金融RAG算法

一. 背景

某学校竞赛题目。相关需求如下:

1. 金融基础信息RAG问答: 能从文档中直接找到相关答案。

   示例:

   Q: “2023年，**基金的基金份额为多少”？

   A: “2023年，**基金在报告期末的基金份额总额为265688785223份。”

2. 金融核心指标统计分析: 从文档中找到多个维度的数据指标，然后做统计分析

   示例:

   Q: "2023年**基金在报告期末的可供分配利润比2022年低多少?"

   A: "2023年**基金在报告期末的可供分配利润比2022年低1212131415.65元。"

3. 结构化信息抽取: 从文档中抽取指定信息，并以Json格式进行序列化

   示例:

   Q:"请以json格式抽取2023年报告期末，**基金的股票名称，需要包含的主键为股票名称，键值为净值比例，以百分数表示，保留2位有效数字。"

   A:“{"中国洪倩":"9.65%"，“绿叶制药”：“6.45%”}”

4. 基金分析报告生成: 抽取文档中相关信息，然后文本分析结果，并绘制对应的图表。

   示例:

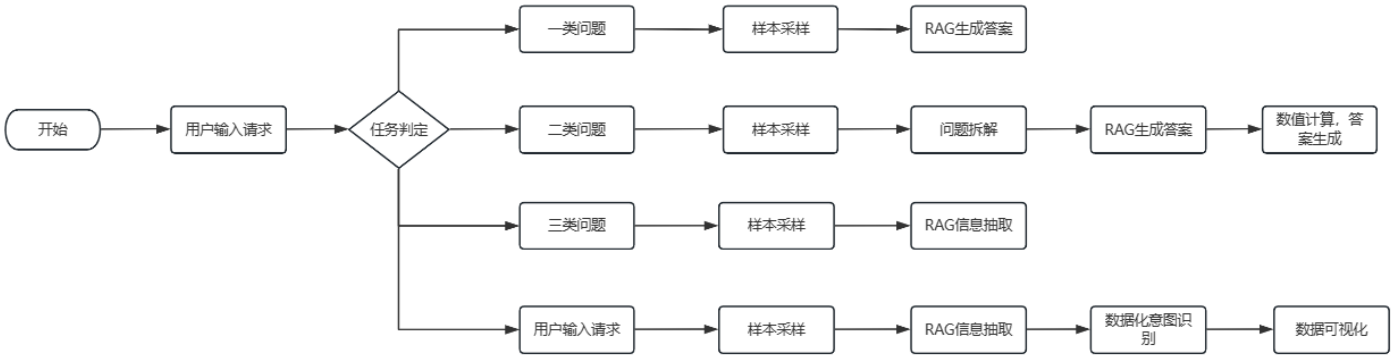
   Q: "请分析2023年报告期末，**基金的基金份额持有人结构信息，并按份额比例绘制饼状图。"

   A: “2023年,****”, 图略。

数据说明:

文件包含本次赛题提供的基金公告文件，共96个pdf文件，涉及16支基金(每支基金4份季报、1份中期报告、1份年度报告)，参赛选手可以对pdf文件进行解析并将文件内容作为构建投研Agent的数据基础。

二. 整体设计



三. 算法说明

3.1 离线部分

需要对few-shot样本(约80个)进行向量化存储，以及对多个文档进行切片向量化存储。用2个不同的索引进行存储，分别用于针对当前问句的few-shot样本采样和RAG问答的建模。

注: 这部分在流程图上没画出来。

3.2 在线部分

整理思路其实就是分类，然后根据不同的子任务分别去做COT拆解，在不同的子任务下，prompt是会有区别的，所以在分类后进行样本的采样，利用few-shot样本提升大模型的推理能力。

3.3 数据结构设计

```
/* 整体接口设计
 不涉及多轮，用最简单的格式定义
 输入直接用String表示query，输出用json格式统一表示。
 注：各个子任务内部的推理输出结果不一定是json格式，数据对齐可以放在外部实现
 API定义如下:
 public Response process(String query);
*/


//返回结果Response 数据定义
//对1-3类子任务，只需要取data中的tts字段即可，对第4类子任务，需要取params中的参数信息
//来完成数据可视化的功能，需要支持哪几种图，需要定义一下。
{
    "status_code": 0,  //结果状态码，0为成果，非零表示异常，竞赛非零状态统一用1就行
    "status_msg": "success" //结果信息说明
    "task": "子任务类型"      //用上述4个类别任务标识
    "data": {
        "tts": "2023年,****”,"
        "params": {
```

```
            "chart_type": "pie chart",
            "x_axis": "key",
            "y_axis": "value",
            "data": {
                "张三": 8.8%,
                "李四": 6.7%,
                "王一": 3.3%
            }
        }
    }
}

//数据样本格式，每个样本定义为json格式，
{
    "query": "2023年**基金在报告期末的可供分配利润比2022年低多少?"
    "category": "task 2",
    "cot": [ // task 1下这里的cot是个[]， 其他场景下有cot内容，具体的文本内容可以利用GPT辅助生成
        "我们可以按下述流程来完成该任务:",
        "首先需要找出2023年和2022年**基金在报告期末的可供分配利润。",
        "根据文档中的信息，可以得到: ",
        "2022年**基金在报告期末的可供分配利润为87亿,",
        "2023年**基金在报告期末的可供分配利润为56亿",
        "所以，我们可以得出结论:"
    ],
    "response":"2023年**基金在报告期末的可供分配利润比2022年低31亿。"
}

//如果是任务3， response 字段用对应的json字符串表示即可.

//如果是任务4， response 用json字符串表示，对应的json格式数据参考Response中的定义
//即:
{
   "tts": "2023年,****”,"
   "params": {
       "chart_type": "pie chart",
       "x_axis": "key",
       "y_axis": "value",
       "data": {
           "张三": 8.8%,
           "李四": 6.7%,
           "王一": 3.3%
        }
   }
}
```

四. 人员分工

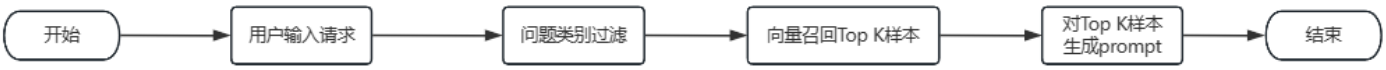
需要同学完成部分工作:

4.1 完成few-shot的样本整理

大概每个类型的子任务，整理10-20个即可，格式参考3.3中的说明，将所有的样本放在一个json文件中，即所有的样本数据构成一个JsonArray的数据。便于程序化导入和few-shot样本的采样。

4.2 完成few-shot样本的采样模块开发

实现一个类，完成few-shot样本的导入，和动态采样模块的编码实现。动态采样模块的算法如下:



该模块的API定义如下:

```
//动态采样模块API
/*
    入参：用户输入的原始query，问题类别信息， 所有候选的样本信息
    出参：利用问题类别一级过滤，然后用语义相似度匹配召回的Top K拼成对应的prompt。
*/
public String sampling(String query, String task, List<Json> samples);

//向量建模的工具可以和RAG中的向量工具保持一只，4.29日确认，有现成的python API工具，无须单独训练模型。
```

4.3 完成数据可视化的模块开发

注: 先确认一下需要支持哪几种图表的类别，如果是折线图，柱状图和饼图，只需要找个python工具，将Response对象中的data["params"]中的数据进行可视化即可，如:

```
 "params": {
       "chart_type": "line chart",
       "x_axis": "key",
       "y_axis": "value",
       "data": {
           "张三": 8.8%,
           "李四": 6.7%,
           "王一": 3.3%
        }
   }
```

需要将data中的key作为x轴，value作为y轴进行绘图即可。