig] = None)
```

Parameters:

? config (PDFConfig, optional) - Configuration options for parsing

parse()

```
parse(
    pdf_path: Union[str, Path],
    category: Optional[DocumentCategory] = None
) -> SimpleDocument
```

Parse a PDF file and extract structured content.

Parameters:

? pdf_path (str or Path) - Path to PDF file
? category (DocumentCategory, optional) - Document category hint

Returns:

? SimpleDocument - Parsed document with extracted content

Raises:

? FileNotFoundError - If PDF file doesn't exist
? PDFParseError - If parsing fails

SimpleDocument

Represents a parsed document with extracted content.

Examples

Extract Tables

```
from pdfparser import PDFParser

parser = PDFParser()
doc = parser.parse("report.pdf")

for table in doc.tables:
    print(f"Table {table.id}:")
    print(f"  Caption: {table.caption}")
    print(f"  Rows: {len(table.data)}")
```

Extract Figures

```
from pdfparser import PDFParser

parser = PDFParser()
doc = parser.parse("paper.pdf")

for i, figure in enumerate(doc.figures):
    print(f"Figure {i+1}: {figure.caption}")
```

Custom Configuration