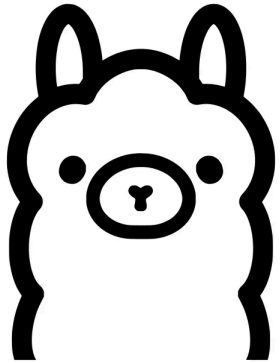


Ollama

本地快速上手大型语言模型

(Llama 3.1, Mistral, Gemma 2, ...)

使用 LangChain 调用 Ollama 运行模型

操作步骤

- 建立 Python LangChain 的开发环境
- 编写 main.py 程序
- 运行确认

API 文档

https://python.langchain.com/v0.2/api_reference/ollama/chat_models/langchain_ollama.chat_models.ChatOllama.html

操作演示

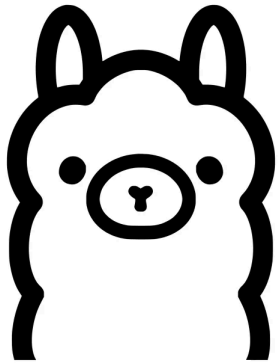
课堂实验

建立 Python LangChain 的开发环境

```
# 虚拟环境
python -V
python -m venv _pvenv_
source _pvenv_/bin/activate
python -m pip install --upgrade pip

# 安装库
pip install \
  langchain==0.2.14 \
  langchain-ollama==0.1.1

# ollama 模型确认
ollama ls
```

main.py

```python
import time
from langchain_ollama import ChatOllama

def evalEndTime(start_time):
    end_time = time.time()  # 获取结束时间
    execution_time = "(程序运行时间：%.2f 秒)" % (
        end_time - start_time
    )  # 计算程序运行时间
    return execution_time

print("=" * 100)
start_time = time.time()  # 获取开始时间

llm = ChatOllama(
    model="llama3.1:8b",
    # model="gemma2:2b",
    # model="gemma2:9b",
    temperature=0.5,
)
print(">", llm)

client_prompt = "桃园结义是几个人？都是谁？"
# client_prompt = "三国演义中的桃园结义是几个人？都是谁？"
# client_prompt = "我想去美国留学,请给我一些建议."

messages = [
    ("system", "你是一个有用的AI聊天机器人."),
    ("human", client_prompt),
]
for chunk in llm.stream(messages):
    print(chunk.content, end="", flush=True)

print(evalEndTime(start_time))
```



下课时间

课件下载： https://github.com/komavideo/lesson_ollama