Python PDF Parsing Tutorial

1. Introduction

This is a sample PDF document created for practicing PDF parsing with Python. PDF (Portable Document Format) is a file format developed by Adobe that presents documents in a manner independent of application software, hardware, and operating systems.

2. Key Features

Text Extraction: Extract plain text from PDF documents.
Metadata Access: Read document properties like author, title, and dates.
Page Manipulation: Split, merge, and rotate PDF pages.
Table Extraction: Parse structured table data from PDF files.

3. Python PDF Libraries

Library	Strengths	Best For
PyPDF2	Simple API, PDF manipulation	Basic text extraction
pdfplumber	Table extraction, Layout	Complex documents
PyMuPDF	High performance, Images	Large files

4. Sample Content for Testing

This page contains sample text for practicing text extraction. You can use various Python libraries to parse this content. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. **Important:** Always verify the extracted text for accuracy. Some PDFs may have encoding issues or non-standard fonts that can affect extraction quality.

Code Example:

```
import PyPDF2

with open('document.pdf', 'rb') as file:
reader = PyPDF2.PdfReader(file)
text = reader.pages[0].extract_text()
print(text)
```