Ollama

大模型快速部署工具Ollama介绍

🌿 在linux上手动安装ollama

Mac和Windows具有成熟的一键安装工具，基本都能安装成功。这里提到如何在linux（内网，配置问题）手动安装linux。
当然你也可以尝试官方给出的一键脚本：`curl -fsSL https://ollama.com/install.sh | sh`

1. 启动autodl代理(非必需)

```
source /etc/network_turbo
```

2. 手动下载安装包

下载ollama linux:

```
# 手动从release 中下载amd64版本的安装文件
https://github.com/ollama/ollama/releases

# 或者使用命令行下载， 注意修改版本号（如v0.1.38)至最新版本
curl -fsSL https://github.com/ollama/ollama/releases/download/v0.1.38/ollama-linux-amd64 -o /root/autodl-tmp/apps/ollama-linux-amd64

或者:
wget -O /root/autodl-tmp/apps/ollama-linux-amd64 https://github.com/ollama/ollama/releases/download/v0.1.38/ollama-linux-amd64
```

3. 挂载软连接

```
sudo ln -s /root/autodl-tmp/apps/ollama-linux-amd64 /usr/local/bin/ollama
chmod +x /root/autodl-tmp/apps/ollama-linux-amd64
```

补充:

- `/bin`：基本命令，必须在系统引导时和单用户模式下可用。ls, rm
- `/usr/bin`：标准用户命令，通常由系统包管理器安装，依赖于 `/usr` 文件系统。 sed, grep
- `/usr/local/bin`：本地安装的软件和脚本，不依赖于系统包管理器。

4. 启动服务

启动服务：

```
ollama server
```

🧢 如何使用ollama

1. 使用对话模式

```
ollama run llama3    # 拉取模型并执行
ollama list          # 显示模型列表
ollama rm llama3     # 删除llama3模型
ollama pull          # 拉取模型，并不执行
```

实际可用的模型，比列出的library更多，推荐去huggingface看该模型的说明文件，确定是否已经被ollama支持。
其他的ollama命令，类似docker，比较简单，自己尝试一下。

2. 如何使用REST api格式调用

下面只展示最常用的api用法，更多参数使用建议阅读官方api文档：
https://github.com/ollama/ollama/blob/main/docs/api.md
注：非流式表示一次性返回所有结果，流式表示每次返回一个新生成的token。

流式调用

```
curl http://localhost:11434/api/generate -d '{
  "model": "llama3",
  "prompt": "Why is the sky blue?"
}'
```

非流式调用

```
curl http://localhost:11434/api/generate -d '{
  "model": "llama3",
  "prompt": "Why is the sky blue?",
  "stream": false
}'
```

带上下文调用（多轮对话）

```
curl http://localhost:11434/api/chat -d '{
  "model": "llama3",
  "messages": [
    {
      "role": "user",
      "content": "why is the sky blue?"
    },
    {
      "role": "assistant",
      "content": "due to rayleigh scattering."
    },
    {
      "role": "user",
      "content": "how is that different than mie scattering?"
    }
  ],
  "stream": false
}'
```

如何从加载本地的模型

ollama支持的模型格式为gguf，如果不是gguf格式需要使用llama.cpp将其转成gguf。
⚠注意：ollama并不能兼容所有的gguf文件！

1. 构建Modelfile
   除了必须制定文件路径，其他参数都是可选项。

```
FROM PATH_TO_YOUR__GGUF_MODEL

# set the temperature to 1 [higher is more creative, lower is more coherent]
PARAMETER temperature 1

# set the system message
SYSTEM """
You are Mario from Super Mario Bros. Answer as Mario, the assistant, only.
```

```
"""
```

2. 加载模型

```
ollama create NAME -f ./Modelfile

# NAME: 在ollama中显示的名称
# ./Modelfile: 绝对或者相对路径
```

3. 使用模型

```
ollama run NAME
```