ollama 安装流程

```
#建议 docker 安装，比较方便
#1、拉取镜像
#docker hub 地址，选则自己需要的版本
#https://hub.docker.com/r/ollama/ollama/tags
#我安装的是最新版
docker pull ollama/ollama:latest

#2、启动 docker
docker run -d --gpus=all -v ollama:/root/.ollama -p 11434:11434 --name ollama ollama/ollama

#3、运行模型
#其中 ollama run llama3 为运行模型的命令，如果需要使用其他模型，可以参考官方提供模型版本 https://github.com/ollama/ollama，将 llama3 替换掉即可
docker exec -it ollama ollama run llama3
```

Langchain 安装

```
conda create --name rag python=3.10
conda activate rag
pip install langchain
pip install streamlit
pip install chromadb
pip install pypdf
pip install langchain_community
pip install chromadb
```

核心代码

```
#核心代码主要包含两个模型，检索模型和生成模型，详见下文备注，模型地址为使用上面 ollama 启动的模型地址

from langchain.chains import RetrievalQA
from langchain.callbacks.streaming_stdout import StreamingStdOutCallbackHandler
from langchain.callbacks.manager import CallbackManager
from langchain_community.llms import Ollama
from langchain_community.embeddings.ollama import OllamaEmbeddings
from langchain_community.vectorstores import Chroma
from langchain.text_splitter import RecursiveCharacterTextSplitter
from langchain_community.document_loaders import PyPDFLoader
from langchain.prompts import PromptTemplate
from langchain.memory import ConversationBufferMemory
import streamlit as st
import os
import time

if not os.path.exists('files'):
    os.mkdir('files')

if not os.path.exists('jj'):
    os.mkdir('jj')

```

```
#执行这个命令运行即可启动服务
streamlit run app.py
#接着在浏览器访问服务地址
http://localhost:8501
```

上传 PDF 即可进行问答对话