Ollama

Ollama est une plateforme open-source qui permet de créer, exécuter et partager des modèles de langage de grande taille localement sur des systèmes macOS et Linux1. Elle offre une interface en ligne de commande simple pour gérer ces modèles et inclut une bibliothèque de modèles pré-construits

- Installation
- Mise à jour
- Modèle personnalisé
- Prompt personnalisé
- Désinstallation
- Créer un modèle
- Supprimer un modèle

Installation

Source : GitHub - ollama/ollama: Get up and running with Llama 3.2, Mistral, Gemma 2, and other large language models.

> **Version :** 0.4.5

Mise à jour du système :

```
sudo apt update && sudo apt upgrade -y && sudo apt full-upgrade -y && sudo apt autoremove -y && sudo apt clean
```

Installer CURL :

```
sudo apt-get install curl
```

Récupération de la source :

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

Ajouter Ollama en service de démarrage :

Créez un utilisateur pour Ollama :

```
sudo useradd -r -s /bin/false -U -m -d /usr/share/ollama ollama
```

Ajoutez votre utilisateur actuel au groupe ollama :

```
sudo usermod -a -G ollama $(whoami)
```

Créez un fichier de service dans `/etc/systemd/system/ollama.service` **:**

```
[Unit]
Description=Ollama Service
After=network-online.target

[Service]
```

```
ExecStart=/usr/bin/ollama serve
User=ollama
Group=ollama
Restart=always
RestartSec=3
Environment="PATH=$PATH"

[Install]
WantedBy=default.target
```

> **Attention** ! Pour la ligne `ExecStart=/usr/bin/ollama serve`, vous devez correctement localiser l'installation de votre ollama, exemple `ExecStart=/usr/local/bin/ollama serve`.

Ensuite, démarrez le service :

```
sudo systemctl daemon-reload
sudo systemctl enable ollama
```

Voir les logs :

```
journalctl -e -u ollama
```

Mise à jour

Pour mettre à jour Ollama, vous pouvez exécuter à nouveau la commande suivante :

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

Cela réinstalle le script d'installation d'Ollama, et peut **écraser des fichiers existants** ou entraîner la perte de configurations personnalisées. Assurez-vous de **sauvegarder vos données et configurations importantes** avant de procéder à une mise à jour pour éviter toute perte de données.

Modèle personnalisé

Importation depuis GGUF

Ollama prend en charge l'importation de modèles au format **GGUF** via un fichier **Modelfile**.

Créer un fichier Modelfile :

Créez un fichier nommé `Modelfile`, qui doit contenir une instruction `FROM` avec le chemin local du modèle que vous souhaitez importer.

```
FROM ./vicuna-33b.Q4_0.gguf
```

Cela indique à Ollama de charger le modèle spécifié depuis le fichier local `vicuna-33b.Q4_0.gguf`.

Créer le modèle dans Ollama :

Une fois le fichier `Modelfile` créé, utilisez la commande suivante pour que le modèle soit intégré dans Ollama :

```
ollama create example -f Modelfile
```

Cette commande utilise le fichier `Modelfile` pour créer un modèle dans Ollama sous le nom "example".

Exécuter le modèle :

Enfin, vous pouvez exécuter le modèle avec la commande suivante :

```
ollama run example
```

Cela lance le modèle que vous venez de créer, ici appelé "example", pour qu'il commence à traiter des demandes ou des tâches.

Prompt personnalisé

Les modèles de la bibliothèque Ollama peuvent être personnalisés avec un **prompt**. Par exemple, pour personnaliser le modèle **llama3.2** :

Télécharger le modèle :

Tout d'abord, vous devez télécharger le modèle avec la commande suivante :

```
ollama pull llama3.2
```

Cela permet de récupérer le modèle **llama3.2** depuis la bibliothèque Ollama.

Créer un fichier Modelfile :

Ensuite, créez un fichier `Modelfile` contenant les instructions de personnalisation. Par exemple :

```
FROM llama3.2

# définir la température à 1 [plus élevé = plus créatif, plus bas = plus cohérent]
PARAMETER temperature 1

# définir le message système
SYSTEM """
You are Mario from Super Mario Bros. Answer as Mario, the assistant, only.
"""
```

- `FROM llama3.2` indique que vous souhaitez utiliser le modèle **llama3.2** comme base.
- La ligne `PARAMETER temperature 1` permet de régler la "température" du modèle, un paramètre qui influence la créativité du modèle. Une température élevée (comme 1) rend le modèle plus créatif, tandis qu'une température plus basse le rend plus cohérent.
- La section `SYSTEM` permet de personnaliser le message système, dans cet exemple, vous demandez au modèle de répondre en tant que **Mario**. uniquement.

Créer et exécuter le modèle :

Après avoir créé le fichier `Modelfile`, vous pouvez créer et exécuter le modèle personnalisé en utilisant ces commandes :

```
ollama create mario -f ./Modelfile
ollama run mario
```

La première commande crée un modèle nommé **Mario** à partir du fichier `Modelfile`, et la deuxième commande lance ce modèle.

Exemple d'interaction avec le modèle personnalisé :

```
>>> hi
Hello! It's your friend Mario.
```

Désinstallation

Supprimez le service Ollama :

```
sudo systemctl stop ollama
sudo systemctl disable ollama
sudo rm /etc/systemd/system/ollama.service
```

Supprimez le binaire `ollama` de votre répertoire `bin` (cela peut être `/usr/local/bin`, `/usr/bin`, ou `/bin`).

```
sudo rm $(which ollama)
```

Supprimez les modèles téléchargés ainsi que l'utilisateur et le groupe ollama créés pour le service :

```
sudo rm -r /usr/share/ollama
sudo userdel ollama
sudo groupdel ollama
```

Cela entraînera la **suppression définitive des fichiers d'Ollama**, y compris les modèles téléchargés et l'utilisateur associé. Assurez-vous de **sauvegarder vos données et configurations** avant de procéder à la désinstallation, car cette action est irréversible et pourrait supprimer des informations importantes.

Créer un modèle

La commande `ollama create` est utilisée pour créer un modèle à partir d'un fichier **Modelfile**.

```
ollama create mymodel -f ./Modelfile
```

Cette commande crée un modèle nommé **mymodel** en utilisant les instructions définies dans le fichier `Modelfile`.

Télécharger un modèle :

Pour télécharger un modèle, utilisez la commande suivante :

```
ollama pull llama3.2
```

Cette commande télécharge le modèle **llama3.2** depuis la bibliothèque Ollama. Elle peut également être utilisée pour **mettre à jour un modèle local**. Seules les différences (diff) entre le modèle local et la version la plus récente seront téléchargées.

Supprimer un modèle :

Si vous souhaitez supprimer un modèle, vous pouvez utiliser la commande suivante :

```
ollama rm llama3.2
```

Cela supprimera le modèle **llama3.2** de votre machine.

Copier un modèle :

Pour copier un modèle, utilisez la commande :

```
ollama cp llama3.2 my-model
```

Cela crée une copie du modèle **llama3.2** sous le nom **my-model**.

Entrée multilignes :

Pour entrer plusieurs lignes de texte, vous pouvez entourer le texte avec des guillemets triples (`"""`), comme suit :

```
>>> """Hello, ... world! ... """
```

Cela permet de saisir un texte réparti sur plusieurs lignes. Par exemple, cette entrée pourrait produire la sortie suivante :

```
I'm a basic program that prints the famous "Hello, world!" message to the console.
```

Modèles multimodaux :

Les modèles multimodaux permettent d'interagir avec des fichiers autres que du texte, comme des images. Par exemple, pour analyser une image avec un modèle multimodal, utilisez la commande :

```
ollama run llava "What's in this image? /Users/jmorgan/Desktop/smile.png"
```

Cela pourrait donner la réponse suivante :

```
The image features a yellow smiley face, which is likely the central focus of the picture.
```

Passer le prompt en argument :

Vous pouvez aussi passer un prompt directement en argument à la commande `ollama run`. Par exemple :

```
$ ollama run llama3.2 "Summarize this file: $(cat README.md)"
```

Cela permet de résumer le contenu d'un fichier, comme le fichier `README.md`. Le modèle peut générer un résumé tel que :

```
Ollama is a lightweight, extensible framework for building and running language models on the local machine. It provides a simple API for creating, running, and managing models, as well as a library of pre-built models that can be easily used in a variety of applications.
```

Afficher les informations sur un modèle :

Pour afficher les informations détaillées d'un modèle, utilisez la commande suivante :

```
ollama show llama3.2
```

Cela affiche des informations sur le modèle **llama3.2**, telles que sa version et ses paramètres.

Lister les modèles sur votre ordinateur :

Pour voir la liste de tous les modèles installés sur votre machine, utilisez cette commande :

```
ollama list
```

Lister les modèles actuellement chargés :

Pour voir quels modèles sont actuellement chargés en mémoire, utilisez la commande :

```
ollama ps
```

Arrêter un modèle en cours d'exécution :

Si vous souhaitez arrêter un modèle qui est en cours d'exécution, utilisez la commande suivante :

```
ollama stop llama3.2
```

Cela arrêtera le modèle **llama3.2** en cours d'exécution.

Démarrer Ollama :

Si vous voulez démarrer Ollama sans utiliser l'application de bureau, vous pouvez utiliser la commande suivante :

```
ollama serve
```

Cela démarre Ollama en mode serveur, permettant ainsi d'interagir avec les modèles via une API sans l'interface graphique.

Supprimer un modèle

Supprimer un modèle dans la list :

```
ollama rm llama3.2
```