

!nvidia-smi

Fri Jul 19 12:14:20 2024

NVIDIA-SMI 535.104.05				Driver Version: 535.104.05		CUDA Version: 12.2	
GPU	Name	Persistence-M	Bus-Id	Disp.A	Volatile Uncorr. ECC		
Fan	Temp	Perf	Pwr:Usage/Cap	Memory-Usage	GPU-Util	Compute M.	MIG M.
0	Tesla T4	Off	00000000:00:04:0	Off	0	0	
N/A	47C	P8	9W / 70W	0MiB / 15360MiB	0%	Default	N/A

Processes:							GPU Memory
GPU	GI	CI	PID	Type	Process name		Usage
ID	ID	ID					
No running processes found							

!pip install -r requirement.txt

```

Requirement already satisfied: accelerate==0.29.3 in /usr/local/lib/python3.10/dist-packages (from -r requirement.txt (line 1)) (0.29.3)
Requirement already satisfied: bitsandbytes==0.43.1 in /usr/local/lib/python3.10/dist-packages (from -r requirement.txt (line 2)) (0.43.1)
Requirement already satisfied: transformers==4.40.0 in /usr/local/lib/python3.10/dist-packages (from -r requirement.txt (line 3)) (4.40.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-packages (from accelerate==0.29.3->-r requirement.txt (line 1)) (1.26.4)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from accelerate==0.29.3->-r requirement.txt (line 2)) (24.1)
Requirement already satisfied: psutil in /usr/local/lib/python3.10/dist-packages (from accelerate==0.29.3->-r requirement.txt (line 3)) (5.9.8)
Requirement already satisfied: pyyaml in /usr/local/lib/python3.10/dist-packages (from accelerate==0.29.3->-r requirement.txt (line 4)) (6.0.1)
Requirement already satisfied: torch>=1.10.0 in /usr/local/lib/python3.10/dist-packages (from accelerate==0.29.3->-r requirement.txt (line 5)) (2.2.2)
Requirement already satisfied: huggingface-hub in /usr/local/lib/python3.10/dist-packages (from accelerate==0.29.3->-r requirement.txt (line 6)) (0.23.4)
Requirement already satisfied: safetensors>=0.3.1 in /usr/local/lib/python3.10/dist-packages (from accelerate==0.29.3->-r requirement.txt (line 7)) (0.4.3)
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-packages (from transformers==4.40.0->-r requirement.txt (line 3)) (3.13.1)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.10/dist-packages (from transformers==4.40.0->-r requirement.txt (line 4)) (2024.5.15)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from transformers==4.40.0->-r requirement.txt (line 5)) (2.32.0)
Requirement already satisfied: tokenizers<0.20,>=0.19 in /usr/local/lib/python3.10/dist-packages (from transformers==4.40.0->-r requirement.txt (line 6)) (0.19.1)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.10/dist-packages (from transformers==4.40.0->-r requirement.txt (line 7)) (4.66.1)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.10/dist-packages (from huggingface-hub->accelerate==0.29.3->-r requirement.txt (line 5)) (2024.6.1)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.10/dist-packages (from huggingface-hub->accelerate==0.29.3->-r requirement.txt (line 6)) (4.11.0)
Requirement already satisfied: sympy in /usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.29.3->-r requirement.txt (line 5)) (1.12.0)
Requirement already satisfied: networkx in /usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.29.3->-r requirement.txt (line 6)) (3.3)
Requirement already satisfied: Jinja2 in /usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.29.3->-r requirement.txt (line 7)) (3.1.3)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.1.105 in /usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.29.3->-r requirement.txt (line 8)) (12.1.105)
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.1.105 in /usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.29.3->-r requirement.txt (line 9)) (12.1.105)
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.1.105 in /usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.29.3->-r requirement.txt (line 10)) (12.1.105)
Requirement already satisfied: nvidia-cudnn-cu12==8.9.2.26 in /usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.29.3->-r requirement.txt (line 11)) (8.9.2.26)
Requirement already satisfied: nvidia-cublas-cu12==12.1.3.1 in /usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.29.3->-r requirement.txt (line 12)) (12.1.3.1)
Requirement already satisfied: nvidia-cufft-cu12==11.0.2.54 in /usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.29.3->-r requirement.txt (line 13)) (11.0.2.54)
Requirement already satisfied: nvidia-curand-cu12==10.3.2.106 in /usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.29.3->-r requirement.txt (line 14)) (10.3.2.106)
Requirement already satisfied: nvidia-cusolver-cu12==11.4.5.107 in /usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.29.3->-r requirement.txt (line 15)) (11.4.5.107)
Requirement already satisfied: nvidia-cusparselt-cu12==12.1.0.106 in /usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.29.3->-r requirement.txt (line 16)) (12.1.0.106)
Requirement already satisfied: nvidia-nccl-cu12==2.20.5 in /usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.29.3->-r requirement.txt (line 17)) (2.20.5)
Requirement already satisfied: nvidia-nvtx-cu12==12.1.105 in /usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.29.3->-r requirement.txt (line 18)) (12.1.105)
Requirement already satisfied: triton==2.3.1 in /usr/local/lib/python3.10/dist-packages (from torch>=1.10.0->accelerate==0.29.3->-r requirement.txt (line 19)) (2.3.1)
Requirement already satisfied: nvidia-nvjitlink-cu12 in /usr/local/lib/python3.10/dist-packages (from nvidia-cusolver-cu12==11.4.5.107->torch>=1.10.0->accelerate==0.29.3->-r requirement.txt (line 15)) (2.3.0)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests->transformers==4.40.0->-r requirement.txt (line 5)) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->transformers==4.40.0->-r requirement.txt (line 6)) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests->transformers==4.40.0->-r requirement.txt (line 7)) (2.2.3)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests->transformers==4.40.0->-r requirement.txt (line 8)) (2024.7.4)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/dist-packages (from Jinja2->torch>=1.10.0->accelerate==0.29.3->-r requirement.txt (line 7)) (2.1.5)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.10/dist-packages (from sympy->torch>=1.10.0->accelerate==0.29.3->-r requirement.txt (line 15)) (1.3.0)

```

```

import json
import torch
from transformers import (AutoTokenizer, AutoModelForCausalLM,BitsAndBytesConfig,pipeline)

```

```

config_data = json.load(open('config.json'))
HF_TOKEN = config_data['HF_TOKEN']

```

```

model_name="meta-llama/Meta-Llama-3-8B"

```

```
bnb_config = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_use_double_quant=True,
    bnb_4bit_quant_type="nf4",
    bnb_4bit_compute_dtype=torch.bfloat16
)
```

```
tokenizer = AutoTokenizer.from_pretrained(model_name, token=HF_TOKEN)
tokenizer.pad_token = tokenizer.eos_token
```

⚡ /usr/local/lib/python3.10/dist-packages/huggingface_hub/file_download.py:1132: FutureWarning: warnings.warn(
 /usr/local/lib/python3.10/dist-packages/huggingface_hub/utils/_token.py:89: UserWarning: The secret `HF_TOKEN` does not exist in your Colab secrets.
 To authenticate with the Hugging Face Hub, create a token in your settings tab (<https://huggingface.co/settings/tokens>). You will be able to reuse this secret in all of your notebooks.
 Please note that authentication is recommended but still optional to access public model

special_tokens_map.json: 100% 73.0/73.0 [00:00<00:00, 1.29kB/s]
 Special tokens have been added in the vocabulary, make sure the associated word embeddir

```
model=AutoModelForCausalLM.from_pretrained(
    model_name,
    token=HF_TOKEN,
    quantization_config=bnb_config,
    device_map="auto"
)
```

⚡ config.json: 100% 654/654 [00:00<00:00, 14.5kB/s]
 model.safetensors.index.json: 100% 23.9k/23.9k [00:00<00:00, 392kB/s]
 Downloading shards: 100% 4/4 [02:50<00:00, 38.27s/it]
 model-00001-of-00004.safetensors: 100% 4.98G/4.98G [00:50<00:00, 20.8MB/s]
 model-00002-of-00004.safetensors: 100% 5.00G/5.00G [00:46<00:00, 268MB/s]
 model-00003-of-00004.safetensors: 100% 4.92G/4.92G [00:54<00:00, 51.8MB/s]
 model-00004-of-00004.safetensors: 100% 1.17G/1.17G [00:18<00:00, 67.8MB/s]

```
text_generator = pipeline(
    'text-generation',
    model=model,
    tokenizer=tokenizer,
    max_new_tokens=128,
)
```

```
def get_response(prompt):
    sequences = text_generator(prompt)
    gen_text = sequences[0]['generated_text']
    return gen_text
```

```
prompt = " what is Llama3 "
```

```
llama3_response = get_response(prompt)
```

```
llama3_response
```

⚡ ' what is Llama3 2.0? It is an extension to Llama2.0 which provides additional features. The Llama3.0 is a new and improved version of Llama2.0, which has been developed with the help of the Llama2.0 developers. Llama3.0 is the first version of Llama2.0 which is a new and improved version of Llama2.0. The Llama3.0 is the first version of Llama2.0 which is a new and improved version of Llama2.0. The Llama3.0 is the first version of Llama2.0.'

