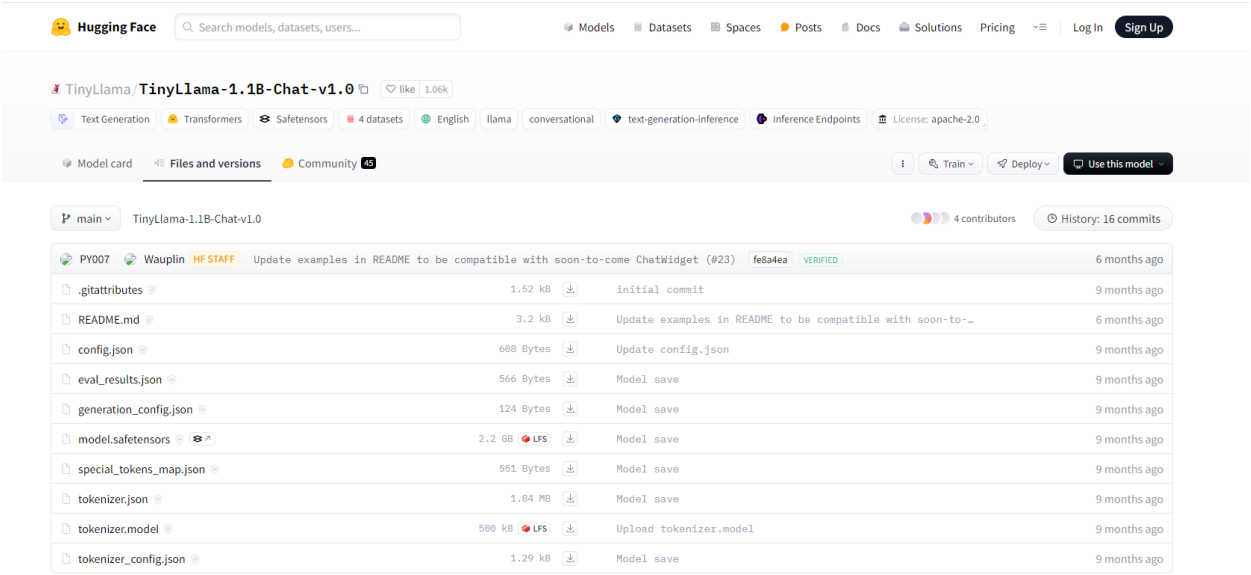
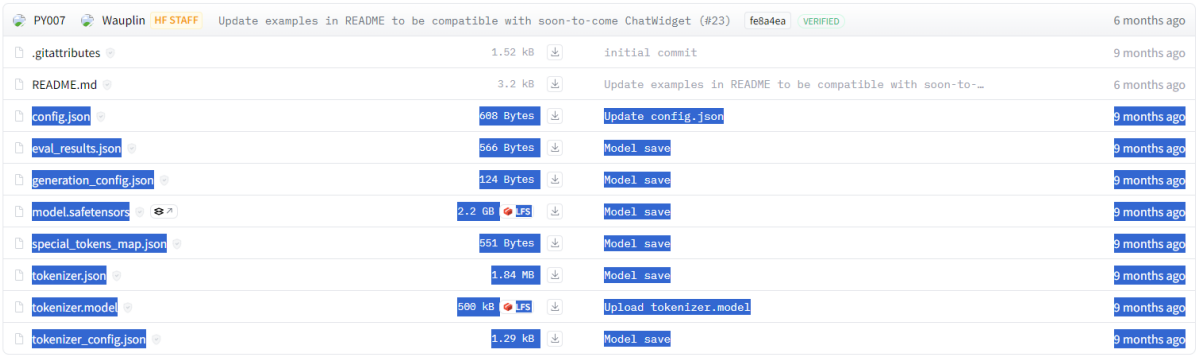
Instructions to running local LLMs

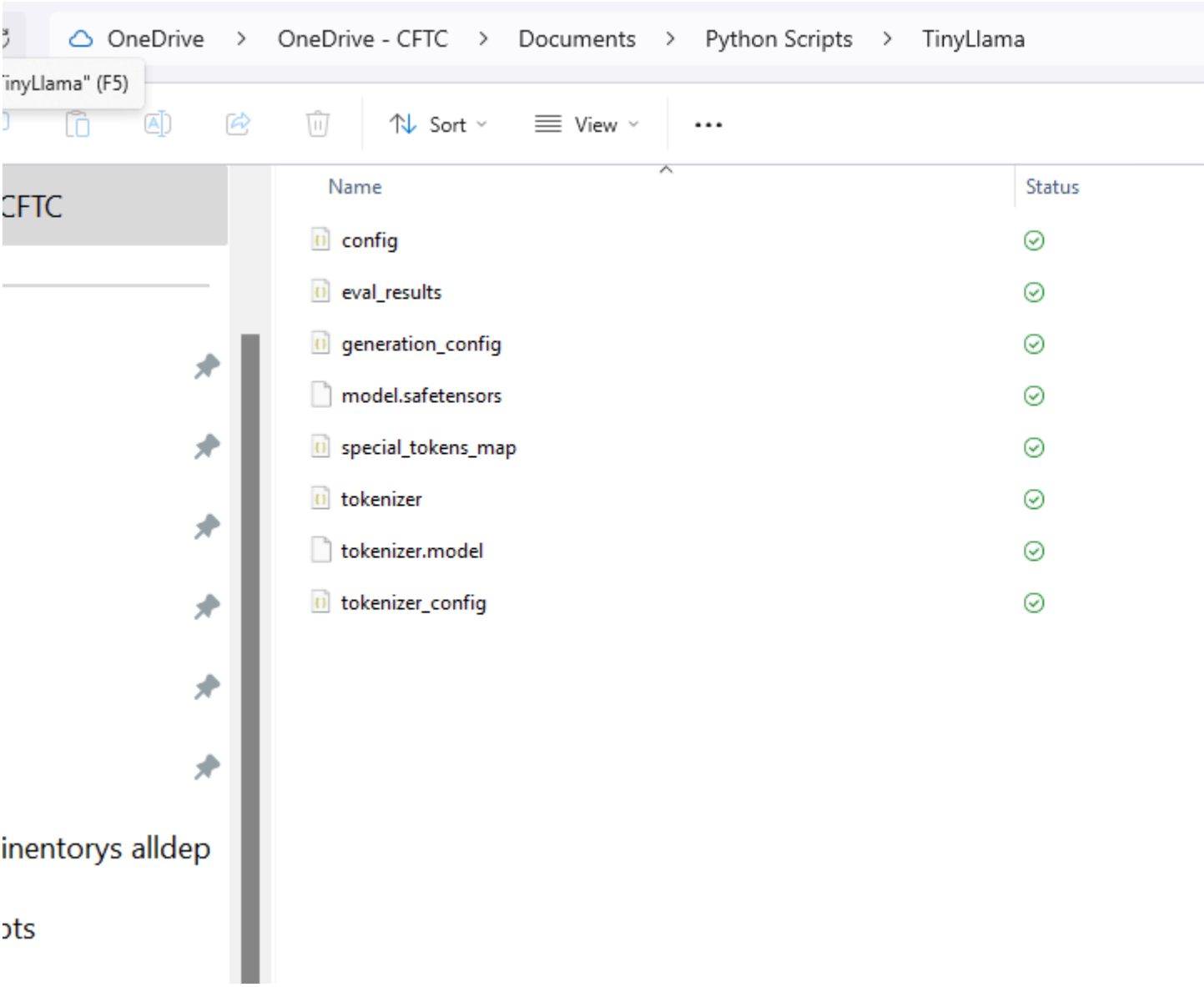
Go to https://huggingface.co/ and find your model.



Download all necessary files.



Choose a directory



Install packages

```
pip install torch transformers
```

May need to use conda-forge method bellow

```
conda install pytorch=1.9.0 torchvision torchaudio -c conda-forge
```

Load Packages

```python
# Packages
import torch
from transformers import AutoTokenizer, AutoModelForCausalLM
```

Set the path

```python
# Set the path to the directory containing the model files
model_path = "D:\Data\OneDrive\Ccantu\OneDrive - CFTC\Documents\Python Scripts\TinyLlama"
```

Load the model and tokenizer

```python
# Load tokenizer and model
tokenizer = AutoTokenizer.from_pretrained(model_path)
model = AutoModelForCausalLM.from_pretrained(model_path, torch_dtype=torch.float16, device_map="auto")
```

Example generation function

```python
def generate_response(prompt, max_length=100): # max_length keeps response short
    inputs = tokenizer(prompt, return_tensors="pt").to(model.device)
    outputs = model.generate(**inputs, max_length=max_length, num_return_sequences=1) # num_return_sequences gives only 1 output sequence
    return tokenizer.decode(outputs[0], skip_special_tokens=True) # Turns output into a readable text and skips padding to a length
```

Example prompt

```python
# Example prompt case
prompt = "Write a short paragraph about commodities:"
response = generate_response(prompt)
print(f"Prompt: {prompt}\n")
print(f"Generated Response:\n{response}")
```