处理PDF文件并通过DeepSeek API分析Markdown文件

### 使用指南

#### 项目结构

1. your\_project/
2. │
3. ├── deepseek\_parse.py # 解析DeepSeek V3 API的Python脚本
4. ├── process\_pdfs.py # 处理PDF文件并转化为Markdown文件的Python脚本
5. ├── pdftomd.py # 用于将PDF转为Markdown的脚本
6. ├── apikey.txt # 存放DeepSeek API密钥的文件
7. ├── markdown/ # 存放转化后的Markdown文件
8. ├── input/ # 存放原始pdf文件
9. └── answer/ # 存放DeepSeek V3 API返回结果的文件夹

#### 依赖

1. **Python 3.x**：本项目依赖Python 3.x版本进行运行。
2. **依赖库**：
   * openai：用于与DeepSeek API交互。
   * subprocess：用于调用其他Python脚本（如pdftomd.py）。

可以通过以下命令安装所需的依赖库：

1. pip install openai

#### 项目功能

本项目包括两个主要任务：

1. **将PDF文件转换为Markdown格式**：
   * process\_pdfs.py脚本会自动扫描当前目录中的所有PDF文件，并使用pdftomd.py脚本将其转换为Markdown文件。转换后的文件将被保存在markdown文件夹中。
2. **通过DeepSeek V3 API分析Markdown文件**：
   * process\_pdfs.py脚本在转换PDF文件后，会继续调用deepseek\_parse.py脚本，使用DeepSeek V3 API分析markdown文件夹中的每一个Markdown文件，并将分析结果保存在answer文件夹中。

#### 配置步骤

1. **获取API密钥**：
   * 访问[SiliconFlow](https://cloud.siliconflow.cn/account/ak)并获取API密钥。
   * 将API密钥保存到apikey.txt文件中。文件中只需要包含API密钥本身。
2. **安装依赖**：
   * 确保你的Python环境中已安装openai库：
3. pip install openai
4. **准备输入文件**：
   * 将需要转换的PDF文件放入input文件夹中。
   * PDF文件将被process\_pdfs.py脚本自动处理。
5. **使用process\_pdfs.py脚本**：
   * 运行process\_pdfs.py脚本，首先将PDF文件转换为Markdown文件，然后调用DeepSeek V3 API处理每个Markdown文件，并保存结果：
6. python process\_pdfs.py

#### 运行流程

1. **PDF转Markdown**：
   * process\_pdfs.py脚本会遍历当前目录下的所有PDF文件，并通过调用pdftomd.py脚本将其转换为Markdown文件。
   * 转换后的Markdown文件将保存到markdown文件夹中。
2. **调用DeepSeek V3 API**：
   * process\_pdfs.py脚本将继续调用deepseek\_parse.py脚本，通过DeepSeek V3 API解析转换后的Markdown文件。
   * 解析结果（文本格式）将被保存到answer文件夹中。

#### 详细步骤

1. **运行process\_pdfs.py脚本**： 这个脚本会自动执行以下操作：
   * 扫描当前目录下所有.pdf文件。
   * 使用pdftomd.py将每个PDF文件转换为Markdown文件，保存到markdown文件夹中。
   * 对每个Markdown文件，使用deepseek\_parse.py脚本调用DeepSeek V3 API，并将结果保存在answer文件夹中。

运行命令：

1. python process\_pdfs.py
2. **DeepSeek V3 API响应格式**：
   * 每个Markdown文件的解析结果将作为纯文本（txt）格式保存。
   * 结果会按照以下格式返回：
3. B73 RefGen\_V4\tZm00001d028850\tP1\tAX-86284808, AX-91425354\tCob color\tDetection of QTNs for kernel moisture concentration and kernel dehydration rate before physiological maturity in maize using multi-locus GWAS
4. **结果存储**：
   * API返回的结果将作为纯文本保存在answer文件夹中，每个文件对应一个Markdown文件。

#### 脚本详情

1. **process\_pdfs.py**：
   * 自动将当前目录下所有PDF文件转化为Markdown文件。
   * 使用pdftomd.py将PDF转化为Markdown。
   * 调用deepseek\_parse.py脚本，通过DeepSeek V3 API分析Markdown文件并保存结果。
2. **pdftomd.py**：
   * 用于将PDF文件转换为Markdown格式。
3. **deepseek\_parse.py**：
   * 使用DeepSeek V3 API对Markdown文件进行分析。
   * API结果以纯文本（txt）格式返回，并保存在answer文件夹中。

#### 示例命令

1. **转换PDF文件并分析Markdown文件**：
2. python process\_pdfs.py
3. **使用DeepSeek V3 API解析Markdown文件**：每个Markdown文件将被传递到DeepSeek V3 API，API将返回解析的内容并保存为纯文本文件。
4. python deepseek\_parse.py --api\_key\_file "apikey.txt" --input\_file "markdown/example.md" --output\_file "answer/example.txt"

#### 注意事项

* **文件路径**：确保input文件夹中包含要转换的PDF文件，markdown和answer文件夹将分别保存转换后的Markdown文件和API响应结果。
* **API密钥**：在运行脚本前，请确保apikey.txt文件中已经正确存储了DeepSeek V3 API密钥。
* **结果存储**：API返回的结果将以纯文本（txt）格式保存在answer文件夹中，每个文件对应一个Markdown文件。