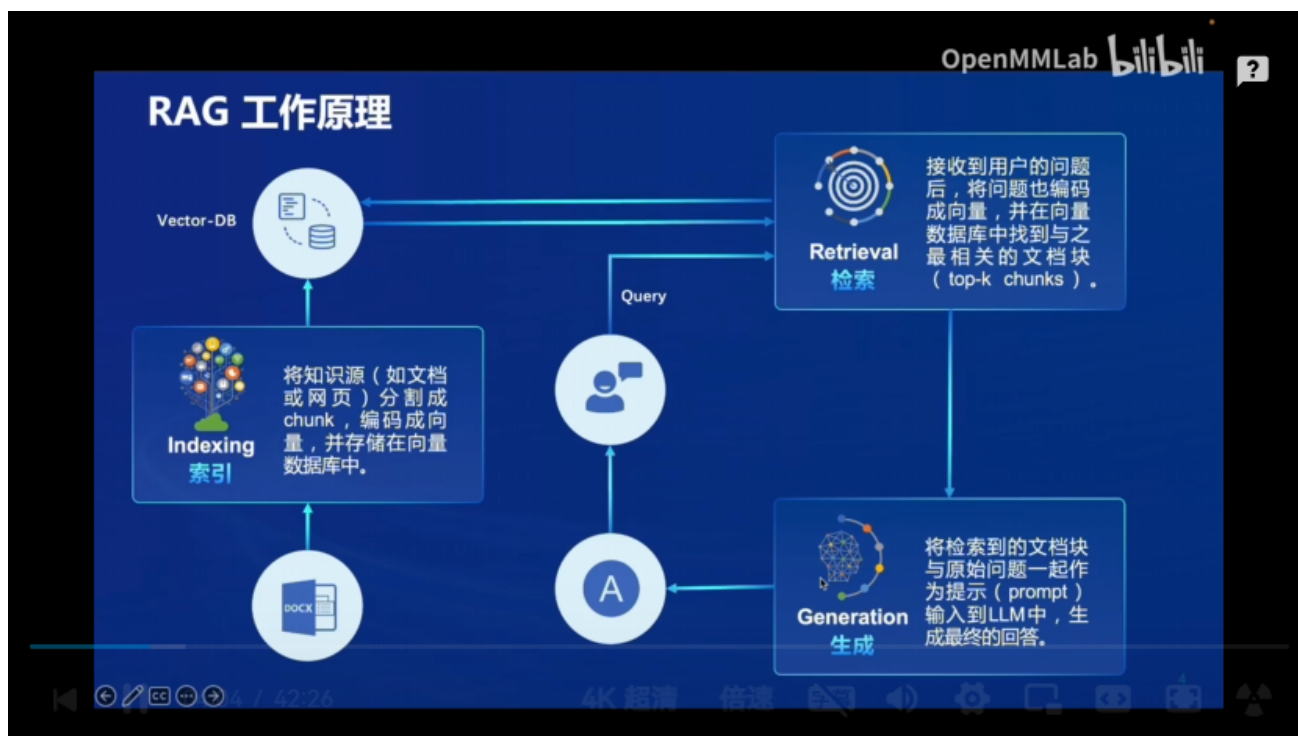


# InternLM 第三课——茴香豆

## 一.线上课程

### 1.RAG (Retrieval Augmented Generation)基础知识

- 结合了检索与生成
- 通过利用**外部数据库**增强LLM的性能-->解决LLMs在处理知识密集型任务时遇到的挑战
- 通过检索与用户输入相关的信息片段，结合这些信息生成更准确丰富的答案
- 应用：问答系统、文本生成、多模态、信息检索等

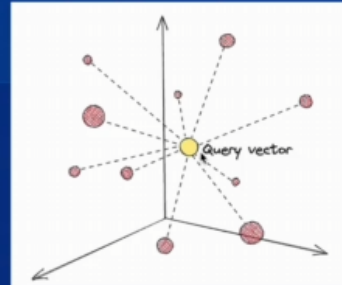


将文本及其他数据通过其他预训练的模型转换为固定长度的向量表示，这些向量能够捕捉文本的语义信息。

根据用户的查询向量，使用向量数据库快速找出最相关的向量的过程。通常通过计算余弦相似度或其他相似性度量来完成。

检索结果根据相似度得分进行排序，最相关的文档将被用于后续的文本生成。

包括使用更高级的文本编码技术，如句子嵌入或段落嵌入，以及对数据库进行优化以支持大规模向量搜索。



按 **Esc** 退出全屏

RAG的概念最早是由Meta ( Facebook ) 的Lewis等人在2020《Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks》中提出的。

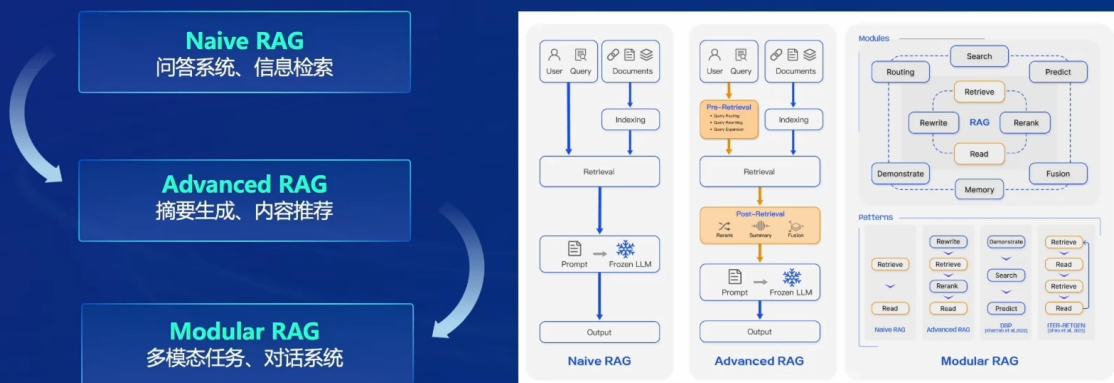



Image from: Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., Wang, H., 2024. Retrieval-Augmented Generation for Large Language Models: A Survey.

- 1.提高**向量数据库**质量：嵌入优化（结合稀疏检索、密集检索等）；索引优化（细粒度分割等）

- 查询优化：多查询；查询转换
- 上下文管理：rerank;上下文选择/压缩
- Retrieval：迭代检索；递归检索(CoT)；自适应检索 (Flare,Self-RAG)
- LLM**微调**：检索微调；生成微调等

OpenMMLab 

## RAG vs. 微调 ( Fine-tuning )

### RAG


- 非参数记忆，利用外部知识库提供实时更新的信息。
- 能够处理知识密集型任务，提供准确的事实性回答。
- 通过检索增强，可以生成更多样化的内容。

#### 适用场景

适用于需要结合最新信息和实时数据的任务：开放域问答、实时新闻摘要等。

**优势：**动态知识更新，处理长尾知识问题。

**局限：**依赖于外部知识库的质量和覆盖范围。依赖大模型能力。



### Fine-tuning

- 参数记忆，通过在特定任务数据上训练，模型可以更好地适应该任务。
- 通常需要大量标注数据来进行有效微调。
- 微调后的模型可能过拟合，导致泛化能力下降。

#### 适用场景

适用于数据可用且需要模型高度专业化的任务，如特定领域的文本分类、情感分析、文本生成等。

**优势：**模型性能针对特定任务优化。

**局限：**需要大量的标注数据，且对新任务的适应性较差。

## 4.RAG的测评框架

## 评估框架和基准测试

### 经典评估指标:

- 准确率 (Accuracy)
- 召回率 (Recall)
- F1分数 (F1 Score)
- BLEU分数 (用于机器翻译和文本生成)
- ROUGE分数 (用于文本生成的评估)

### RAG 评测框架:

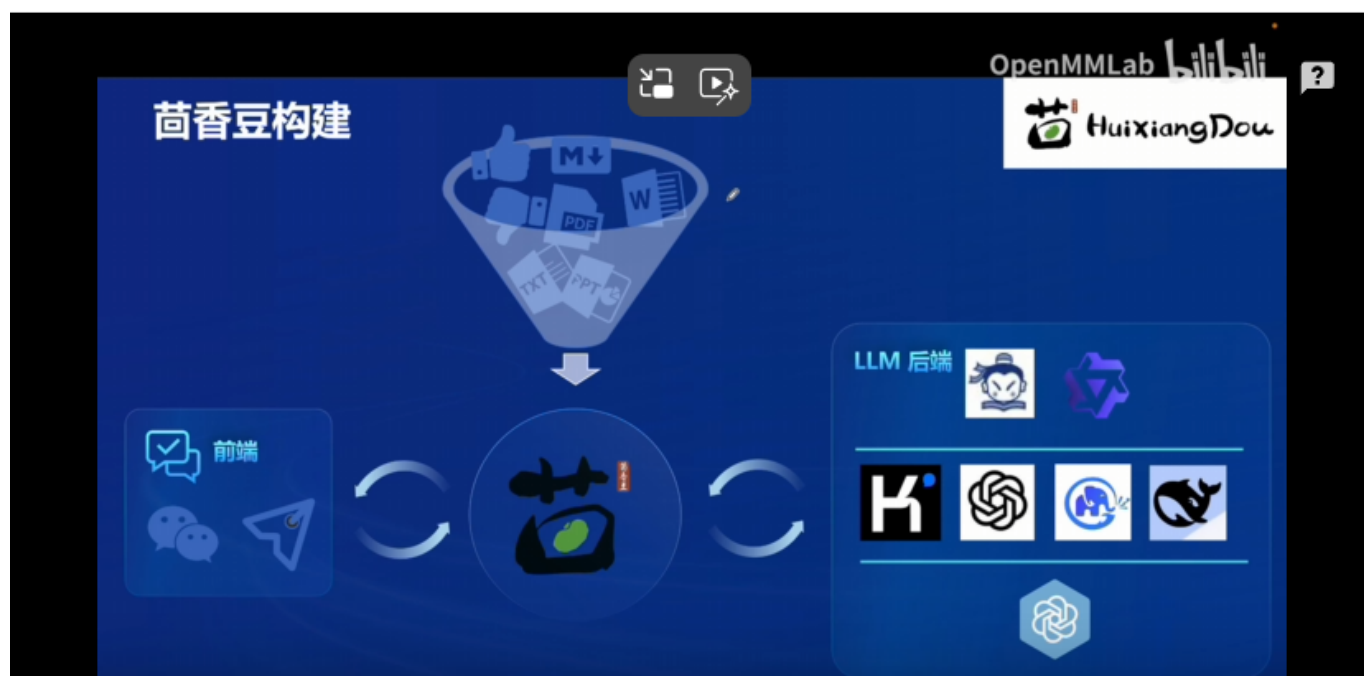
- 基准测试 - RGB、RECALL、CRUD
- 评测工具 - RAGAS、ARES、TruLens

Evaluation Framework	Evaluation Targets	Evaluation Aspects	Quantitative Metrics
RGB <sup>†</sup>	Retrieval Quality	Noise Robustness	Accuracy
	Generation Quality	Negative Rejection	EM
RECALL <sup>‡</sup>	Retrieval Quality	Information Integration	Accuracy
	Generation Quality	Counterfactual Robustness	Accuracy
RAGAS <sup>‡</sup>	Retrieval Quality	Context Relevance	R-Rate (Reappearance Rate)
	Generation Quality	Faithfulness	*
ARES <sup>‡</sup>	Retrieval Quality	Answer Relevance	Cosine Similarity
	Generation Quality	Context Relevance	Accuracy
TruLens <sup>‡</sup>	Retrieval Quality	Faithfulness	Accuracy
	Generation Quality	Answer Relevance	Accuracy
CRUD <sup>†</sup>	Retrieval Quality	Context Relevance	*
	Generation Quality	Faithfulness	*
	Retrieval Quality	Answer Relevance	*
	Generation Quality	Creative Generation	BLEU
	Retrieval Quality	Knowledge-intensive QA	ROUGE-L
	Generation Quality	Error Correction	BertScore
	Retrieval Quality	Summarization	RAGQuestEval
	Generation Quality		

Image from: Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., Wang, H., 2024. Retrieval-Augmented Generation for Large Language Models: A Survey.

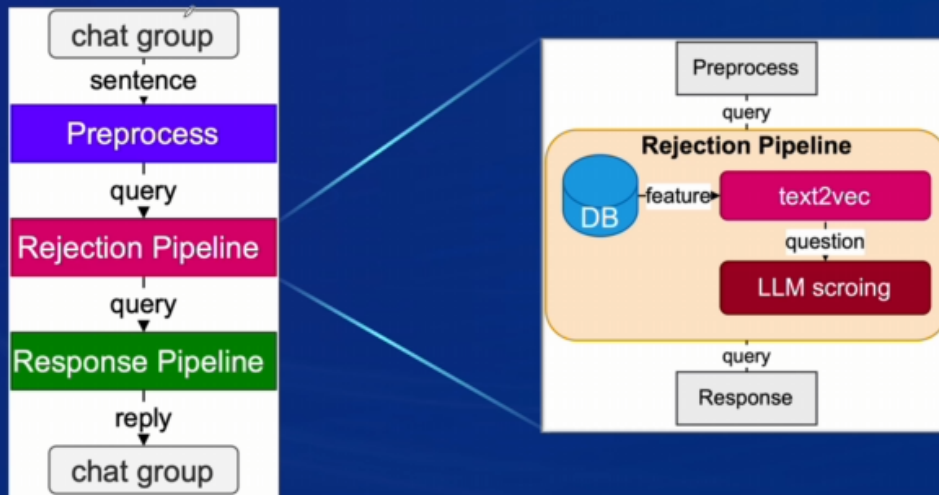
11

## 5. 茴香豆



- 支撑Markown,txt,Word,PPT等
- 设置接收问题和拒答问题
- 支持本地调用大模型（书生浦语，通义千问）；远端API

## 茴香豆的工作流



## 二.基础作业

### 1.在 InternLM Studio 上部署茴香豆技术助手

1. **创建开发机:** 50% A100 cuda11.7-conda
2. **配置环境:** studio-conda -o internlm-base -t InternLM2\_Huixiangdou (创建 InternLM2\_Huixiangdou环境)
3. **进入环境:** conda activate InternLM2\_Huixiangdou
4. **加载基础文件:**

```

# 创建模型文件夹
cd /root && mkdir models

# 复制BCE模型(软链接)
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

# 复制大模型参数(软链接)
ln -s /root/share/new_models/Shanghai_AI_Laboratory/internlm2-chat-7b /root/models/internlm2-chat-7b

```

## 5. 下载安装茴香豆所需依赖

```
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
```

## 6. 下载茴香豆：

```
cd /root
# 下载 repo
git clone https://github.com/internlm/huixiangdou && cd huixiangdou
git checkout 447c6f7e68a1657fce1c4f7c740ea1700bde0440
```

## 7. 修改配置文件

a.修改向量数据库和词嵌入模型（使用网易的BCE模型，不管有无Reranker，都比bge-m3-large好，这么厉害！）：`sed -i '6s#.*#embedding_model_path = "/root/models/bce-embedding-base_v1"' /root/huixiangdou/config.ini`

注：BCE模型地址：[netease-youdao/BCEembedding: Netease Youdao's open-source embedding and reranker models for RAG products. \(github.com\)](https://github.com/netease-youdao/BCEembedding)

b.修改用于检索的重排序模型（同样使用BCE模型，配上各种不同的embedding，效果都挺好）：`sed -i '7s#.*#reranker_model_path = "/root/models/bce-reranker-base_v1"' /root/huixiangdou/config.ini`

c.修改本次选用的模型（选用chat-7b模型）：`sed -i '29s#.*#local_llm_path = "/root/models/internlm2-chat-7b"' /root/huixiangdou/config.ini`

d.检查一下：`vim /root/huixiangdou/config.ini`

## 8. 创建知识库：

- 下载茴香豆语料

```
cd /root/huixiangdou && mkdir repodir
```

```
# 仅仅clone最新的提交历史，文件保存在repodir/huixiangdou文件夹中  
git clone https://github.com/internlm/huixiangdou --depth=1  
repodir/huixiangdou
```

- 增加接受问题列表（放在 `huixiangdou/resource/good_questions.json` 中）



```
cd /root/huixiangdou\

# 将 good_questions.json 文件重命名为 good_questions_bk.json
mv resource/good_questions.json resource/good_questions_bk.json

# 覆盖写入good_questions.json
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

```

- 查看拒绝问题列表: `vim huixiangdou/resource/bad_questions.json`

默认内置了一些拒绝回答的问题

- 测试拒答有没有效果:

```

cd /root/huixiangdou

echo '[
    "huixiangdou 是什么?",
    "你好, 介绍下自己"
]' > ./test_queries.json

# 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

```

1) \*\*第一个问题`huixiangdou是什么`不在bad\_question里面, 得到了回复(但是看不懂)\*\*



2) \*\*第二个问题`你好, 介绍下自己`直接就拒绝了:\*\*



\*\*3) 验证: 此时`/root/huixiangdou/workdir/negative.txt`和  
`/root/huixiangdou/workdir/postive.txt`各新增一条记录\*\*

## 9. 运行茴香豆知识助手

```
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
```

```
cd /root/huixiangdou/ python3 -m huixiangdou.main --standalone
```

第一行命令使用了sed工具来编辑文件，具体来说编辑了 /root/huixiangdou/huixiangdou/main.py 文件的第74行，将原来的内容替换为 queries = ["huixiangdou 是什么?", "茴香豆怎么部署到微信群", "今天天气怎么样?"]。

**工作流程 (推测):** 在加载完模型之后，应该是结

合 /root/huixiangdou/huixiangdou/service/worker.py 中设置的prompt来完成的:\*\*

- o self.SCORING\_QUESTION\_TEMPLTE = "{}\n请仔细阅读以上内容，判断句子是否是个有主题的疑问句，结果用 0~10 表示。直接提供得分不要解释。判断标准：有主语谓语宾语并且是疑问句得 10 分；缺少主谓宾扣分；陈述句直接得 0 分；不是疑问句直接得 0 分。直接提供得分不要解释。'->得分
- o self.TOPIC\_TEMPLATE = '告诉我这句话的主题，直接说主题不要解释：{}'->得到一个主题
- o 进入retriever进行检索 ->输出top1 file:
- o self.SCORING\_RELAVANCE\_TEMPLATE = '问题：{}\n材料：{}\n请仔细阅读以上内容，判断问题和材料的关联度，用0~10表示。判断标准：非常相关得 10 分；完全没关联得 0 分。直接提供得分不要解释。'\n'->得分
- o self.GENERATE\_TEMPLATE = '材料：{}\n问题：{}\n 请仔细阅读参考材料回答问题。'->得到结果
- o 将结果输出INFO

## 10. 结果:

```
04/10/2024 01:20:17 - [INFO] - aiohttp.access->>> 127.0.0.1 [10/Apr/2024:01:20:06 +0800] "POST /inference HTTP/1.1" 200 2369 "-" "python-requests/2.31.0"
2024-04-10 01:20:17.447 | INFO | _main :lark send_only:79 - ErrorCode.SUCCESS, huixiangdou 是什么?, HuixiangDou 是一个基于 LLM (大型语言模型) 的群组聊天助手。它使用两阶段管道来处理群组聊天场景，以回答用户问题，而不会导致信息过载。该工具的优点包括:
1. 两阶段管道: HuixiangDou 使用一个两阶段管道来处理群组聊天。首先，它使用一个预训练的模型来快速判断一个消息是否需要被处理。如果消息需要被处理，它将使用另一个模型来生成回答。这种设计可以避免信息过载，并提高处理效率。
2. 低成本: HuixiangDou 只需要 1.5GB 的内存，并且不需要进行训练，因此它的成本相对较低。
3. 完整套件: HuixiangDou 提供了 Web, Android 和管道源代码的完整套件，这些代码是工业级和商业上可行的。
4. 运行场景: HuixiangDou 已经在多个场景中运行，包括 Wechat 和 Feishu 群组。
5. 开源: HuixiangDou 的源代码是开源的，并且可以在 [OpenXLab](https://openxlab.org.cn/apps/detail/tpoisonooo/huixiangdou-web) 上使用，用户可以轻松地构建自己的知识助手，而无需编写任何代码。
HuixiangDou 的更多信息可以在 [arxiv2401.08772](https://arxiv.org/abs/2401.08772) 论文中找到。，["README.md"]
```

```
04/10/2024 01:20:36 - [INFO] -aiohttp.access->>> 127.0.0.1 [10/Apr/2024:01:20:22 +0800] "POST /inference HTTP/1.1" 200 3315 "-" "python-requests/2.31.0"
2024-04-10 01:20:36.514 | 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。 , ['REA
DME_zh.md']

2024-04-10 01:20:37.961 | INFO | __main__:lark_send_only:79 - ErrorCode.UNRELATED, 今天天气怎么样? , , ['HuixiangDou.pdf']
```

## 2.在茴香豆 Web 版中创建自己领域的知识问答助手

我上传了一本《financial machine learning》的书，跟茴香豆进行对话。

感觉整体效果一般：

- 1) 我没有设置拒绝回答的问题，但是不知道为什么询问算法都会提示**问题敏感**，我也不太清楚原因。
- 2) 我试了一下让茴香豆回答某一章的内容，也完全文不对题。
- 3) 大海捞针我也试了一下，效果也比较一般。
- 4) 具体对话内容如下：



这本书的主题是什么



这本书的主题是金融机器学习。它涵盖了各种机器学习技术在金融领域的应用，包括资产定价、投资组合优化、风险建模和预测市场动态等方面。它强调了机器学习如何成为金融研究的一个重要工具，帮助研究人员更好地理解金融市场现象和机制。

参考文档:

financial machine learning.pdf



金融机器学习有哪些应用途径



金融机器学习可以应用于以下途径:

1. 资产定价: 机器学习模型可以用于预测股票、债券等资产的未来价格变化。这有助于投资者进行资产配置和交易决策。
2. 风险管理: 机器学习模型可以用于评估不同投资组合的风险水平, 并帮助投资者优化投资组合以最大化收益和降低风险。

化收益和降低风险。

3. 高频交易: 机器学习模型可以分析市场微观结构数据, 识别交易机会并进行高频交易。
4. 信用评分: 机器学习模型可以分析借款人的历史数据和其他特征, 为银行和金融机构提供更准确的信用评分和贷款决策。
5. 保险精算: 机器学习模型可以分析保险索赔数据和风险因素, 帮助保险公司确定保费和理赔金额。
6. 欺诈检测: 机器学习模型可以分析交易数据和用户行为, 检测信用卡欺诈、网络钓鱼等欺诈行为。
7. 投资组合优化: 机器学习模型可以分析不同资产的历史表现和相关性, 帮助投资者构建最优的投资组合以实现收益最大化和风险最小化。

总的来说, 金融机器学习可以在投资决策、风险管理、高频交易、信贷评估、保险精算、欺诈检测和投资组合优化等多个领域发挥作用, 提高金融市场的效率和参与者的利益。

参考文档:

financial machine learning.pdf



讲讲这一节:  
1.6 Economic Content (Two Cultures of Financial Economics)



本节内容主要讨论了金融经济学的两种文化, 即经济学和金融学。虽然两者都属于金融经济学的范畴, 但它们在研究方法和理论基础上有很大差异。

1. **经济学**: 经济学是一种社会科学, 它研究的是人类如何分配稀缺资源以满足不断增长的需求。经济学家通常使用供求模型来解释市场现象, 并通过微观和宏观经济理论来分析个人和整体经济行为。在金融经济学中, 经济学家关注的问题包括资产定价、风险管理、市场效率和公司财务等。
2. **金融学**: 金融学是一种研究金融市场、金融机构和金融产品的学科。金融学家通常使用金融市场理论和公司财务理论来解释金融现象, 并关注投资组合管理、风险管理和金融工程等领域。在金融经济学中, 金融学家关注的问题包括证券市场、衍生品市场、金融中介和公司治理等。

尽管经济学和金融学在某些方面存在重叠, 但它们的视角和方法有所不同。经济学家更注重理论分析和实证研究, 而金融学家则更注重实践应用和市场动态。了解这两种文化对于理解金融经济学的不同研究方向和实际应用非常重要。

[Empty]: query is not a question

讲讲这一节：Economic Content (Two Cultures of Financial Economics)

很抱歉，由于之前我提供的信息有误，我将重新回答您的问题。

本节内容讨论了金融经济学的两个不同文化，即经济学和金融学。尽管它们都属于金融经济学的范畴，但在研究方法和理论基础方面存在差异。

1. **经济学**：经济学是一门社会科学，研究如何分配有限的资源以满足不断增长的需求。经济学家通常使用供求模型来解释市场现象，并通过微观经济理论和宏观经济理论来分析个人和整体经济行为。在金融经济学中，经济学家关注的问题包括资产定价、风险管理、市场效率和公司财务等。

2. **金融学**：金融学是研究金融市场、金融机构和金融产品的学科。金融学家通常使用金融市场理论和公司财务理论来解释金融现象，并关注投资组合管理、风险管理和金融工程等领域。在金融经济学中，金融学家关注的问题包括证券市场、衍生品市场、金融中介和公司治理等。

尽管经济学和金融学在某些方面存在重叠，但它们的视角和方法有所不同。经济学家更注重理论分析和实证研究，而金融学家则更注重实践应用和市场动态。了解这两种文化对于理解金融经济学的不同研究方向和实际应用非常重要。

[Empty]: query is not a question

这本书讲了哪些算法？

检测到敏感内容，无法显示

参考文档：

financial machine learning.pdf

[Empty]: Reply has a high relevance to prohibited topics.

这本书讲了哪些机器学习算法？

这本书讲述了以下几种机器学习算法：

1. 线性回归
2. 神经网络
3. 决策树
4. 支持向量机
5. 随机森林
6. AdaBoost

5. 随机森林
6. AdaBoost
7. Gradient Boosting
8. K-近邻算法
9. 聚类算法
10. 深度学习

这些算法被应用于金融市场预测、投资组合优化、风险管理等多个领域。

参考文档：

financial machine learning.pdf

## 三.进阶作业（布置到飞书）

参考[茴香豆零编程接入飞书 - 飞书云文档 \(feishu.cn\)](https://feishu.cn) 进行创建

具体而言，我创建了一个名字叫“金融工程机器学习”的知识库，上传了42篇论文，参考文档的做法创造了一个名为“文献”的机器人，效果如图所示，还不错：



zyf

茴香豆是什么



1 条回复



文献 机器人 | 文献查阅

| 回复 zyf: 茴香豆是什么

茴香豆是一种豆类食品，通常用于制作各种菜肴和调味品。它们被广泛用于中国和日本等地的烹饪中，具有独特的味道和香气。在烹饪中，茴香豆可以用于增加菜肴的深度和复杂度，例如在红烧肉类或炖菜中使用。此外，茴香豆还可以作为佐料，用于炒饭、炒面或其他面食中，以增加风味。在某些地区，茴香豆也被用作制作豆腐、豆腐皮或其他豆制品的原材料。总之，茴香豆是一种多功能且受欢迎的食材，可以在各种烹饪中发挥重要作用。



zyf

@文献 这篇文章“Antweiler\_Frank\_2004\_Is all that talk just noise.pdf”讲了什么?



1 条回复



文献 机器人 | 文献查阅 17:49

| 回复 zyf: @文献 这篇文章“Antweiler\_Frank\_2004\_Is all that talk just noise.pdf”讲了什么?

Answer: "Is all that talk just noise?" is a paper written by Walter Antweiler and Michael Frank published in 2004. The paper investigates the information content of internet stock message boards and explores whether the discussions on these boards actually provide valuable insights for investors or if they are just "noise". Through their analysis, the authors find that while message board discussions can sometimes provide useful information, they are not consistently reliable and may lead to incorrect investment decisions. The paper concludes that investors should exercise caution when relying solely on message board discussions for investment advice.

Some key points from the paper:

- The authors analyze over 3 million messages from 17 message boards covering 1,600 stocks over a 4-year period.
- They find that message boards do sometimes provide useful information that can help predict stock returns, though the signals are noisy and unreliable.
- The impact of message board discussions on stock prices tends to be short-lived, lasting only a few days.
- The authors develop a "bullishness" index based on message board discussions, which they find to be somewhat predictive of future returns. However, they caution that this index is not always accurate and should be used with care.
- Overall, the paper concludes that while message board discussions can occasionally provide valuable insights, they are not







zyl

finbert是什么? @文献



1 条回复



文献 机器人 | 文献查阅 17:52

| 回复 zyl: finbert是什么? @文献

FinBERT是一个专门为金融领域设计的大型语言模型，基于Google的BERT算法。它是通过在大量金融文本上进行预训练和微调来开发的。与一般的BERT模型相比，FinBERT更擅长分析金融文本，并且在情感分类、环境、社会和治理（ESG）讨论识别等任务中表现更优秀。与传统的自然语言处理算法相比，FinBERT在识别金融文本中的情感信息方面具有优势，这主要是因为FinBERT熟悉金融领域的词汇和语境。



zyl

和传统的bert有什么区别? FINBERT的训练数据集是什么?



1 条回复



文献 机器人 | 文献查阅 17:54

| 回复 zyl: 和传统的bert有什么区别? FINBERT的训练数据集是什么?

FINBERT是谷歌的BERT算法的一种变体，它被特别设计用于处理金融文本。与传统BERT相比，FINBERT有以下几个区别：

1. 训练数据集不同：FINBERT是在大量金融文本数据集上预训练的，包括公司年报、季度报告、分析师报告和财报电话会议记录等。这使得FINBERT在处理金融文本时能够更好地捕捉金融词汇和行业术语。
2. 任务适应性：FINBERT是通过在金融文本数据集上进行预训练，并针对金融领域的NLP任务进行微调而设计的。这使得FINBERT在处理金融文本任务（如情感分类和ESG讨论标记）时表现更出色。
3. 考虑上下文信息：FINBERT使用了Transformer编码器，可以考虑到单词的上下文信息，这有助于FINBERT在处理金融文本时更好地捕捉语义和逻辑关系。
4. 词汇表不同：FINBERT使用了自己的金融词汇表（FinVocab），其中包含金融领域特有的词汇，这些词汇在传统BERT的词汇表中可能不常见。
5. 性能优势：通过实验，作者发现FINBERT在金融文本情感分类任务上比传统BERT和其他NLP算法具有显著的优势，特别是在训练样本较小时。

所以，FINBERT的训练数据集主要是金融文本，并且通过针对金融领域的NLP任务进行微调，FINBERT能够更好地处理金融文本数据。

