一、环境配置

1.1配置基础环境：

进入开发机后，从官方环境复制运行 InternLM 的基础环境，命名为 `InternLM2_Huixiangdou`，在命令行模式下运行：

```
# studio-conda -t <target-conda-name> -o <origin-conda-name> 将预设的conda环境拷贝到指定的conda环境
studio-conda -o internlm-base -t InternLM2_Huixiangdou
```

使用conda命令激活InternLM2_Huixiangdou环境

```
# 查看所有conda环境名称
conda env list
# 激活所需要的环境
conda activate InternLM2_Huixiangdou  # yourself_env == InternLM2_Huixiangdou
```

1.2下载基础的文件

所有的作业和教程涉及的模型都已经存放再Intern Studio 开发共享文件中

```
cd /root && mkdir models  # 进入到root的文件下，并且创建models文件夹

# 复制BCE模型
"""
ln 是用来创建连接的命令
-s 选项表示创建符号连接
复制文件到一个目录
"""
ln -s /root/share/new_models/maidalun1020/bce-embedding-base_v1 /root/models/bce-embedding-base_v1
ln -s /root/share/new_models/maidalun1020/bce-reranker-base_v1 /root/models/bce-reranker-base_v1
ln -s /root/share/new_models/Shanghai_AI_Laboratory/internlm2-chat-7b /root/models/internlm2-chat-7b
```

1.3下载安装茴香豆

先安装茴香豆所需要的依赖包。

```
# pip install -r requirements.txt  直接安装requirements.txt
pip install protobuf==4.25.3 accelerate==0.28.0 aiohttp==3.9.3 auto-gptq==0.7.1
bcembedding==0.1.3 beautifulsoup4==4.8.2 einops==0.7.0 faiss-gpu==1.7.2 langchain==0.1.14
loguru==0.7.2 lxml_html_clean==0.1.0 openai==1.16.1 openpyxl==3.1.2 pandas==2.2.1
pydantic==2.6.4 pymupdf==1.24.1 python-docx==1.1.0 pytoml==0.1.21 readability-lxml==0.8.1
redis==5.0.3 requests==2.31.0 scikit-learn==1.4.1.post1 sentence_transformers==2.2.2
textract==1.6.5 tiktoken==0.6.0 transformers==4.39.3 transformers_stream_generator==0.0.5
unstructured==0.11.2
```

从茴香豆官方仓库下载茴香豆。

```
cd /root
# 克隆代码仓库
git clone https://github.com/internlm/huixiangdou && cd huixiangdou
git checkout 447c6f7e68a1657fce1c4f7c740ea1700bde0440
```

二、使用茴香豆搭建RAg助手

2.1修改配置文件

从github上下载的模型。进入到huixiangdou/config.ini中修改

```
embedding_model_path = "/root/models/bce-embedding-base_v1"
reranker_model_path = "/root/models/bce-reranker-base_v1"
local_llm_path = "/root/models/internlm2-chat-7b"
```

```
sed -i '6s#.*#embedding_model_path = "/root/models/bce-embedding-base_v1"#'
/root/huixiangdou/config.ini
sed -i '7s#.*#reranker_model_path = "/root/models/bce-reranker-base_v1"#'
/root/huixiangdou/config.ini
sed -i '29s#.*#local_llm_path = "/root/models/internlm2-chat-7b"#'
/root/huixiangdou/config.ini
```

```
[feature_store]
reject_throttle = 0.3612170956128262
embedding_model_path = "/root/models/bce-embedding-base_v1"
reranker_model_path = "/root/models/bce-reranker-base_v1"
work_dir = "workdir"

[web_search]
x_api_key = "${YOUR-API-KEY}"
domain_partial_order = ["openai.com", "pytorch.org", "readthedocs.io", "nvidi
save_dir = "logs/web_search_result"

[llm]
enable_local = 1
enable_remote = 0
client_url = "http://127.0.0.1:8888/inference"

[llm.server]
local_llm_path = "/root/models/internlm2-chat-7b"
local_llm_max_text_length = 3000
local_llm_bind_port = 8888
remote_type = "kimi"
remote_api_key = "${YOUR-API-KEY}"
remote_llm_max_text_length = 128000
remote_llm_model = "moonshot-v1-128k"
rpm = 500
```

2.2创建知识库

使用InternLM的Huixiangdou文档作为新知识检索来源，再不重新训练的情况下，打造一个huixiangdou技术问答助手

首先下载Huixiangdou语料：

```
cd /root/huixiangdou && mkdir repodir  # 进入到文件中并创建repodir文件夹
git clone https://github.com/internlm/huixiangdou --depth=1 repodir/huixiangdou  # 克隆github的仓库到repodir/huixiangdou中  depth=1 的意思是值克隆最新的一个提交历史
```

茴香豆建立接受和拒答两个数据库，用来再检索的过程中更加精确的判断提问的相关性，折两个数据库的来源分别是：

```
huixiangdou/resource/good_questions.json
huixiangdou/resource/bad_questions.json
```

增加茴香豆相关问题到接受问题示例中：

```
cd /root/huixiangdou  #  进入文件夹
mv resource/good_questions.json resource/good_questions_bk.json  # 复制文件
```

在good_questions.json中写一些内容

```
echo '[
    "mmpose中怎么调用mmyolo接口",
    "mmpose实现姿态估计后怎么实现行为识别",
    "mmpose执行提取关键点命令不是分为两步吗，一步是目标检测，另一步是关键点提取，我现在目标检测这部分的代码是demo/topdown_demo_with_mmdet.py demo/mmdetection_cfg/faster_rcnn_r50_fpn_coco.py checkpoints/faster_rcnn_r50_fpn_1x_coco_20200130-047c8118.pth    现在我想把这个mmdet的checkpoints换位yolo的，那么应该怎么操作",
    "在mmdetection中，如何同时加载两个数据集，两个dataloader",
    "如何将mmdetection2.28.2的retinanet配置文件改为单尺度的呢？",
    "1.MMPose_Tutorial.ipynb、inferencer_demo.py、image_demo.py、bottomup_demo.py、body3d_pose_lifter_demo.py这几个文件和topdown_demo_with_mmdet.py的区别是什么，\n2.我如果要使用mmdet是不是就只能使用topdown_demo_with_mmdet.py文件，",
    "mmpose 测试 map 一直是 0 怎么办？",
    "如何使用mmpose检测人体关键点？",
    "我使用的数据集是labelme标注的，我想知道mmpose的数据集都是什么样式的，全都是单目标的数据集标注，还是里边也有多目标然后进行标注",
    "如何生成openmmpose的c++推理脚本",
    "mmpose",
    "mmpose的目标检测阶段调用的模型，一定要是demo文件夹下的文件吗，有没有其他路径下的文件",
    "mmpose可以实现行为识别吗，如果要实现的话应该怎么做",
    "我在mmyolo的v0.6.0（15/8/2023)更新日志里看到了他新增了支持基于 MMPose 的 YOLOX-Pose，我现在是不是只需要在mmpose/project/yolox-Pose内做出一些设置就可以，换掉demo/mmdetection_cfg/faster_rcnn_r50_fpn_coco.py 改用mmyolo来进行目标检测了",
    "mac m1从源码安装的mmpose是x86_64的",
    "想请教一下mmpose有没有提供可以读取外接摄像头，做3d姿态并达到实时的项目呀？",
    "huixiangdou 是什么？",
    "使用科研仪器需要注意什么？",
    "huixiangdou 是什么？",
    "茴香豆 是什么？",
    "茴香豆 能部署到微信吗？",
    "茴香豆 怎么应用到飞书",
    "茴香豆 能部署到微信群吗？",
    "茴香豆 怎么应用到飞书群",
    "huixiangdou 能部署到微信吗？",
    "huixiangdou 怎么应用到飞书",
    "huixiangdou 能部署到微信群吗？",
    "huixiangdou 怎么应用到飞书群",
    "huixiangdou",
    "茴香豆",
    "茴香豆 有哪些应用场景",
    "huixiangdou 有什么用",
    "huixiangdou 的优势有哪些？",
    "茴香豆 已经应用的场景",
    "huixiangdou 已经应用的场景",
    "huixiangdou 怎么安装",
    "茴香豆 怎么安装",
    "茴香豆 最新版本是什么",
    "茴香豆 支持哪些大模型",
    "茴香豆 支持哪些通讯软件",
    "config.ini 文件怎么配置",
    "remote_llm_model 可以填哪些模型?"
```

```
]' > /root/huixiangdou/resource/good_questions.json
```

再创建一个测试用的问询列表，用来测试拒答流程是否起效：

```
cd /root/huixiangdou  # 进入文件夹
echo '["huixiangdou 是什么？","你好，介绍下自己"]' > ./test_queries.json  # 写内容到
test_queries.json
```

创建 RAG 检索过程中使用的向量数据库：

```
# 创建向量数据库存储目录
cd /root/huixiangdou && mkdir workdir
# 分别向量化知识语料、接受问题和拒绝问题中后保存到 workdir
python3 -m huixiangdou.service.feature_store --sample ./test_queries.json
```

2.3运行茴香豆知识助手

```
# 填入问题
sed -i '74s/.*/    queries = ["huixiangdou 是什么？", "茴香豆怎么部署到微信群", "今天天气怎么
样？"]/' /root/huixiangdou/huixiangdou/main.py
# 运行茴香豆
cd /root/huixiangdou/
python3 -m huixiangdou.main --standalone
```

结果：

```
04/16/2024 14:03:57 - [INFO] -aiohttp.access->>>  127.0.0.1 [16/Apr/2024:14:03:17 +0800] "POST /inference HTTP/1.1" 200 3315 "-" "python-requests/2.31.0"
2024-04-16 14:03:57.088 | INFO     | __main__:lark_send_only:79 - ErrorCode.SUCCESS, 茴香豆怎么部署到微信群，要部署茴香豆到微信群，请按照以下步骤操作：

1. **安装茴香豆**：
   - 首先，您需要从 GitHub 上下载茴香豆的源代码。
   - 在您的计算机上安装 Python 3.8 或更高版本。
   - 使用 `pip` 安装茴香豆的依赖项：`pip install -r requirements.txt`。

2. **准备数据**：
   - 您需要为茴香豆提供一些数据，例如问题、答案和相关知识。
   - 将数据整理成 JSON 格式，并保存为 `data.json` 文件。

3. **配置茴香豆**：
   - 打开 `config.ini` 文件，并根据您的需求配置参数。
   - 配置 `model` 参数，以选择您要使用的模型。
   - 配置 `data` 参数，以指定您要使用的数据文件。
   - 配置 `log` 参数，以指定日志文件的路径。

4. **运行茴香豆**：
   - 在终端中，导航到茴香豆的根目录。
   - 运行 `python huixiangdou.py` 启动茴香豆。
   - 茴香豆将开始处理数据，并准备回答用户的问题。

5. **集成到微信群**：
   - 在您的微信中，创建一个新的群聊。
   - 将茴香豆的 QQ 号添加到该群聊中。
   - 茴香豆将开始接收来自微信群的消息，并尝试回答用户的问题。

6. **测试和优化**：
   - 测试茴香豆的性能，并根据需要进行调整。
   - 您可以通过添加更多数据、调整参数或使用更高级的模型来提高茴香豆的性能。

请注意，以上步骤仅提供了一个基本的部署流程，您可能需要根据您的具体需求进行一些调整。同时，为了确保茴香豆的正常运行，请确保您的计算机具有足够的计算资源，并且您的数据集是干净、有组织且格式正确的。，['README_zh.md']
```