Readme

一、方案简介：本方案以检索增强上下文学习（RAG-ICL）为核心，以Qwen14B-Finance（后文简称为Qwen大模型）为主要工具，进行问题分类与回答。

二、技术路线：
1.数据预处理
1.1. Text 文件：使用正则表达式抽取公司名称。将举办方提供的每个text格式招股书文件分为1000字每段，相邻两段具有200字重叠的片段，使用Qwen大模型的tokenizer进行词频统计备用。
1.2. PDF文件：使用pdfplumber 包进行表格抽取，主要补充了举办方提供的text文件中缺失的表格信息。
1.3. Questions：将1000个Questions中不含有公司名称的问题（即潜在的使用SQL查询进行回答的问题）使用tokenizer统计词频，两两计算余弦相似度后进行谱聚类，类别个数75，可以看到每类问题基本具有相同格式。从每类问题中随机选择2个，构成RAG-ICL的样本库，编写并校对SQL语句和回答。
2.工作流
2.1. 新问题进入，通过20个示例问题-分类结果构成的prompt，利用Qwen大模型分为SQL查询和文本理解两类。
对于SQL查询问题，新问题使用Qwen大模型的tokenizer统计词频，与RAG-ICL样本库中的问题比较，选取最相似的2-4个问题-SQL语句对加入prompt，利用Qwen大模型做“填空与替换”，生成高可解释性与可靠性的SQL查询。运行查询，利用Qwen大模型将查询结果和问题生成为答案。
对于文本理解问题，新问题使用Qwen大模型的tokenizer统计词频，与分段的Text+表格文件计算总词频加权的余弦相似度，选出与问题最相关的20个文本片段，利用他们生成答案。

三、项目亮点：
1.充分利用了Qwen大模型的tokenizer部分包含的金融语境信息，在外来embedding检索效果较差的情况下，不训练新的embedding，达到了较好的检索效果。
2.RAG-ICL为核心的方法使得我们可以在仅标注较少数据集（不到200条），不精调模型的情况下快速得到一个效果较好的问答系统，同时本方案具有较高的可解释性（只需修改RAG-ICL样本库中的例子就可以修改生成效果），对于可见的新问题类型也无需调整模型，只需补充RAG-ICL样本即可。