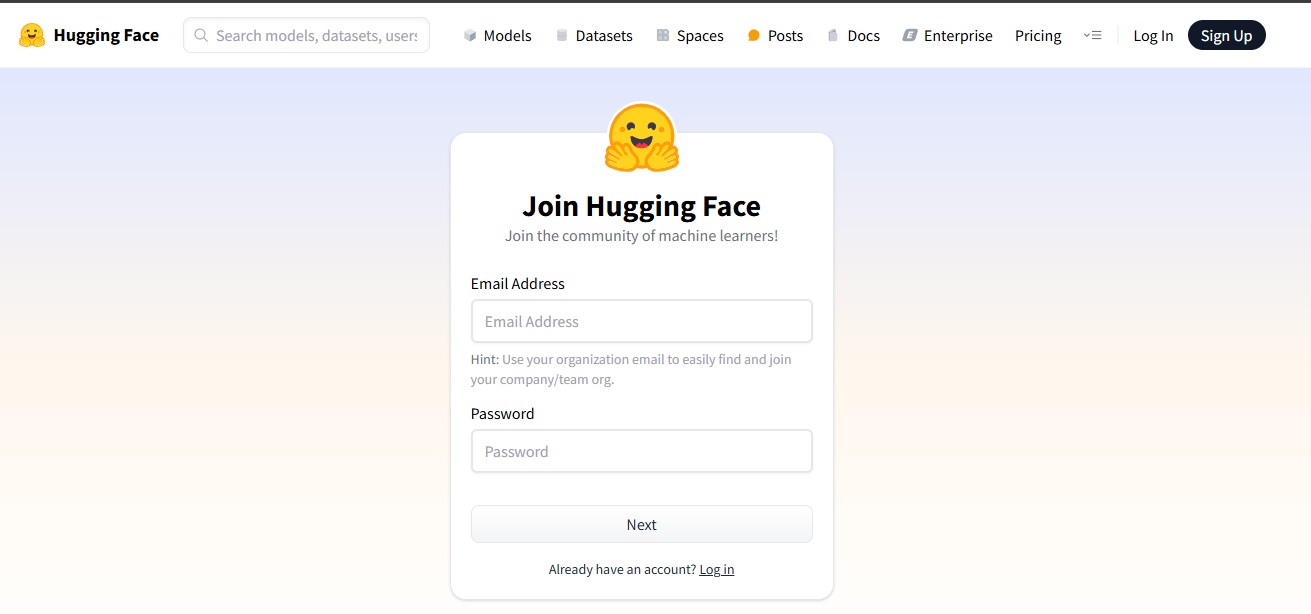
**Executing LLM model using LangChain and hugging face**

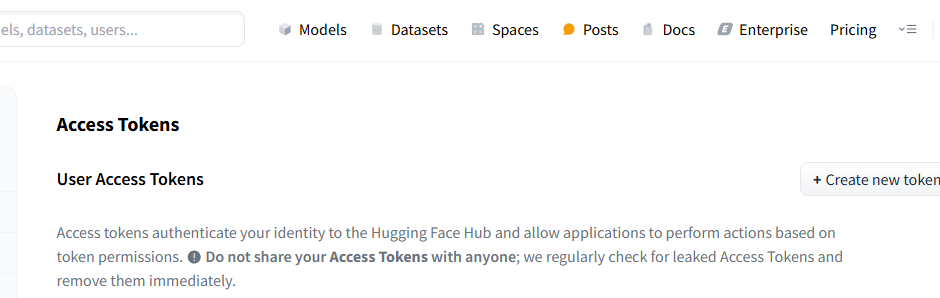
With all the hype of Generative AI, I want to try how to use the LLM model for a simple application. I built a program to read pdf files using models available on the hugging face site (https://huggingface.co/) . There are multiple articles and videos available for this work and I am not going to do something new, but I am writing this mostly because of problems I faced during the setup. Maybe this will help someone who is also trying a similar setup.

I used the Nvidia L4 system to run this code.

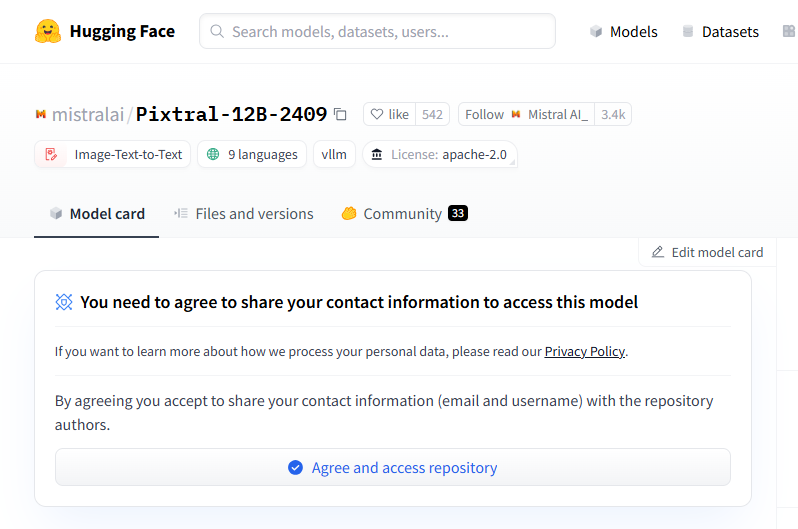
1. Create an account in hugging face site:



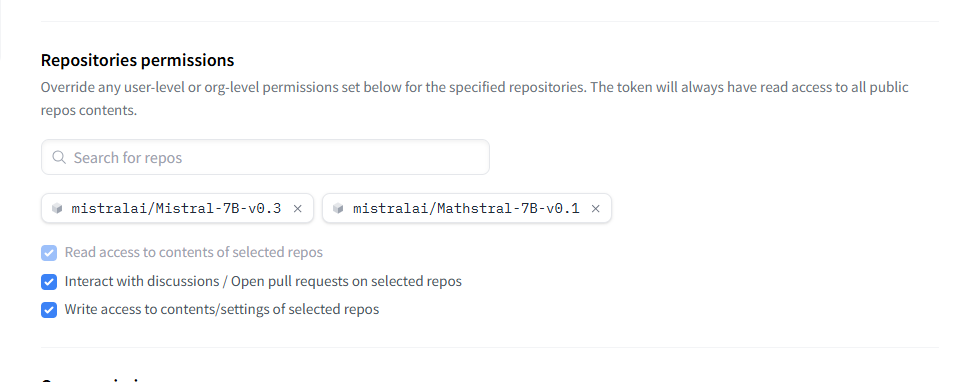
1. Create the API token.



1. search the model you wanted to use, in my case, I used **mistralai/Mistral-7B-v0.3**. select the agreement for the model.



Go to the token→edit permission and provide access.



1. Install python >=3.9 if you don’t have one already.
2. Install hugging face hub and cli using pip - pip install -U "huggingface\_hub[cli]"
3. Install below packages:

Pip install transformers

Pip install protobuf

Pip install sentencepiece

pip install langchain-community

1. Now the problem I faced was to download the model and execute. I used the code which is available at model page in huggingface site:

*from huggingface\_hub import snapshot\_download*

*from pathlib import Path*

*mistral\_models\_path = Path.home().joinpath('mistral\_models', '7B-v0.3')*

*mistral\_models\_path.mkdir(parents=True, exist\_ok=True)*

*snapshot\_download(repo\_id="mistralai/Mistral-7B-v0.3", allow\_patterns=["params.json", "consolidated.safetensors", "tokenizer.model.v3"], local\_dir=mistral\_models\_path)*

First you need to login to hugginface cli then only this download works and later the problem will occur. Once you try to use model (with the correct path where you installed the model) the program will always throw **Can't load tokenizer If you were trying to load it from ‘Models - Hugging Face’, make sure you don’t have a local directory with the same name.**

how to login to huggingface cli? In one of the setup, I was able to without any issues with command huggingface-cli login. In another setup (on mac), I had to run the command echo "export PATH=\"`python3 -m site --user-base`/bin:\$PATH\"" >> ~/.bashrc

source ~/.bashrc

and then called huggingface-cli login.This will ask huggingface api token.

1. On the same terminal, where you logged in to huggingface-cli, download the model and tokenizer. Use the code I have attached in git LLMDownloader.py, change the path where you want to save the model and tokenizer. Run using

Python LLMDownloader.py

The path will be used in another file to use the downloaded model. Notice we can change the default path for hugging face home, which is under ./cache for model, dataset etc.

1. Now use the code PDFQA.py to get answers from pdf. Update the code for the location of pdf files as well as for the model and tokenizer file. You can notice how the tokenizer and model are used in this file. I am not going into details of each method/code as there is already plenty of material available online.

Hope you will be able to execute a LLM model for your project.