Documentation & Installation Guide

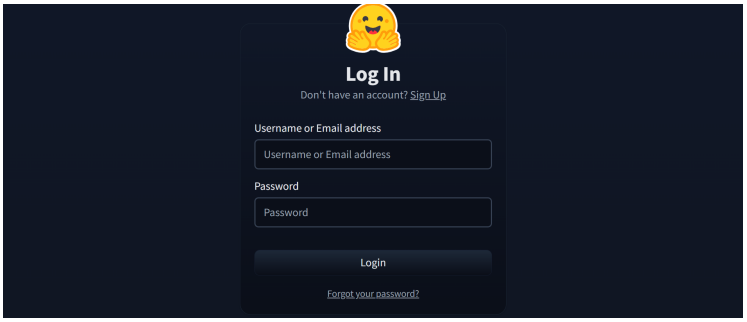
Mistral 7b Model

1. Installed Required Libraries

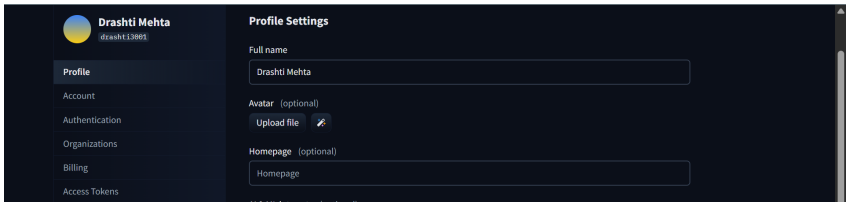
```
!pip install --upgrade pip
!pip install -q kagglehub huggingface_hub transformers accelerate torch
```

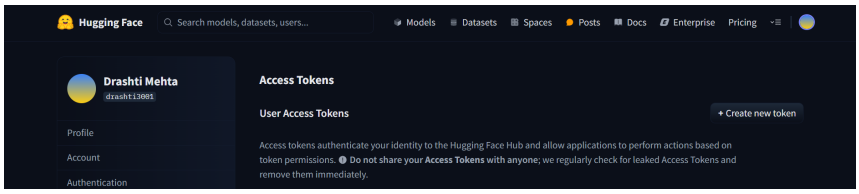
```
Requirement already satisfied: pip in /usr/local/lib/python3.11/dist-packages (24.1.2)
Collecting pip
  Downloading pip-25.0.1-py3-none-any.whl.metadata (3.7 kB)
Downloading pip-25.0.1-py3-none-any.whl (1.8 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.8/1.8 MB 26.7 MB/s eta 0:00:00
Installing collected packages: pip
  Attempting uninstall: pip
    Found existing installation: pip 24.1.2
    Uninstalling pip-24.1.2:
      Successfully uninstalled pip-24.1.2
Successfully installed pip-25.0.1
```

2. Hugging Face Token Setup
    1. Login / Create the Hugging Face Account - https://huggingface.co/login



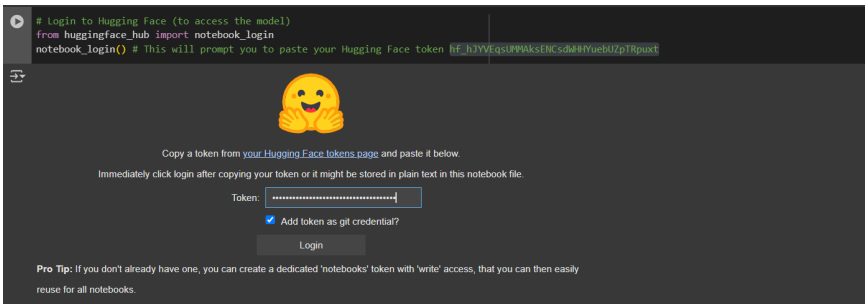
    2. Go to your account's settings -> Access token's : https://huggingface.co/settings/tokens



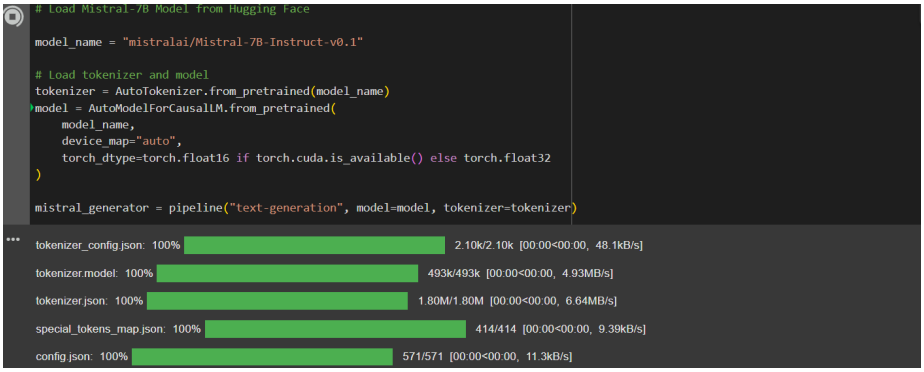


    3. Click on **"New token"** in the top right. In the popup:
        a. **Name the token** something like Mistral Model Access.
        b. Under **Role**, choose Read (this is enough to load public models).
        c. Click **Generate token**.

4. Once generated, **copy the token** and paste it when prompted in the code



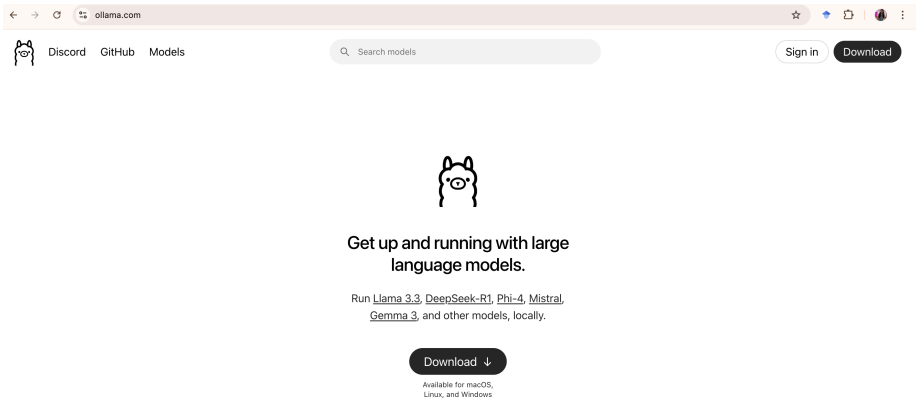
5. Load the model



Llama 2-7B model

1. Download Ollama and install it. Ollama allows us to run the LLMs locally with a minimal setup.

   Link to download ollama : https://ollama.com/



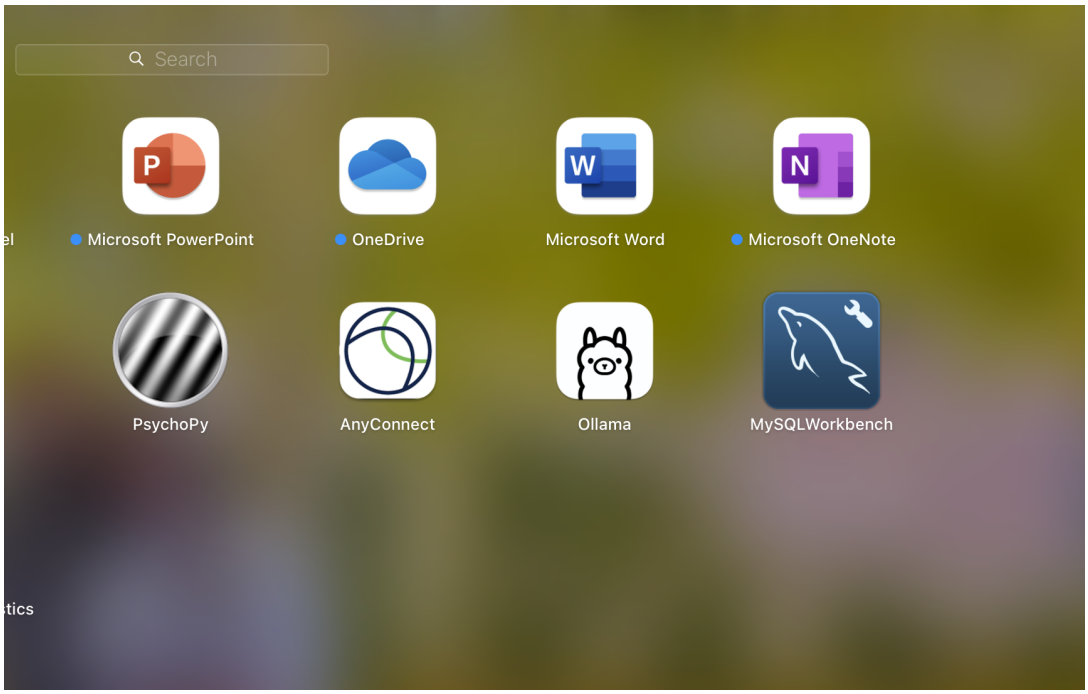
2. Next open the terminal and give the following commands to pull the model you want to work with and use them locally.

   Start the ollama server: ollama serve

   Pull the model: ollama pull llama2:7b-chat

3. Every time you want to run the python code with ollama and specific model,

   open the ollama application it will start running and we can execute the code.



Once all the work is done and you close the jupyter notebook, quit the ollama in the tool bar.