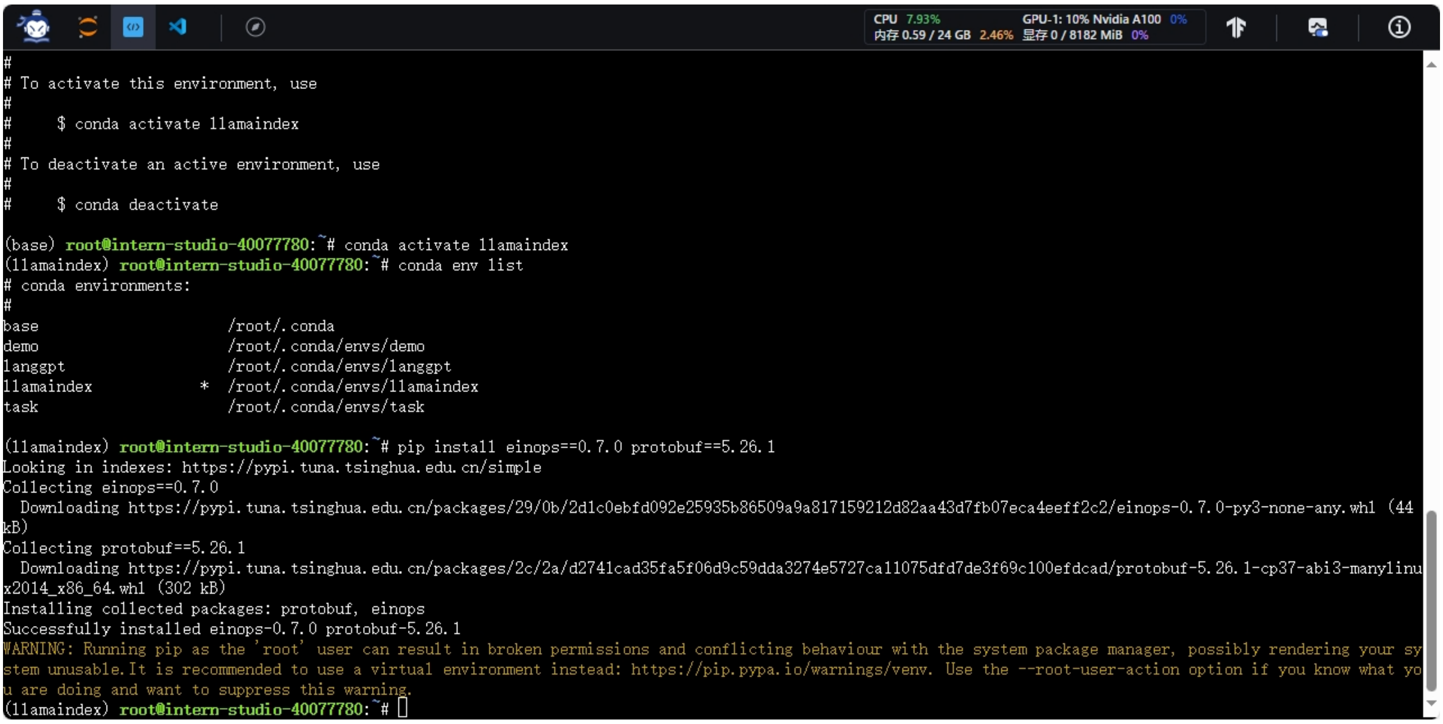
基础岛任务笔记4-InternLM + LlamaIndex RAG 实践

- **任务要求**：基于 LlamaIndex 构建自己的 RAG 知识库，寻找一个问题 A,这个问题在使用 LlamaIndex 之前InternLM2-Chat-1.8B模型不会回答，借助 LlamaIndex 后 InternLM2-Chat-1.8B 模型具备回答 A 的能力，截图保存。

环境，模型准备

配置基础环境

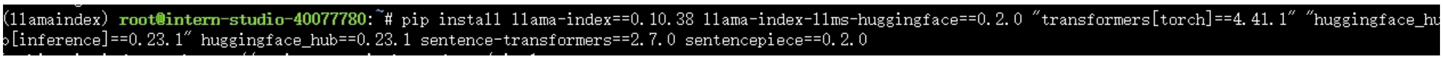
```
#创建新的conda环境，命名为 llamaindex，在命令行模式下运行：
conda create -n llamaindex python=3.10

#运行 conda 命令，激活 llamaindex 然后安装相关基础依赖 python 虚拟环境：
conda activate llamaindex
conda install pytorch==2.0.1 torchvision==0.15.2 torchaudio==2.0.2 pytorch-cuda=11.7 -c pytorch -c nvidia

#安装python 依赖包
pip install einops==0.7.0 protobuf==5.26.1
```



```
#
# To activate this environment, use
#
#     $ conda activate llamaindex
#
# To deactivate an active environment, use
#
#     $ conda deactivate

(base) root@intern-studio-40077780:~# conda activate llamaindex
(llamaindex) root@intern-studio-40077780:~# conda env list
# conda environments:
#
base                     /root/.conda
demo                     /root/.conda/envs/demo
langgpt                  /root/.conda/envs/langgpt
llamaindex            *  /root/.conda/envs/llamaindex
task                     /root/.conda/envs/task

(llamaindex) root@intern-studio-40077780:~# pip install einops==0.7.0 protobuf==5.26.1
Looking in indexes: https://pypi.tuna.tsinghua.edu.cn/simple
Collecting einops==0.7.0
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/29/0b/2d1c0ebfd092e25935b86509a9a817159212d82aa43d7fb07eca4eeff2c2/einops-0.7.0-py3-none-any.whl (44
kB)
Collecting protobuf==5.26.1
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/2c/2a/d2741cad35fa5f06d9c59dda3274e5727ca11075dfd7de3f69c100efdcad/protobuf-5.26.1-cp37-abi3-manylinu
x2014_x86_64.whl (302 kB)
Installing collected packages: protobuf, einops
Successfully installed einops-0.7.0 protobuf-5.26.1
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager, possibly rendering your sy
stem unusable.It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv. Use the --root-user-action option if you know what yo
u are doing and want to suppress this warning.
(llamaindex) root@intern-studio-40077780:~# 
```

安装 llamaindex

```
pip install llama-index==0.10.38 llama-index-llms-huggingface==0.2.0 "transformers[torch]==4.41.1" "huggingface_hub[inference]==0.23.1" huggingface_hub==0.23.1 sentence-transformers==2.7.0 sentencepiece==0.2.0
```



```
(llamaindex) root@intern-studio-40077780:~# pip install llama-index==0.10.38 llama-index-llms-huggingface==0.2.0 "transformers[torch]==4.41.1" "huggingface_hu
b[inference]==0.23.1" huggingface_hub==0.23.1 sentence-transformers==2.7.0 sentencepiece==0.2.0
```

下载Sentence Transformer模型

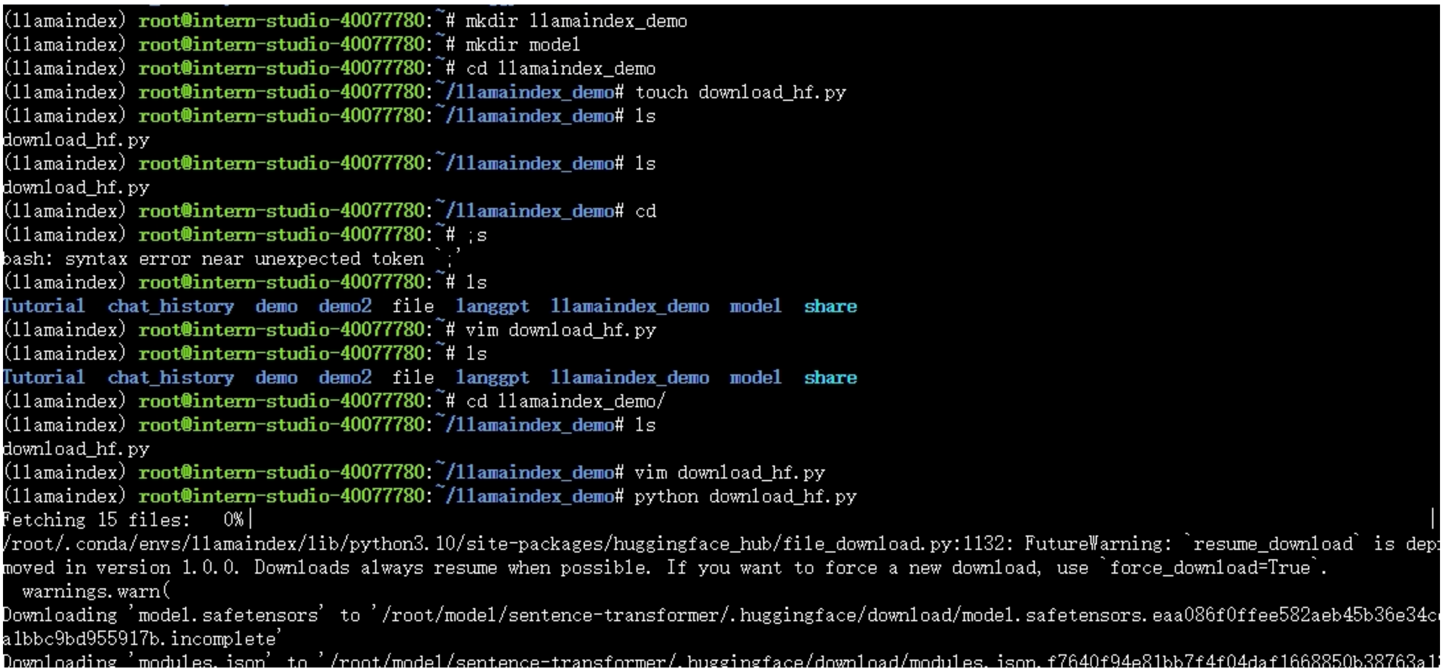
```
cd ~
mkdir llamaindex_demo
mkdir model
cd ~/llamaindex_demo
touch download_hf.py
```

打开 `download_hf.py` 贴入以下代码

```python
import os
# 设置环境变量
os.environ['HF_ENDPOINT'] = 'https://hf-mirror.com'
# 下载模型
os.system('huggingface-cli download --resume-download sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 --local-dir /root/model/sentence-transformer')
```

然后，在 `/root/llamaindex_demo` 目录下执行该脚本：

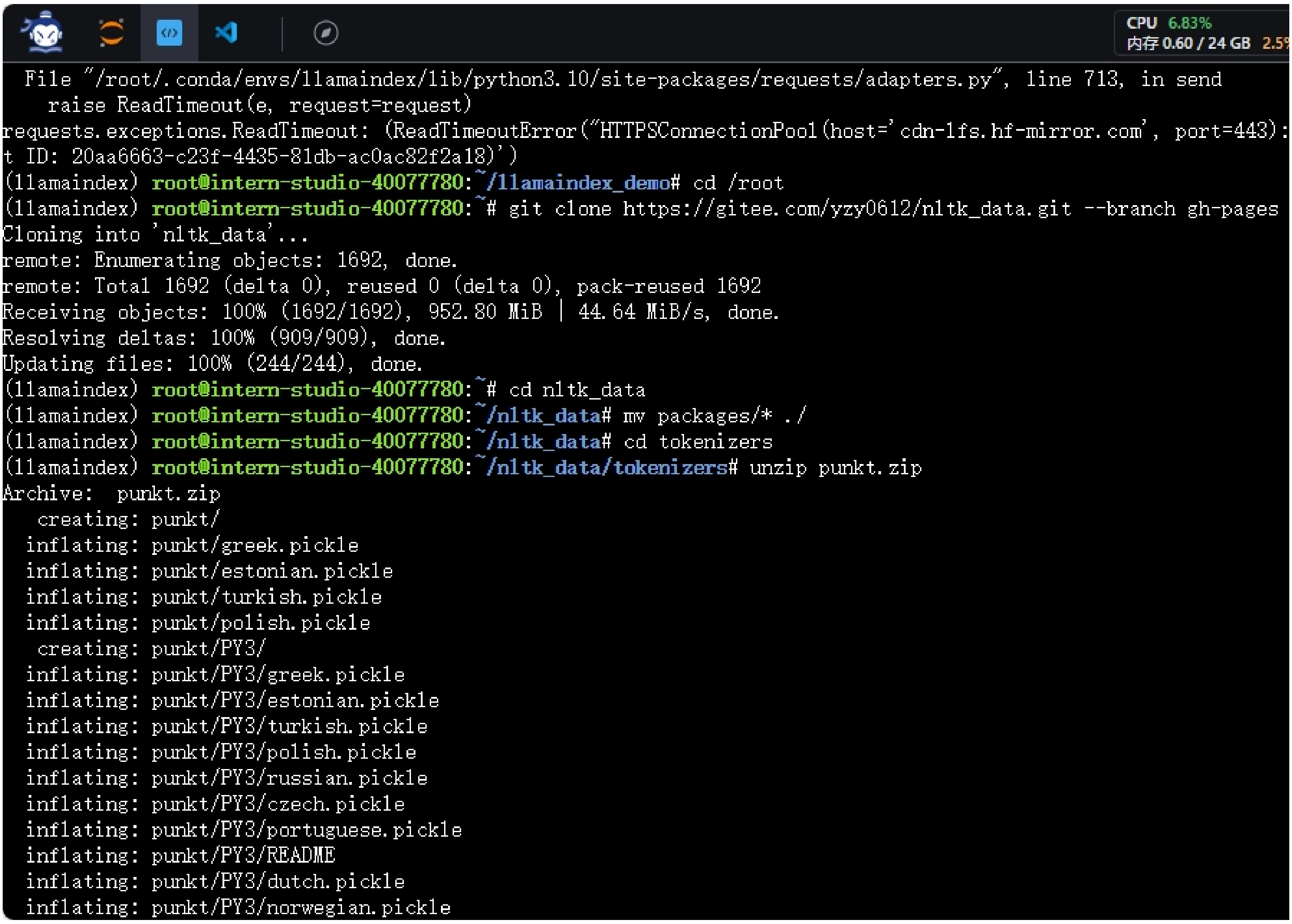
```
cd /root/llamaindex_demo
conda activate llamaindex
python download_hf.py
```



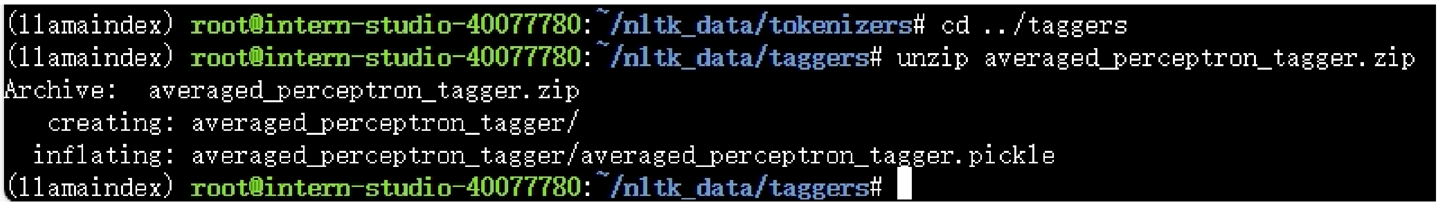
```
(llamaindex) root@intern-studio-40077780:~# mkdir llamaindex_demo
(llamaindex) root@intern-studio-40077780:~# mkdir model
(llamaindex) root@intern-studio-40077780:~# cd llamaindex_demo
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# touch download_hf.py
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# ls
download_hf.py
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# ls
download_hf.py
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# cd
(llamaindex) root@intern-studio-40077780:~# ;s
bash: syntax error near unexpected token `;'
(llamaindex) root@intern-studio-40077780:~# ls
Tutorial  chat_history  demo  demo2  file  langgpt  llamaindex_demo  model  share
(llamaindex) root@intern-studio-40077780:~# vim download_hf.py
(llamaindex) root@intern-studio-40077780:~# ls
Tutorial  chat_history  demo  demo2  file  langgpt  llamaindex_demo  model  share
(llamaindex) root@intern-studio-40077780:~# cd llamaindex_demo/
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# ls
download_hf.py
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# vim download_hf.py
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# python download_hf.py
Fetching 15 files:   0%|                                                                              |
/root/.conda/envs/llamaindex/lib/python3.10/site-packages/huggingface_hub/file_download.py:1132: FutureWarning: `resume_download` is dep
moved in version 1.0.0. Downloads always resume when possible. If you want to force a new download, use `force_download=True`.
  warnings.warn(
Downloading 'model.safetensors' to '/root/model/sentence-transformer/.huggingface/download/model.safetensors.eaa086f0ffee582aeb45b36e34c
a1bbc9bd955917b.incomplete'
Downloading 'modules.json' to '/root/model/sentence-transformer/.huggingface/download/modules.json.f7640f94e81bb7f4f04daf1668850b38763a1
```

下载NLTK相关资源

```
cd /root
git clone https://gitee.com/yzy0612/nltk_data.git  --branch gh-pages
cd nltk_data
mv packages/*  ./
cd tokenizers
unzip punkt.zip
cd ../taggers
unzip averaged_perceptron_tagger.zip
```



```
  File "/root/.conda/envs/llamaindex/lib/python3.10/site-packages/requests/adapters.py", line 713, in send
    raise ReadTimeout(e, request=request)
requests.exceptions.ReadTimeout: (ReadTimeoutError("HTTPSConnectionPool(host='cdn-lfs.hf-mirror.com', port=443):
t ID: 20aa6663-c23f-4435-81db-ac0ac82f2a18)')
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# cd /root
(llamaindex) root@intern-studio-40077780:~# git clone https://gitee.com/yzy0612/nltk_data.git --branch gh-pages
Cloning into 'nltk_data'...
remote: Enumerating objects: 1692, done.
remote: Total 1692 (delta 0), reused 0 (delta 0), pack-reused 1692
Receiving objects: 100% (1692/1692), 952.80 MiB | 44.64 MiB/s, done.
Resolving deltas: 100% (909/909), done.
Updating files: 100% (244/244), done.
(llamaindex) root@intern-studio-40077780:~# cd nltk_data
(llamaindex) root@intern-studio-40077780:~/nltk_data# mv packages/* ./
(llamaindex) root@intern-studio-40077780:~/nltk_data# cd tokenizers
(llamaindex) root@intern-studio-40077780:~/nltk_data/tokenizers# unzip punkt.zip
Archive:  punkt.zip
   creating: punkt/
  inflating: punkt/greek.pickle
  inflating: punkt/estonian.pickle
  inflating: punkt/turkish.pickle
  inflating: punkt/polish.pickle
   creating: punkt/PY3/
  inflating: punkt/PY3/greek.pickle
  inflating: punkt/PY3/estonian.pickle
  inflating: punkt/PY3/turkish.pickle
  inflating: punkt/PY3/polish.pickle
  inflating: punkt/PY3/russian.pickle
  inflating: punkt/PY3/czech.pickle
  inflating: punkt/PY3/portuguese.pickle
  inflating: punkt/PY3/README
  inflating: punkt/PY3/dutch.pickle
  inflating: punkt/PY3/norwegian.pickle
```



```
(llamaindex) root@intern-studio-40077780:~/nltk_data/tokenizers# cd ../taggers
(llamaindex) root@intern-studio-40077780:~/nltk_data/taggers# unzip averaged_perceptron_tagger.zip
Archive:  averaged_perceptron_tagger.zip
   creating: averaged_perceptron_tagger/
  inflating: averaged_perceptron_tagger/averaged_perceptron_tagger.pickle
(llamaindex) root@intern-studio-40077780:~/nltk_data/taggers# 
```

LlamaIndex HuggingFaceLLM

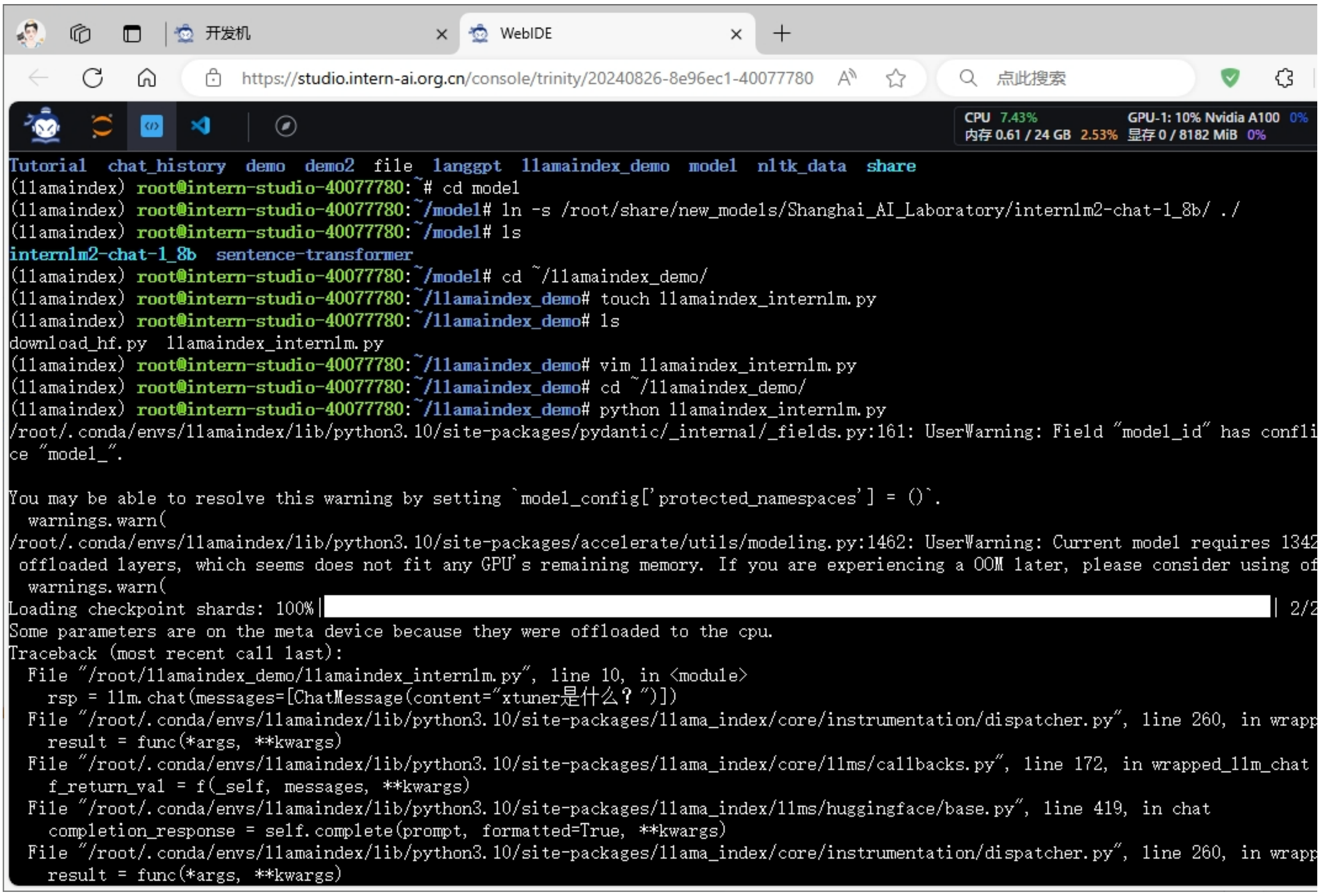
```
#运行以下指令，把 InternLM2 1.8B 软连接出来
cd ~/model
ln -s /root/share/new_models/Shanghai_AI_Laboratory/internlm2-chat-1_8b/ ./

#运行以下指令，新建一个python文件
cd ~/llamaindex_demo
touch llamaindex_internlm.py
```

```
#打开llamaindex_internlm.py 贴入以下代码
from llama_index.llms.huggingface import HuggingFaceLLMfrom
llama_index.core.llms import ChatMessagellm = HuggingFaceLLM(
    model_name="/root/model/internlm2-chat-1_8b",
```

```python
    tokenizer_name="/root/model/internlm2-chat-1_8b",
    model_kwargs={"trust_remote_code":True},
    tokenizer_kwargs={"trust_remote_code":True}
)

rsp = llm.chat(messages=[ChatMessage(content="xtuner是什么? ")])
print(rsp)
```

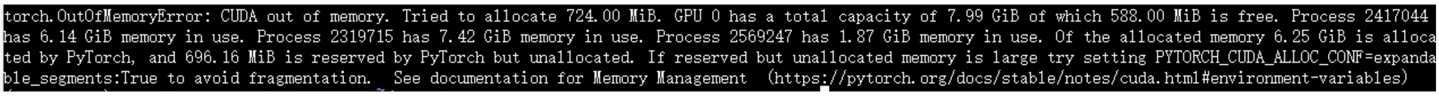
```
#运行
conda activate llamaindex
cd ~/llamaindex_demo/
python llamaindex_internlm.py
```

问了一个“xtuner是什么？”的问题，回答的不咋地


开发机
WebIDE
https://studio.intern-ai.org.cn/console/trinity/20240826-8e96ec1-40077780
点此搜索
CPU 7.43%
内存 0.61 / 24 GB 2.53%
GPU-1: 10% Nvidia A100 0%
显存 0 / 8182 MiB 0%

```
Tutorial  chat_history  demo  demo2  file  langgpt  llamaindex_demo  model  nltk_data  share
(llamaindex) root@intern-studio-40077780:~# cd model
(llamaindex) root@intern-studio-40077780:~/model# ln -s /root/share/new_models/Shanghai_AI_Laboratory/internlm2-chat-1_8b/ ./
(llamaindex) root@intern-studio-40077780:~/model# ls
internlm2-chat-1_8b  sentence-transformer
(llamaindex) root@intern-studio-40077780:~/model# cd ~/llamaindex_demo/
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# touch llamaindex_internlm.py
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# ls
download_hf.py  llamaindex_internlm.py
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# vim llamaindex_internlm.py
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# cd ~/llamaindex_demo/
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# python llamaindex_internlm.py
/root/.conda/envs/llamaindex/lib/python3.10/site-packages/pydantic/_internal/_fields.py:161: UserWarning: Field "model_id" has confli
ce "model_".

You may be able to resolve this warning by setting `model_config['protected_namespaces'] = ()`.
  warnings.warn(
/root/.conda/envs/llamaindex/lib/python3.10/site-packages/accelerate/utils/modeling.py:1462: UserWarning: Current model requires 1342
 offloaded layers, which seems does not fit any GPU's remaining memory. If you are experiencing a OOM later, please consider using of
  warnings.warn(
Loading checkpoint shards: 100%|██████████████████████████████████████████████████| 2/2
Some parameters are on the meta device because they were offloaded to the cpu.
Traceback (most recent call last):
  File "/root/llamaindex_demo/llamaindex_internlm.py", line 10, in <module>
    rsp = llm.chat(messages=[ChatMessage(content="xtuner是什么？")])
  File "/root/.conda/envs/llamaindex/lib/python3.10/site-packages/llama_index/core/instrumentation/dispatcher.py", line 260, in wrapp
    result = func(*args, **kwargs)
  File "/root/.conda/envs/llamaindex/lib/python3.10/site-packages/llama_index/core/llms/callbacks.py", line 172, in wrapped_llm_chat
    f_return_val = f(_self, messages, **kwargs)
  File "/root/.conda/envs/llamaindex/lib/python3.10/site-packages/llama_index/llms/huggingface/base.py", line 419, in chat
    completion_response = self.complete(prompt, formatted=True, **kwargs)
  File "/root/.conda/envs/llamaindex/lib/python3.10/site-packages/llama_index/core/instrumentation/dispatcher.py", line 260, in wrapp
    result = func(*args, **kwargs)
```

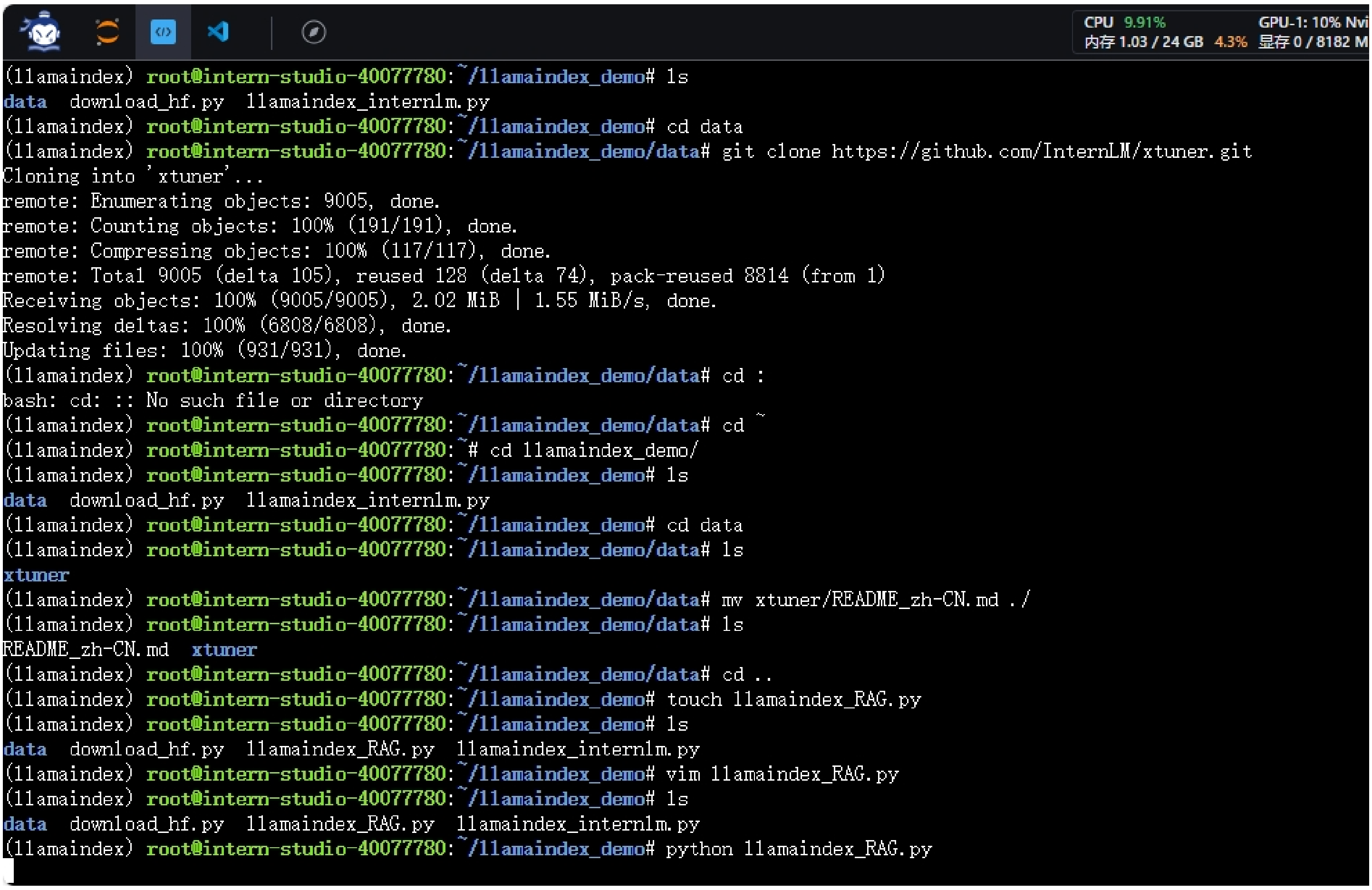
很无奈，过载了，开发机10%的cpu果然有点不够



```
torch.OutOfMemoryError: CUDA out of memory. Tried to allocate 724.00 MiB. GPU 0 has a total capacity of 7.99 GiB of which 588.00 MiB is free. Process 2417044
has 6.14 GiB memory in use. Process 2319715 has 7.42 GiB memory in use. Process 2569247 has 1.87 GiB memory in use. Of the allocated memory 6.25 GiB is alloca
ted by PyTorch, and 696.16 MiB is reserved by PyTorch but unallocated. If reserved but unallocated memory is large try setting PYTORCH_CUDA_ALLOC_CONF=expanda
ble_segments:True to avoid fragmentation.  See documentation for Memory Management  (https://pytorch.org/docs/stable/notes/cuda.html#environment-variables)
```

LlamaIndex RAG

```
#安装 LlamaIndex 词嵌入向量依赖
conda activate llamaindex
pip install llama-index-embeddings-huggingface==0.2.0 llama-index-embeddings-instructor==0.1.3

#运行以下命令，获取知识库
cd ~/llamaindex_demo
mkdir data
cd data
git clone https://github.com/InternLM/xtuner.git
mv xtuner/README_zh-CN.md ./

#运行以下指令，新建一个python文件
cd ~/llamaindex_demo
touch llamaindex_RAG.py
```

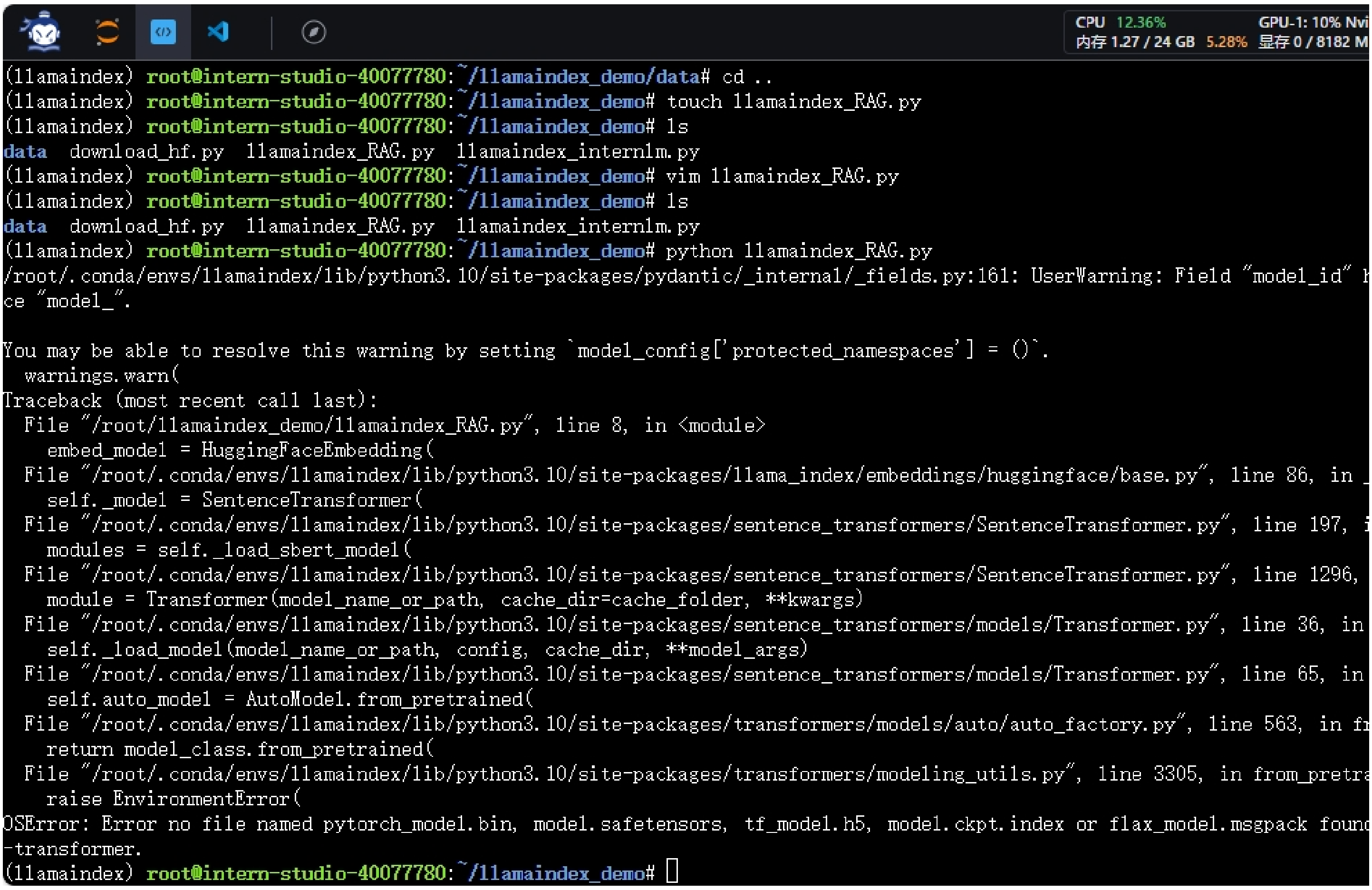


```
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# ls
data  download_hf.py  llamaindex_internlm.py
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# cd data
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo/data# git clone https://github.com/InternLM/xtuner.git
Cloning into 'xtuner'...
remote: Enumerating objects: 9005, done.
remote: Counting objects: 100% (191/191), done.
remote: Compressing objects: 100% (117/117), done.
remote: Total 9005 (delta 105), reused 128 (delta 74), pack-reused 8814 (from 1)
Receiving objects: 100% (9005/9005), 2.02 MiB | 1.55 MiB/s, done.
Resolving deltas: 100% (6808/6808), done.
Updating files: 100% (931/931), done.
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo/data# cd :
bash: cd: :: No such file or directory
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo/data# cd ~
(llamaindex) root@intern-studio-40077780:~# cd llamaindex_demo/
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# ls
data  download_hf.py  llamaindex_internlm.py
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# cd data
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo/data# ls
xtuner
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo/data# mv xtuner/README_zh-CN.md ./
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo/data# ls
README_zh-CN.md  xtuner
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo/data# cd ..
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# touch llamaindex_RAG.py
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# ls
data  download_hf.py  llamaindex_RAG.py  llamaindex_internlm.py
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# vim llamaindex_RAG.py
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# ls
data  download_hf.py  llamaindex_RAG.py  llamaindex_internlm.py
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# python llamaindex_RAG.py
```

llamaIndex_RAG.py写入如下代码，使大模型可以从词嵌入向量获得信息

```

from llama_index.core import VectorStoreIndex, SimpleDirectoryReader, Settings

from llama_index.embeddings.huggingface import HuggingFaceEmbedding
from llama_index.llms.huggingface import HuggingFaceLLM

#初始化一个HuggingFaceEmbedding对象，用于将文本转换为向量表示
embed_model = HuggingFaceEmbedding(
#指定了一个预训练的sentence-transformer模型的路径
    model_name="/root/model/sentence-transformer"
)
#将创建的嵌入模型赋值给全局设置的embed_model属性，
#这样在后续的索引构建过程中就会使用这个模型。
Settings.embed_model = embed_model

llm = HuggingFaceLLM(
    model_name="/root/model/internlm2-chat-1_8b",
    tokenizer_name="/root/model/internlm2-chat-1_8b",
    model_kwargs={"trust_remote_code":True},
    tokenizer_kwargs={"trust_remote_code":True}
)
#设置全局的llm属性，这样在索引查询时会使用这个模型。
Settings.llm = llm

#从指定目录读取所有文档，并加载数据到内存中
documents = SimpleDirectoryReader("/root/llamaindex_demo/data").load_data()
#创建一个VectorStoreIndex，并使用之前加载的文档来构建索引。
# 此索引将文档转换为向量，并存储这些向量以便于快速检索。
index = VectorStoreIndex.from_documents(documents)
# 创建一个查询引擎，这个引擎可以接收查询并返回相关文档的响应。
query_engine = index.as_query_engine()
response = query_engine.query("xtuner是什么?")

print(response)
```

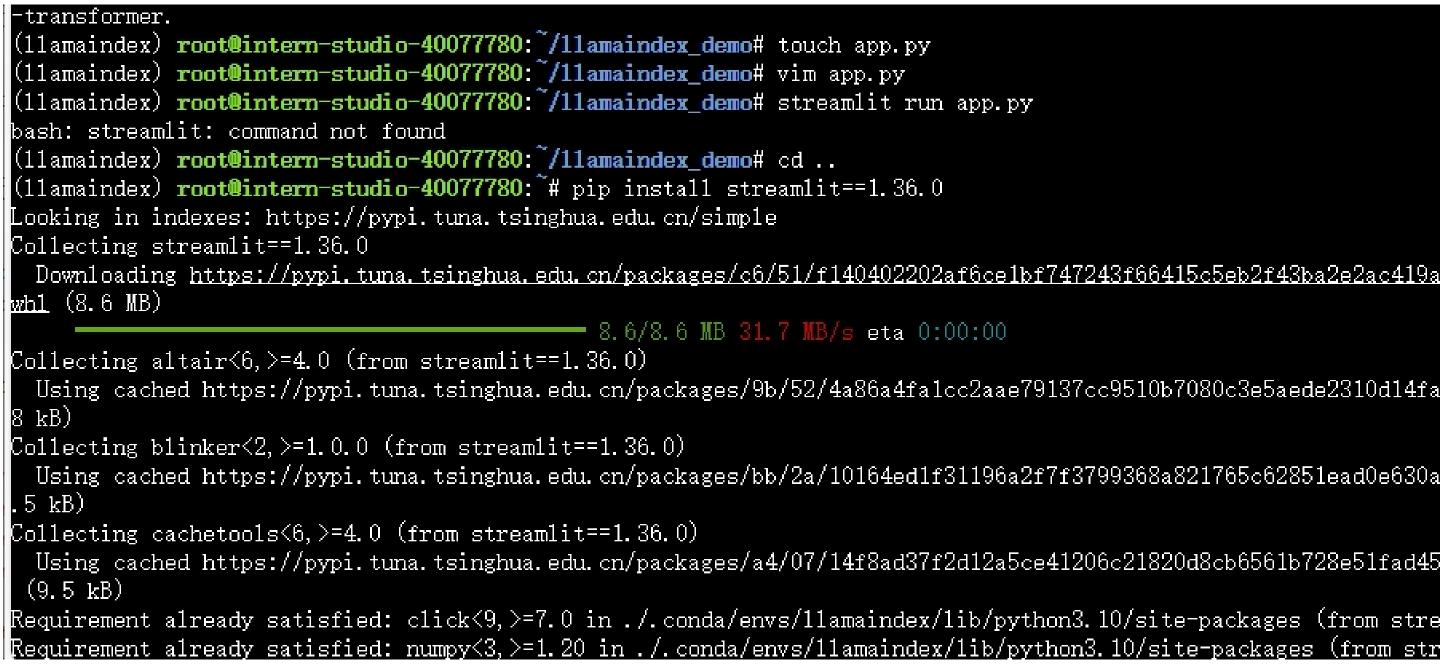
再次过载，无奈



```
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo/data# cd ..
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# touch llamaindex_RAG.py
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# ls
data  download_hf.py  llamaindex_RAG.py  llamaindex_internlm.py
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# vim llamaindex_RAG.py
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# ls
data  download_hf.py  llamaindex_RAG.py  llamaindex_internlm.py
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# python llamaindex_RAG.py
/root/.conda/envs/llamaindex/lib/python3.10/site-packages/pydantic/_internal/_fields.py:161: UserWarning: Field "model_id" h
ce "model_".

You may be able to resolve this warning by setting `model_config['protected_namespaces'] = ()`.
  warnings.warn(
Traceback (most recent call last):
  File "/root/llamaindex_demo/llamaindex_RAG.py", line 8, in <module>
    embed_model = HuggingFaceEmbedding(
  File "/root/.conda/envs/llamaindex/lib/python3.10/site-packages/llama_index/embeddings/huggingface/base.py", line 86, in _
    self._model = SentenceTransformer(
  File "/root/.conda/envs/llamaindex/lib/python3.10/site-packages/sentence_transformers/SentenceTransformer.py", line 197, i
    modules = self._load_sbert_model(
  File "/root/.conda/envs/llamaindex/lib/python3.10/site-packages/sentence_transformers/SentenceTransformer.py", line 1296,
    module = Transformer(model_name_or_path, cache_dir=cache_folder, **kwargs)
  File "/root/.conda/envs/llamaindex/lib/python3.10/site-packages/sentence_transformers/models/Transformer.py", line 36, in
    self._load_model(model_name_or_path, config, cache_dir, **model_args)
  File "/root/.conda/envs/llamaindex/lib/python3.10/site-packages/sentence_transformers/models/Transformer.py", line 65, in
    self.auto_model = AutoModel.from_pretrained(
  File "/root/.conda/envs/llamaindex/lib/python3.10/site-packages/transformers/models/auto/auto_factory.py", line 563, in fr
    return model_class.from_pretrained(
  File "/root/.conda/envs/llamaindex/lib/python3.10/site-packages/transformers/modeling_utils.py", line 3305, in from_pretra
    raise EnvironmentError(
OSError: Error no file named pytorch_model.bin, model.safetensors, tf_model.h5, model.ckpt.index or flax_model.msgpack found
-transformer.
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# 
```

LlamaIndex web

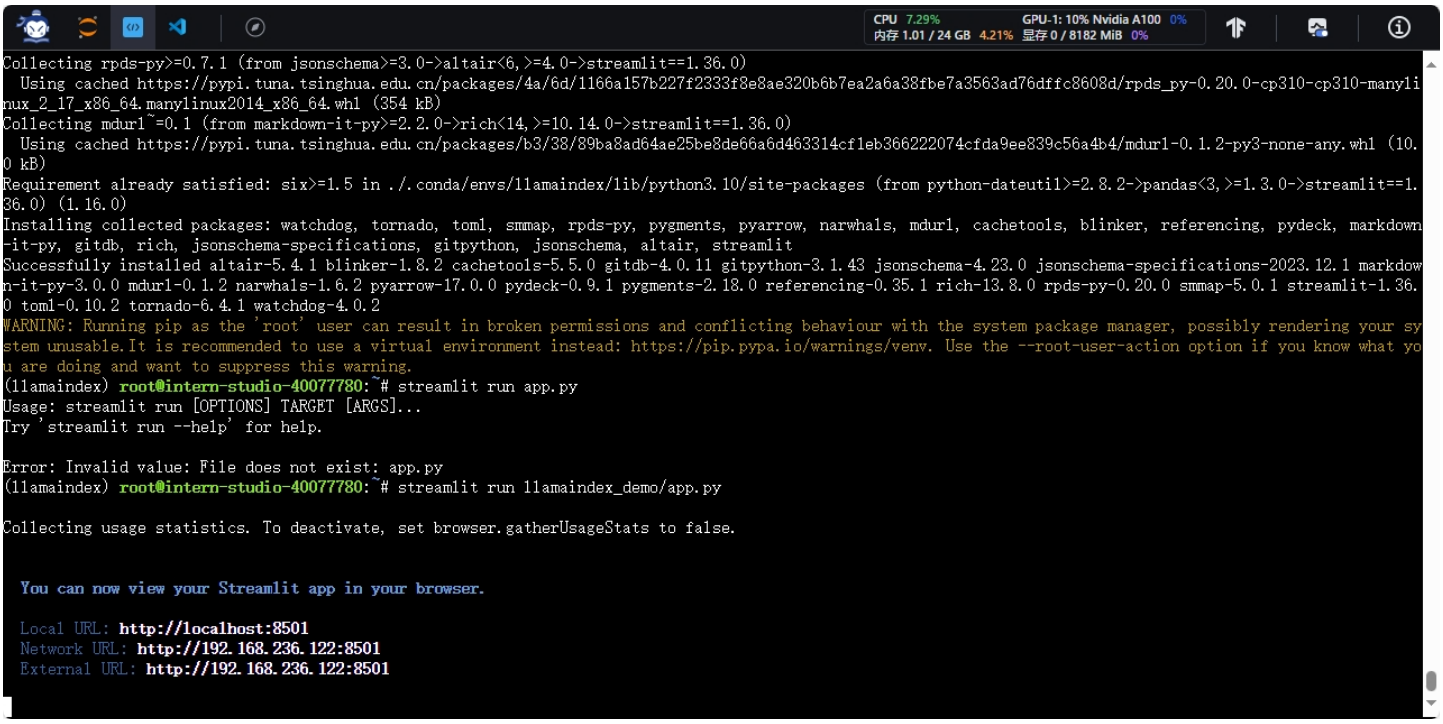
```
#运行之前首先安装依赖
pip install streamlit==1.36.0

#运行以下指令，新建一个python文件
cd ~/llamaindex_demo
touch app.py
```



```
-transformer.
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# touch app.py
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# vim app.py
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# streamlit run app.py
bash: streamlit: command not found
(llamaindex) root@intern-studio-40077780:~/llamaindex_demo# cd ..
(llamaindex) root@intern-studio-40077780:~# pip install streamlit==1.36.0
Looking in indexes: https://pypi.tuna.tsinghua.edu.cn/simple
Collecting streamlit==1.36.0
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/c6/51/f140402202af6ce1bf747243f66415c5eb2f43ba2e2ac419a
whl (8.6 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 8.6/8.6 MB 31.7 MB/s eta 0:00:00
Collecting altair<6,>=4.0 (from streamlit==1.36.0)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/9b/52/4a86a4fa1cc2aae79137cc9510b7080c3e5aede2310d14fa
8 kB)
Collecting blinker<2,>=1.0.0 (from streamlit==1.36.0)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/bb/2a/10164ed1f31196a2f7f3799368a821765c62851ead0e630a
.5 kB)
Collecting cachetools<6,>=4.0 (from streamlit==1.36.0)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/a4/07/14f8ad37f2d12a5ce41206c21820d8cb6561b728e51fad45
 (9.5 kB)
Requirement already satisfied: click<9,>=7.0 in ./.conda/envs/llamaindex/lib/python3.10/site-packages (from stre
Requirement already satisfied: numpy<3,>=1.20 in ./.conda/envs/llamaindex/lib/python3.10/site-packages (from str
```

```python
#app.py文件
import streamlit as st
from llama_index.core import VectorStoreIndex, SimpleDirectoryReader, Settings
from llama_index.embeddings.huggingface import HuggingFaceEmbedding
from llama_index.llms.huggingface import HuggingFaceLLM

st.set_page_config(page_title="llama_index_demo", page_icon="🦜🔗")
st.title("llama_index_demo")

# 初始化模型
@st.cache_resource
def init_models():
    embed_model = HuggingFaceEmbedding(
        model_name="/root/model/sentence-transformer"
    )
    Settings.embed_model = embed_model

    llm = HuggingFaceLLM(
        model_name="/root/model/internlm2-chat-1_8b",
        tokenizer_name="/root/model/internlm2-chat-1_8b",
        model_kwargs={"trust_remote_code": True},
        tokenizer_kwargs={"trust_remote_code": True}
    )
    Settings.llm = llm

    documents = SimpleDirectoryReader("/root/llamaindex_demo/data").load_data()
    index = VectorStoreIndex.from_documents(documents)
    query_engine = index.as_query_engine()

    return query_engine

# 检查是否需要初始化模型
if 'query_engine' not in st.session_state:
    st.session_state['query_engine'] = init_models()

def greet2(question):
    response = st.session_state['query_engine'].query(question)
    return response


# Store LLM generated responses
if "messages" not in st.session_state.keys():
    st.session_state.messages = [{"role": "assistant", "content": "你好，我是你的助手，有什么我可以帮助你的吗？"}]

    # Display or clear chat messages
for message in st.session_state.messages:
    with st.chat_message(message["role"]):
        st.write(message["content"])

def clear_chat_history():
    st.session_state.messages = [{"role": "assistant", "content": "你好，我是你的助手，有什么我可以帮助你的吗？"}]

st.sidebar.button('Clear Chat History', on_click=clear_chat_history)

# Function for generating LLaMA2 response
def generate_llama_index_response(prompt_input):
    return greet2(prompt_input)

# User-provided prompt
if prompt := st.chat_input():
    st.session_state.messages.append({"role": "user", "content": prompt})
    with st.chat_message("user"):
        st.write(prompt)

# Gegenerate_llama_index_response last message is not from assistant
if st.session_state.messages[-1]["role"] != "assistant":
    with st.chat_message("assistant"):
        with st.spinner("Thinking..."):
            response = generate_llama_index_response(prompt)
            placeholder = st.empty()
            placeholder.markdown(response)
    message = {"role": "assistant", "content": response}
    st.session_state.messages.append(message)
```

运行`streamlit run app.py`



CPU 7.29% 内存 1.01 / 24 GB 4.21%
GPU-1: 10% Nvidia A100 0% 显存 0 / 8182 MiB 0%

```
Collecting rpds-py>=0.7.1 (from jsonschema>=3.0->altair<6,>=4.0->streamlit==1.36.0)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/4a/6d/1166a157b227f2333f8e8ae320b6b7ea2a6a38fbe7a3563ad76dffc8608d/rpds_py-0.20.0-cp310-cp310-manyli
nux_2_17_x86_64.manylinux2014_x86_64.whl (354 kB)
Collecting mdurl~=0.1 (from markdown-it-py>=2.2.0->rich<14,>=10.14.0->streamlit==1.36.0)
  Using cached https://pypi.tuna.tsinghua.edu.cn/packages/b3/38/89ba8ad64ae25be8de66a6d463314cf1eb366222074cfda9ee839c56a4b4/mdurl-0.1.2-py3-none-any.whl (10.
0 kB)
Requirement already satisfied: six>=1.5 in ./.conda/envs/llamaindex/lib/python3.10/site-packages (from python-dateutil>=2.8.2->pandas<3,>=1.3.0->streamlit==1.
36.0) (1.16.0)
Installing collected packages: watchdog, tornado, toml, smmap, rpds-py, pygments, pyarrow, narwhals, mdurl, cachetools, blinker, referencing, pydeck, markdown
-it-py, gitdb, rich, jsonschema-specifications, gitpython, jsonschema, altair, streamlit
Successfully installed altair-5.4.1 blinker-1.8.2 cachetools-5.5.0 gitdb-4.0.11 gitpython-3.1.43 jsonschema-4.23.0 jsonschema-specifications-2023.12.1 markdow
n-it-py-3.0.0 mdurl-0.1.2 narwhals-1.6.2 pyarrow-17.0.0 pydeck-0.9.1 pygments-2.18.0 referencing-0.35.1 rich-13.8.0 rpds-py-0.20.0 smmap-5.0.1 streamlit-1.36.
0 toml-0.10.2 tornado-6.4.1 watchdog-4.0.2
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager, possibly rendering your sy
stem unusable.It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv. Use the --root-user-action option if you know what yo
u are doing and want to suppress this warning.
(llamaindex) root@intern-studio-40077780:~# streamlit run app.py
Usage: streamlit run [OPTIONS] TARGET [ARGS]...
Try 'streamlit run --help' for help.

Error: Invalid value: File does not exist: app.py
(llamaindex) root@intern-studio-40077780:~# streamlit run llamaindex_demo/app.py

Collecting usage statistics. To deactivate, set browser.gatherUsageStats to false.


  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8501
  Network URL: http://192.168.236.122:8501
  External URL: http://192.168.236.122:8501
```