Tutorial: LLMs locais com ollama

Por Eduardo Maçan

A ideia deste tutorial de instalação é ajudar a desmistificar os Large Language Models, que não tem nada do “gênio da lâmpada” superinteligente que mora em alguma nuvem mítica, que é como me parece que o cidadão comum os enxerga, impusionado por uma boa dose de click-baits, sensacionalismo e antropomorfização exagerada na mídia.

Enquanto é verdade que os requisitos de hardware para se treinar ou executar LLMs em escala estão muito longe do alcance do cidadão comum (e da esmagadora maioria das empresas), existem versões dos modelos “open source” que foram simplificados de maneira a ser possível carregá-los e executá-los dentro do espaço de memória dos computadores pessoais e que podem se beneficiar de GPUs domésticas.

O principal elemento de simplificação dos modelos é chamado de “quantização”. Os “pesos” calculados pelo treinamento tem sua precisão reduzida e deixam de ser representados por números em ponto flutuante de alta precisão para serem representados por apenas 4 bits, reduzindo drasticamente a necessidade de memória, mas também a qualidade dos textos que o modelo é capaz de produzir. As chamadas “alucinações” serão muito mais evidentes, bem como as demais limitações deste tipo de tecnologia, mas não será menos divertido e didático para iniciar sua jornada de aprofundamento no tema, além de ser muito mais barato.

Eu executei esses exemplos em dois computadores: um Laptop Intel i7-7500U com 16Gb de RAM, executando os modelos 100% em CPU+RAM e também em uma máquina mais moderna, com CPU Intel i9-13900H, 16Gb de RAM e uma GPU NVidia GeForce RTX 3050 6Gb. Também executei em Macs com chip M1 e M3 com 16Gb RAM. Nos equipamentos com hardware mais limitado a performance é inferior, mas ainda é bem aceitável para fins didáticos.

O fundamental é que os modelos caibam em memória.

Instalando o ollama

O ollama é uma aplicação open source capaz de executar localmente diferentes modelos, tudo o que você precisa saber está muito bem documentado no *site do projeto*[1], inclusive como instalar a ferramenta, executá-la e gerenciar seus modelos. Este pequeno guia pretende encurtar ainda mais esse caminho, de forma que não haja desculpas para não experimentar.

1 https://ollama.com/

Instalando o ollama no Microsoft Windows

De todos a instalação mais simples, no melhor estilo NNF (Next, Next, Finish). Baixe e execute o *instalador para windows*[2] , a sequência será como segue:

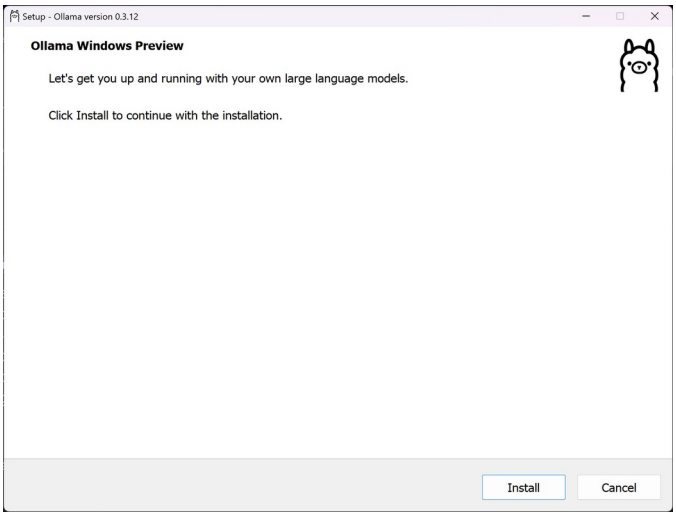


Figure 1: Tela inicial da Instalação do ollama para windows

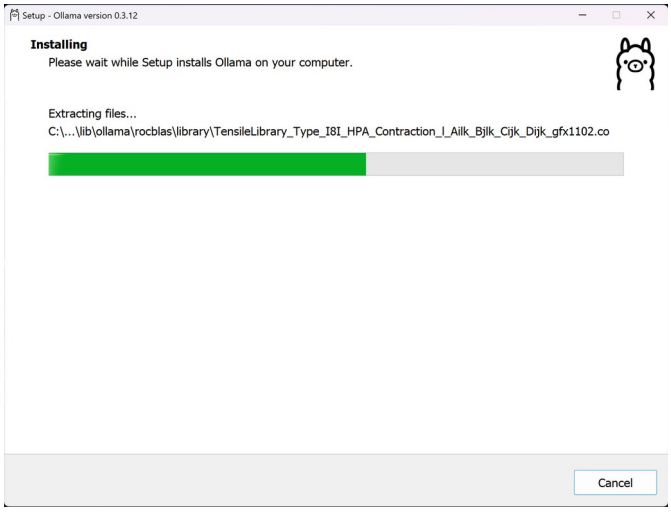


Figure 2: Instalação em Execução no Windows

2 https://ollama.com/download/windows

Ao final da instalação, o ollama irá abrir a linha de comando e imprimirá uma mensagem instruindo a instalação do seu primeiro modelo. O que é importante saber é que o ollama instala um serviço que irá iniciar junto com seu computador e que será responsável por gerenciar os modelos e executá-los. Você irá usar uma interface de linha de comando que irá interagir com o serviço, seja para gestão dos modelos, seja para abrir uma sessão de chat. Se você chegou até aqui você pode ir para a sessão “Baixando e executando seu primeiro modelo” do tutorial.

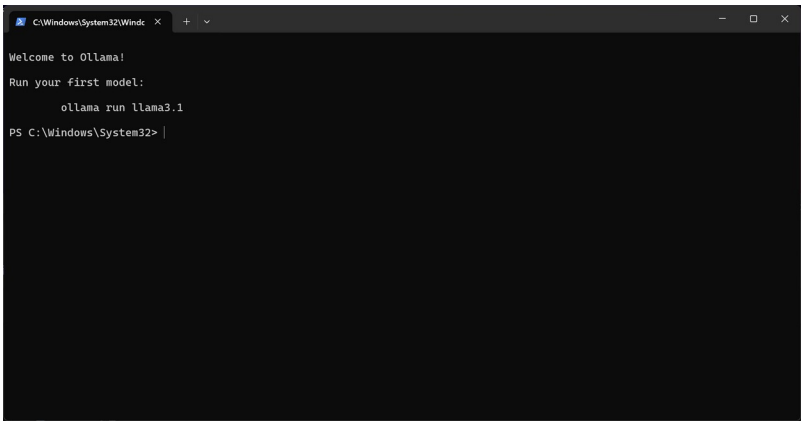


Figure 3: Instalação Windows Concluída, Prompt de Comando aberto.

Instalando o ollama no MacOS

Baixe o zip contendo a aplicação *para sua versão do Mac*[3] e abra seu conteúdo na pasta /Applications

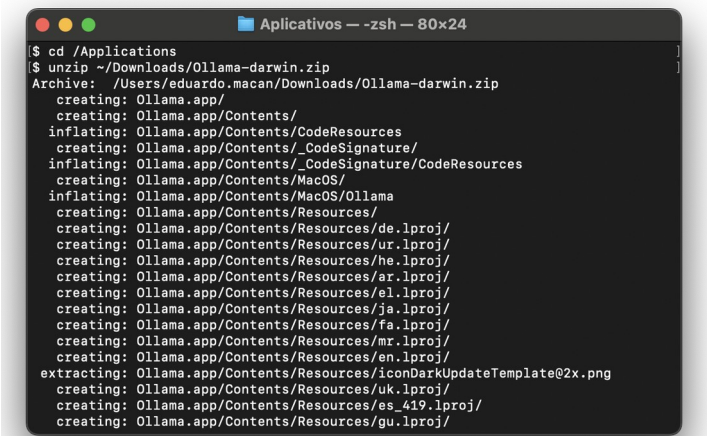


Figure 4: Descompactando o instalador/app no MacOS

3 https://ollama.com/download/mac

Agora execute o programa *ollama*, por exemplo usando o atalho **command+espaço** e digitando “ollama” na busca do spotlight. O mac vai pedir algumas confirmações de segurança e de privilégio de acesso para instalação. Em alguns momentos pedirá a senha do seu usuário, confirme e siga adiante.

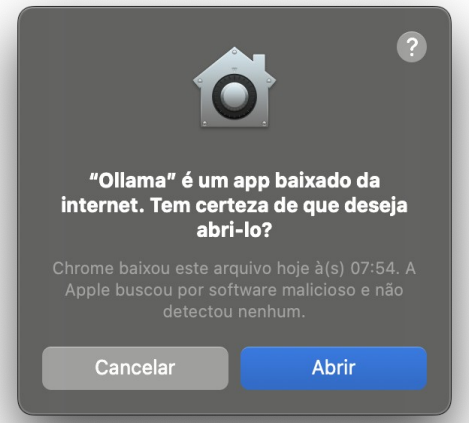


Figure 5: Confirmação de Segurança do Mac

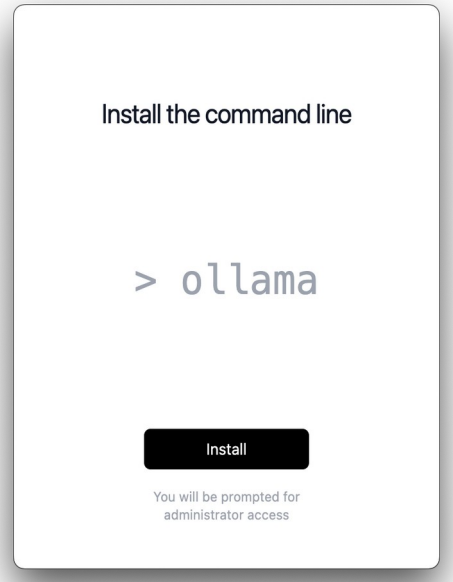


Figure 6: Programa de Instalação do Ollama para Mac

Figure 7: Senha para Instalação do ollama em /usr/local/bin

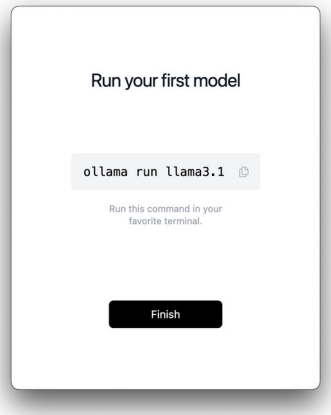
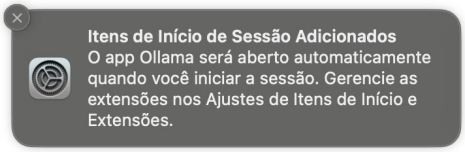


Figure 8: Instalação Concluída.

O ollama é uma aplicação em duas partes: O servidor/serviço que será iniciado junto com o seu computador e irá gerenciar os modelos e sua execução, e uma interface em linha de comando que será usada para interagir com o serviço. Por isso o aviso a seguir.



Instalando ollama no Linux

A maneira mais simples de se instalar o ollama no linux é *a descrita no site*[4], através da execução da linha de comando que baixa e executa o script de instalação, pedindo senha quando necessário para colocar os arquivos em /usr/local/bin e também para instalar o serviço.

O ollama é uma aplicação em duas partes: O servidor/serviço que será iniciado junto com o seu computador e irá gerenciar os modelos e sua execução, e uma interface em linha de comando que será usada para interagir com o serviço.

Os drivers nvidia são famosos por terem uma instalação não trivial no linux, eu não vou entrar no mérito dessa instalação. A única coisa a ficar atento é que, mesmo que você tenha uma placa nvidia, se os drivers não estiverem propriamente instalados você pode acabar não se beneficiando do hardware e acabar executando o modelo na CPU. Você vai gerar menos tokens por segundo, porém conseguirá brincar mesmo assim.

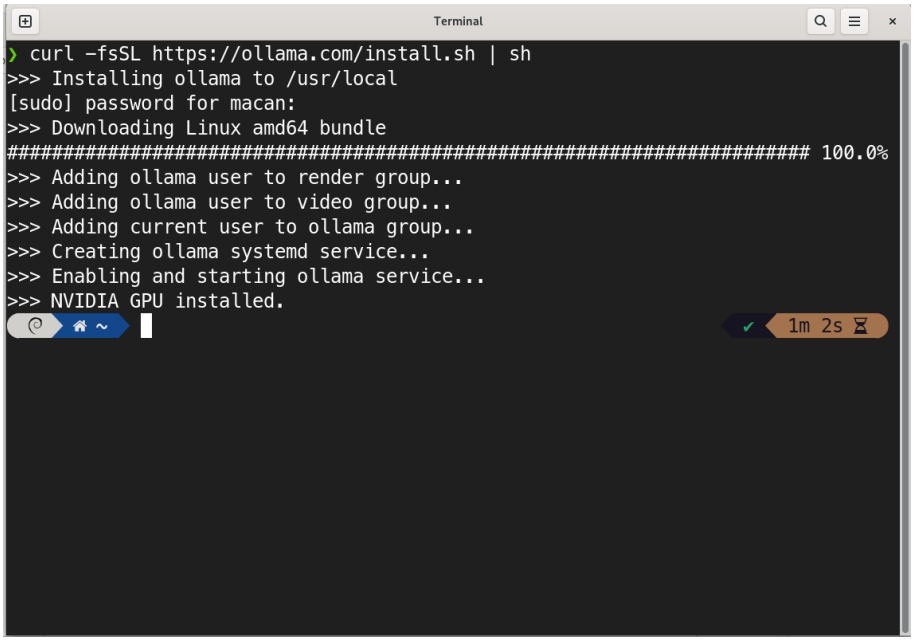


Figure 9: baixando e rodando o script de instalação no Linux

4 https://ollama.com/download/linux

Instalando ollama com homebrew (Linux ou MacOS)

Uma outra alternativa para Linux ou MacOS é instalar utilizando o *homebrew*[5]. Se você quiser optar por esta modalidade instale o homebrew em seu computador, se você já não o tem através das instruções do site e execute o seguinte comando:

```
brew install ollama
```

E você deve ver algo como a tela a seguir:

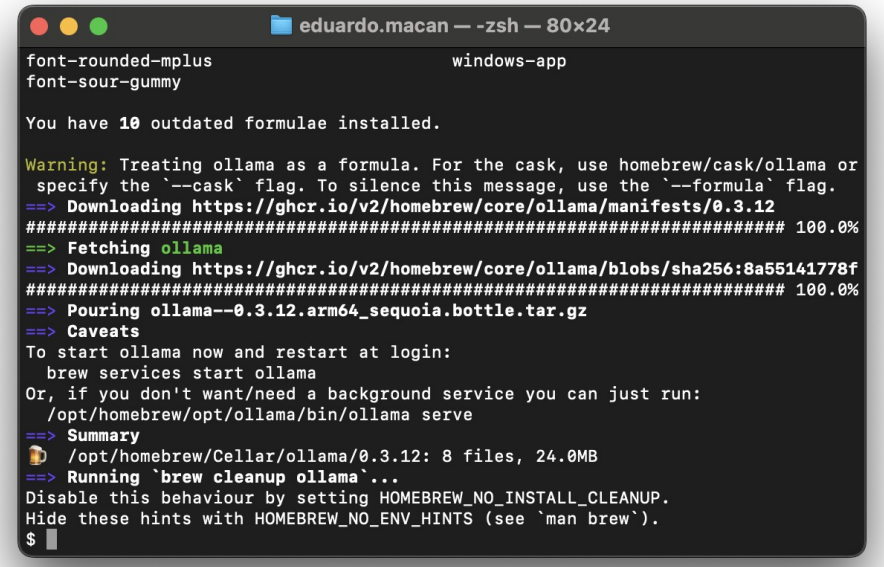


Figure 10: Instalação do ollama usando homebrew

O homebrew não irá por default instalar o serviço do ollama para iniciar automaticamente no boot. Se você quiser que o ollama se inicie sempre que seu computador ligar, você deve executar o comando abaixo:

```
brew services start ollama
```

Eu prefiro executar o serviço do ollama apenas quando vou interagir com ele, evitando assim mais um serviço rodando em backgroud o tempo todo. Neste caso, antes de instalar e interagir com modelos você terá que executar o comando "*ollama serve*" em outra janela antes de seguir para a próxima sessão deste documento. Também é interessante para acompanhar as interações que estão acontecendo com o serviço em tempo real.

```
ollama serve
```

5 https://brew.sh/

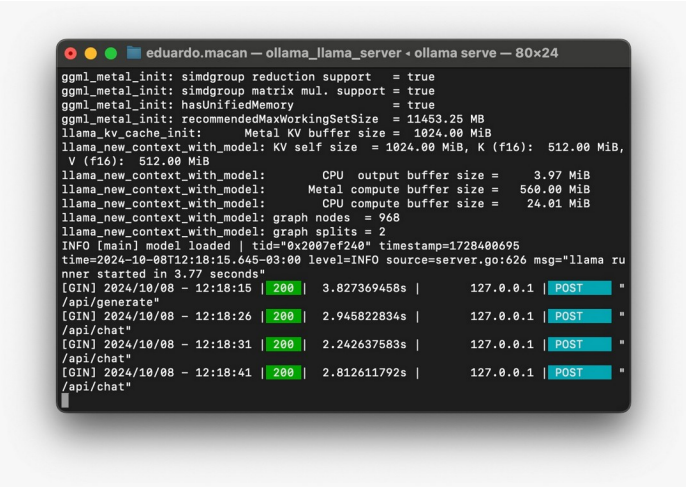


Figure 11: Executando o serviço manualmente com ollama serve

Executando seu primeiro Modelo

Neste momento você concluiu a instalação do serviço (ollama server) e da interface de linha de comando, mas ainda não tem nenhum modelo instalado. Os modelos são grandes (LARGE Language models, não é mesmo?) e o tempo de download vai variar da sua conexão com a internet e do modelo selecionado. Os menores modelos ocupam algumas centenas de Megabytes, tipicamente vários Gigabytes de armazenamento, e precisarão ser carregados em memória para serem executados. Não vou entrar em detalhes aqui sobre as razões por trás disso, mas imagine que um modelo contém para cada token (“palavra” por simplicidade) tanto informações que representam as probabilidades com que ela aparece próximo a cada uma das demais palavras do vocabulário quanto a uma representação das diversas nuances de significado que aquele token pode ter, esses valores “aprendidos” durante o treinamento são chamados de “parâmetros” e esta é uma das principais características dos modelos. Os principais termos que você deve se habituar a ouvir são: Tokens, Parâmetros, Contexto e Quantização. Todas estas informações estão bem detalhadas nas páginas de descrição dos diversos modelos suportados pelo ollama.

Para este exercício vamos utilizar o modelo “aya”, que me pareceu um dos melhores meios-termos entre tamanho, qualidade do que é gerado e suporte ao português, entre todos. No fim deixarei sugestões de outros modelos para você experimentar, comparar e tirar suas próprias conclusões.

Para instalar o aya, na linha de comando, digite a seguinte instrução:

```
ollama run aya
```

Você deve ver o download (~4.8Gb) do modelo durante a primeira execução e ao final o prompt de chat, para interação. A screenshot a seguir foi feita no Linux, mas a saída será semelhante nos outros sistemas operacionais.

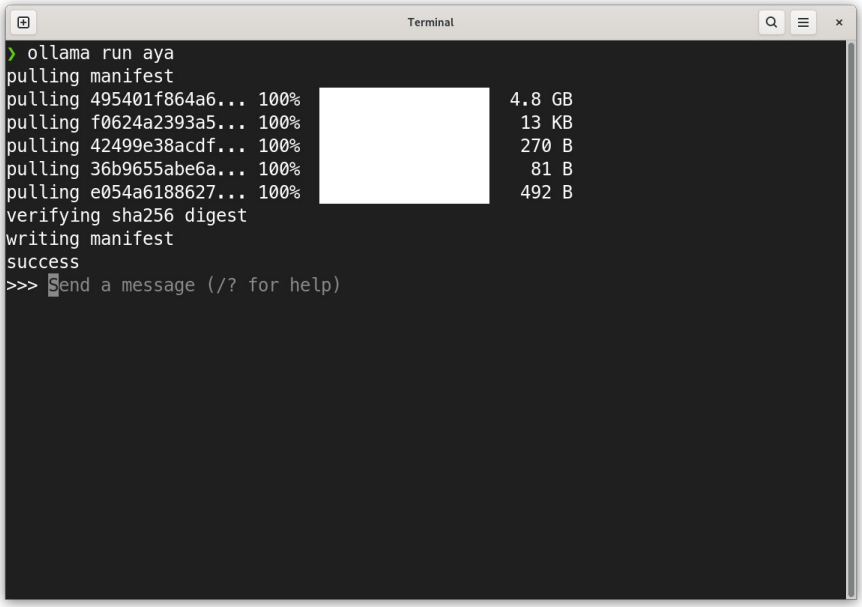


Figure 12: Interagindo com um modelo através do ollama

Tudo pronto! A partir de agora você já pode interagir com o modelo através desta interface de chat, experimente falar em português, mas também em inglês ou outras línguas. Aqui algumas saídas para alguns prompts simples:

```
>>> Tell me a joke.
Why did the chicken cross the road?
To get to the other side!

>>> Conte-me uma piada.
Por que o frango cruzou a estrada?
Para chegar ao outro lado!

>>> Por que o céu é azul?
O céu é azul devido à maneira como a luz solar interage com a atmosfera
terrestre. Quando a luz do sol passa pela atmosfera, as moléculas de ar e
partículas dispersam a luz. No entanto, as moléculas de ar dispersam todos
os comprimentos de onda da luz igualmente. A cor azul, tendo um
comprimento de onda menor, é dispersa em várias direções, dando ao céu a
sua cor característica. Este fenómeno é conhecido como dispersão de
Rayleigh, nomeado em homenagem a Lord Rayleigh, que o descreveu pela
primeira vez. Então, na próxima vez que você olhar para o céu azul,
lembre-se do poder da dispersão de Rayleigh!
```

Divertido, não? O que antes era a mágica de “um gênio da lâmpada na nuvem” agora está acontecendo no seu próprio computador ao lado de aplicativos e ferramentas bem mais mundanas, como editores de texto e planilhas. Sem dúvidas que esta é uma aplicação ordens de grandeza mais complexa em sua construção e funcionamento do que as demais, mas não deixa de ser uma ferramenta como as outras. Não é mesmo?

Informação detalhada sobre os modelos pode ser encontrada no próprio site do ollama, aqui estão as informações sobre o modelo aya[6], por exemplo.

Agora explore outros modelos! Seguem algumas sugestões e o tamanho de download de cada um:

Nome	Tamanho	URL
Phi3:medium-128k	7.9 Gb	https://ollama.com/library/phi3
Gemma:7b	5.2 Gb	https://ollama.com/library/gemma
Aya	4.8 Gb	https://ollama.com/library/aya
Llama:3.1	4.7 Gb	https://ollama.com/library/llama3
phi3	2.3 Gb	https://ollama.com/library/phi3

6 https://ollama.com/library/aya

Conclusão

Espero que esse mini tutorial tenha ajudado você a superar o momento mais complicado de qualquer jornada, que sempre é começar e extrair um primeiro resultado frente a um assunto inicialmente complexo. Ter um entendimento amplo, mesmo que superficial de como essas coisas funcionam por dentro não é tão difícil assim e dará um senso muito mais concreto de como essas ferramentas podem ser úteis para você e quais as limitações da tecnologia.

Se você achou esse material útil, repasse para um amigo que está na mesma situação. Me marque no linkedin ou em outra rede qualquer e me deixe saber que foi útil pra você. Eu costumo ser @eduardomacan em todos os lugares online :)

Obrigado,

Maçan

Revisões e Errata

2024-10-08 Versão inicial.