风险源RAG系统部署指南

1.环境配置

```
conda create -n flash python=3.10
```

```
conda activate flash
#下载transformer库，解压，安装，在transformer目录下执行
#https://github.com/huggingface/transformers/tree/7a25f8dfdba4c710d278d8312ef2522c5996a894
pip install -e .
```

```
pip install gradio==5.4.0 gradio_client==1.4.2 qwen-vl-utils==0.0.10
transformers-stream-generator==0.0.4  accelerate av
```

```
pip install torch==2.6.0 torchvision==0.21.0 torchaudio==2.6.0 --index-url
https://download.pytorch.org/whl/cu124
```

```
pip install flash-attn==2.6.1 --no-build-isolation
```

```
pip install -i https://pypi.tuna.tsinghua.edu.cn/simple langchain
langchain_community langchain_core sentence-transformers faiss-cpu pypdf fastapi
uvicorn python-multipart Pillow
pip install --upgrade autoawq
pip install chardet openpyxl
```

2.下载模型

```
pip install modelscope
modelscope download --model Qwen/Qwen2.5-VL-32B-Instruct-AWQ
modelscope download --model AI-ModelScope/bge-base-zh-v1.5
```

3.创建知识库索引

将《风险源清单.xlsx》文件上传到服务器目录kb_risk

```
#创建向量数据库
python vector1.py
```

这会在指定的 VECTOR_STORE_PATH 生成 FAISS 索引文件。

4.运行RAG脚本

将“风险源列表_清洗后.txt”上传至服务器

```
python qwen_32b_flash.py
```

```
收到图片：162_6157030_1.png，开始进行多步 RAG 分析...
步骤 1：生成图像描述...
生成描述（部分）：这是一张展示电梯门的照片。画面中可以看到一扇关闭的不锈钢材质电梯门，表面光滑且具有金属光泽，反射着光线。电梯门设计简洁现代，由两部分组成并以黑色边框装饰，显得坚固耐用。背景为浅色墙面与地板瓷砖，整体环...
步骤 1.5：生成候选风险源列表...
生成的候选列表：['办公', '机械', '机械设备', '电梯']
步骤 1.6：LLM 视觉验证候选列表...
LLM 验证后的风险源：['电梯']
最终选定的风险源：['电梯']
步骤 2：基于验证后的风险源检索法规...
检索到 30 篇原始文档。
步骤 2.5：过滤文档...
过滤后剩余 14 篇相关文档。
步骤 3a：生成风险识别与描述报告...
风险识别与描述报告生成完毕。
步骤 3b：生成综合管理措施...
综合管理措施生成完毕。
```

5.测试

将测试图片上传至服务器

```
bash analyze_images.sh
```

格式优化

```
python format_risk_json.py
```

输出为doc文档

```
python generate_doc_with_images.py
```