Parsing PDF files in Python

Parsing PDF files in Python can be useful in many contexts, such as extracting data from documents or indexing content. In this article, we'll see how to do it using one of the most common libraries: **PyMuPDF** (also known as `fitz`).

Installing the Library

First, you need to install the `PyMuPDF` library. It can be easily installed via pip:

```
pip install pymupdf
```

Parsing a PDF File

Once the library is installed, we can start extracting text from a PDF. The following code shows a basic example:

```
import fitz  # PyMuPDF

def extract_text_from_pdf(file_path):
    doc = fitz.open(file_path)
    pdf_text = ""

    for page in doc:
        pdf_text += page.get_text()

    doc.close()
    return pdf_text
```

```
# Example usage
pdf_path = "document.pdf"
text = extract_text_from_pdf(pdf_path)
print(text)
```

Code Explanation

- `fitz.open`: opens the PDF file.
- The loop `for page in doc` iterates over all pages in the PDF.
- `page.get_text()`: extracts the text from each page.
- `doc.close()`: closes the file to free resources.

Conclusions

Using PyMuPDF is a simple and fast way to read the content of PDF files in Python. The library also offers more advanced features such as extracting images, searching text, and manipulating layout, which can be useful in more complex projects.