Importing Open Source Models to Ollama

Clone from huggingface.co:

```
apt install git-lfs
```

```
git clone https://huggingface.co/nvidia/Llama-3.1-Nemotron-70B-Instruct-HF
```

Import to ollama

Create Modelfile

```
# Modelfile
FROM "./"
PARAMETER stop "<|im_start|>"
PARAMETER stop "<|im_end|>"
TEMPLATE """
<|im_start|>system
{{ .System }}<|im_end|>
<|im_start|>user
{{ .Prompt }}<|im_end|>
<|im_start|>assistant
"""
```

```
ollama create Llama-3.1-Nemotron-70B-Instruct-HF
```

To also quantize the model:

```
ollama create Llama-3.1-Nemotron-70B-Instruct-HF:q4_0 --quantize q4_0
```

See more Optiones here: https://github.com/ollama/ollama/blob/main/docs/import.md#supported-quantizations

---

Revision #4
Created 2024-10-16 17:44:37 UTC by joscha.mijailovic
Updated 2024-11-05 00:57:35 UTC by joscha.mijailovic