Ollama e Modelli Locali

Cosa imparerai 💡

- Cos'è Ollama e perché è utile
- Installazione di Ollama su diversi sistemi operativi
- Organizzazione dei modelli
- Utilizzo di Ollama da terminale (CLI)
- Utilizzo di Ollama tramite API (Python e `requests`)

Ollama e Modelli Locali

Cos'è Ollama?

Ollama è un progetto OPEN-SOURCE che esegue modelli linguistici di grandi dimensioni (LLM) direttamente su una macchina locale.

Questo progetto nato nel 2023, ha lo scopo principale di mantenere il controllo sulla ***privacy dei dati***.

Difatto uno dei maggiori problemi nell'utilizzo di modelli LLM pubblici è quello della condivizione di informazioni sensibili che vengono utilizzate anche per riaddestrare i modelli nel processo di tuning.

Questo comportamento rende estremamente pericolo l'utilizzo di modelli pubblici per funzioni che richiedono l'analisi di dati sensibili.

Ollama e Modelli Locali

Cos'è Ollama?

- Ollama è uno strumento che ti permette di **eseguire modelli di linguaggio di grandi dimensioni (LLM) direttamente sul tuo computer**.
- **Vantaggi:**
  - **Privacy:** I dati non lasciano mai la tua macchina.
  - **Velocità:** Nessuna latenza di rete.
  - **Costo:** Nessun costo per l'utilizzo di API esterne.
  - **Offline:** Funziona anche senza connessione internet.

Ollama e Modelli Locali

Installazione di Ollama

- **macOS:**
  1. Scarica l'installer da https://ollama.com/.
  2. Apri l'installer e segui le istruzioni.
- **Linux:**
  1. Esegui lo script di installazione:

```
curl -fsSL https://ollama.com/install.sh | sh
```

- **Windows:**
  1. Segui le istruzioni sul sito di Ollama.

Ollama e Modelli Locali

Installazione di Ollama

Di seguito alcuni modelli di esempio:

Modello	Descrizione	Dimensione	Note
Llama2	Potente modello general-purpose di Meta	7B, 13B, 70B	Buona performance in diversi task, richiede GPU potente per 70B.
Mistral	Modello efficiente e performante	7B	Ottimo compromesso tra dimensioni e performance.
CodeLlama	Ottimizzato per la generazione di codice	7B, 13B, 34B	Ideale per sviluppatori.
StableLM	Modello sviluppato da Stability AI	3B, 7B	Alternativa open source.
...	...	...	...

Ollama e Modelli Locali

Organizzazione dei modelli

Ollama gestisce i modelli in modo strutturato per garantire riproducibilità e facilità d'uso.

- Un `Modelfile` è un file di testo che definisce **come costruire un modello**. (il DockerFile degli LLM 😉)
- Contiene istruzioni su:
  - **Modello di base (FROM):** Quale modello usare come punto di partenza (es. `llama2`).
  - **Personalizzazioni (SYSTEM, TEMPLATE):** Come adattare il modello alle tue esigenze.
  - **Dipendenze (INSTALL):** Eventuali pacchetti o librerie necessarie.

Ollama e Modelli Locali

Organizzazione dei modelli

Esempio:

```
FROM phi4:latest

SYSTEM "Sei un assistente AI che risponde sempre e solo in italiano. Anche se la domanda è in un'altra lingua. Sei esperto di allenamenti di corsa su medie / lunghe distanze. Fai sempre attenzione a dare consigli su come evitare di infortunarsi e conosci tutti i trucchi per migliorare le tecniche di corsa. Devi rispondere in italiano."
```

> ***NOTA:*** In questo esempio stiamo personalizzando il modello *phi4:latest* con un prompt per farlo diventare *coach* di corsa.

Ollama e Modelli Locali

Organizzazione dei modelli

Livelli (Layers)

Ollama utilizza un sistema di "livelli" (layers) per memorizzare i modelli e le loro modifiche in modo efficiente.

- **Ogni istruzione in un `Modelfile` crea un nuovo livello.**
- I livelli sono **immutabili e memorizzati nella cache.** Se due modelli condividono lo stesso livello di base, Ollama lo scaricherà solo una volta.

Questo permette di **risparmiare spazio su disco e velocizzare la creazione di nuovi modelli.**

> ***NOTA:*** I file dei modelli (pesi, configurazioni, ecc.) sono memorizzati in una directory nascoste specifiche di Ollama, ma è possibile gestirli tramite gli appositi comandi ollama.

Ollama e Modelli Locali

Utilizzo da Terminale (CLI)

Comando	Descrizione	Esempio
`ollama run`	Esegue un modello.	`ollama run llama2`
`ollama list`	Elenca i modelli installati.	`ollama list`
`ollama pull`	Scarica un modello.	`ollama pull mistralai/Mistral-7B`
`ollama rm`	Rimuove un modello.	`ollama rm llama2`
`ollama show`	Mostra informazioni su un modello (Modelfile).	`ollama show llama2`
`ollama serve`	Avvia il server Ollama (per l'API).	`ollama serve`
`ollama stop`	Ferma il server Ollama.	`ollama stop`

Ollama e Modelli Locali

Utilizzo tramite API (Python e `requests`)

```
import requests
import json
url = "http://localhost:11434/api/generate" #attenzione alla porta
headers = {'Content-Type': 'application/json'}
data = {
    "model": "llama2",
    "prompt": "Scrivi una breve poesia sull'autunno.",
    "stream": False # importante per ricevere la risposta completa in una
volta
}
response = requests.post(url, headers=headers, data=json.dumps(data),
stream=False)
if response.status_code == 200:
    print(response.json()['response'])
else:
    print(f"Errore: {response.status_code} - {response.text}")
```

Ollama e Modelli Locali

Utilizzo tramite API (Python e `requests`)

Spiegazione:

- **url:** corrisponda all'indirizzo dove Ollama sta girando.
- **`stream: False`:** Riceve l'intera risposta in una volta sola. Se impostato a `True`, la risposta viene creata come un flusso di "pezzi" (singoli token).
- **data:** Il dizionario JSON con il nome del modello e il prompt.
- **response:** Il testo generato si trova in `response.json()['response']`.

Ollama e Modelli Locali

QUIZ

1. Qual è uno dei principali vantaggi di usare Ollama?

- A) Richiede una connessione internet costante.
- B) I dati sono sempre inviati a server esterni.
- C) Permette di eseguire modelli di linguaggio offline e in privato.
- D) È più costoso delle API cloud.

2. Come si installa Ollama su Linux?

- A) Scaricando un file .exe
- B) Usando `apt-get install ollama`
- C) Eseguendo uno script di installazione con `curl`.
- D) Non è supportato su Linux.

Ollama e Modelli Locali

QUIZ

3. Quale comando si usa per eseguire un modello in Ollama?

- A) `ollama start`
- B) `ollama execute`
- C) `ollama run`
- D) `ollama open`

4. Cosa è un `Modelfile`?

- A) Un file eseguibile per installare Ollama.
- B) Un file di configurazione per definire come scaricare ed eseguire un modello.
- C) Un file contenente il modello stesso.
- D) Un file di log di Ollama.

Ollama e Modelli Locali

QUIZ

5. In quale chiave del JSON di risposta si trova il testo generato quando si usa l'API di Ollama?

- A) `text`
- B) `generated_text`
- C) `response`
- D) `output`

6. Cosa fa il parametro `stream: False` nell'API di Ollama?

- A) Invia la richiesta a un server di streaming.
- B) Disabilita l'audio durante la generazione.
- C) Riceve l'intera risposta in una volta sola invece di un flusso di dati.
- D) Comprime la risposta per una trasmissione più veloce.

Ollama e Modelli Locali

Approfondimenti

Vedere le seguenti guide di ollama presenti sul repository GITHUB ufficiale:

- Documentazione Ollama
- Documentazione modelfile

 Esercizi

1. **ese1**: Installa Ollama sul tuo sistema operativo.
2. **ese2**: Scarica ed esegui il modello `phi4:latest` da terminale.
3. **ese3**: Crea un `Modelfile` per far diventare il tuo assistente un bravo cuoco esperto in ricette vegetariane e vegane.
4. **ese4**: Scrivi uno script Python che usa l'API di Ollama per generare testo sfruttando i template