Extractext

1. Importing Libraries

~~~
```python
import fitz  # PyMuPDF
~~~

- This line imports the `fitz` module, which is part of the PyMuPDF library. PyMuPDF is a powerful library for working with PDF files. You'll need to install it if you haven't already: `pip install pymupdf`

2. Defining the Extraction Function

```
def extract_text_from_pdf(pdf_path):
    doc = fitz.open(pdf_path)
    text = "\n".join([page.get_text() for page in doc])
    return text
```

- **`def extract_text_from_pdf(pdf_path):`**: This defines a function named `extract_text_from_pdf` that takes one argument:
    - `pdf_path`: The file path to the PDF file you want to extract text from.
- **`doc = fitz.open(pdf_path)`**: This line opens the PDF file specified by `pdf_path` using `fitz.open()`. The opened PDF document is stored in the `doc` variable.
- **`text = "\n".join([page.get_text() for page in doc])`**: This is the core of the text extraction. Let's break it down:
    - `[page.get_text() for page in doc]`: This is a list comprehension. It iterates through each `page` in the `doc` (the PDF document). For each page, it calls the `page.get_text()` method to extract the text content of that page. This creates a list where each element is the text from a single page.
    - `"\n".join(...)`: The `join()` method takes the list of strings (the page texts) and concatenates them into a single string. `"\n"` specifies that the pages' text

should be joined with a newline character (`\n`) in between them. This preserves the page breaks and formatting to some extent.

- **`return text`**: The function returns the complete extracted text as a single string.

3. Example Usage

```
pdf_text = extract_text_from_pdf("/Users/vinod/Desktop/mike/sample.pdf")
print(pdf_text[:500])  # Print first 500 characters
```

- **`pdf_text = extract_text_from_pdf("/Users/vinod/Desktop/mike/sample.pdf")`**: This line calls the `extract_text_from_pdf` function, passing the path to your PDF file. **Important:** Replace `"/Users/vinod/Desktop/mike/sample.pdf"` with the actual path to your PDF file. The returned text is stored in the `pdf_text` variable.
- **`print(pdf_text[:500])`**: This line prints the first 500 characters of the extracted text. This is useful for quickly previewing the extracted content. You can remove the `[:500]` to print the entire text.

This code provides a clean and efficient way to extract text from PDFs using PyMuPDF. Remember to install the library and replace the placeholder PDF path with your actual file path.