文档：基于 RAG 的知识库构建与验证过程

本文档详细解释了如何使用代码实现基于检索增强生成（RAG）的知识库构建与验证。主要内容包括系统结构、关键组件及其功能、流程，以及最终的评估方法。

---

1. 系统概述

RAG 系统的目标是将外部知识库与生成式模型相结合，提升问答任务的准确性。通过解析 PDF 文档，构建向量化的知识库，并提供检索功能，RAG 系统能够根据查询返回最相关的知识内容。此外，系统还内置评估工具，以确保知识库的有效性和模型的表现。

---

2. 系统架构

2.1 核心模块

- **PDFParser**：负责从 PDF 文档中提取文本和表格数据。
- **TextVectorizer**：利用嵌入模型将文本数据转化为向量表示。
- **VectorStore**：存储和管理知识库向量，支持基于相似度的检索。
- **RAGEvaluator**：评估系统性能，包括 RAGAs 指标和 GPT 打分。
- **TestCaseGenerator**：基于上下文生成测试案例。

---

3. 流程说明

3.1 知识库构建

1. **解析 PDF 文档**：
    - 使用 `PDFParser` 解析 PDF 文本和表格数据。
    - 将提取的文本存储为 `text_data`，表格数据存储为 `table_data`。
    - 目前未处理公式数据，但可扩展支持。
2. **向量化数据**：
    - 调用 `TextVectorizer` 使用指定的嵌入模型（如 `sentence-transformers/all-MiniLM-L6-v2`）将文本和表格数据转化为向量。
    - 文本数据向量化后存储为 `text_embeddings`，表格数据向量化后存储为 `table_embeddings`。
3. **构建向量存储**：
    - 将所有向量存储到 `VectorStore` 中，构建向量索引。
    - 使用 `VectorStore.save_vector_store` 将索引保存至本地。

3.2 文档检索

1. 接收用户输入的查询（`query`），通过嵌入模型生成查询向量。

2. 利用向量数据库执行近似最近邻搜索，返回与查询向量最接近的文档内容及其相似度得分。
3. 根据检索结果返回文档内容（`retrieved_docs`）。

3.3 测试与验证

测试数据生成：

- 使用 `TestCaseGenerator` 基于样本文本生成测试案例。每个案例包含一个问题（`query`）及其理想答案（`ideal_answer`）。

性能评估：

- **RAGAs 指标**：
  - 调用 `RAGEvaluator` 的 `evaluate_rag_with_ragas` 方法，输入测试案例和系统实例。
  - 输出 RAGAs 指标（如召回率、精确率等）。
- **GPT 评分**：
  - 对每个测试案例，通过系统检索答案（`retrieved_answer`）。
  - 使用 GPT 模型对检索答案与理想答案的相关性进行评分。
  - 记录 GPT 评分结果。

报告生成：

- 汇总 RAGAs 指标和 GPT 评分，生成完整评估报告（HTML 格式）。

---

4. 示例代码分析

4.1 初始化 RAG 系统

```
pdf_path = "examples/power_textbook.pdf"
index_path = "examples/knowledge_base.index"
rag_system = RAGSystem(pdf_path, index_path)
rag_system.build_knowledge_base()
```

上述代码初始化 `RAGSystem`，并调用 `build_knowledge_base` 构建知识库。

4.2 文档检索

```
query = "What is the principle of transformer operation?"
retrieved_docs = rag_system.retrieve(query, top_k=1)
print("Retrieved Docs:", retrieved_docs)
```

系统根据用户输入的查询，检索最相关的文档内容。

4.3 测试集生成与评估

```
testcase_generator = TestCaseGenerator(api_key)
generated_testcases =
testcase_generator.generate_test_cases(sample_context, 5)

evaluator = RAGEvaluator(api_key)
ragas_metrics = evaluator.evaluate_rag_with_ragas(test_cases, rag_system)
```

测试案例由 `TestCaseGenerator` 自动生成，随后由 `RAGEvaluator` 对系统性能进行量化评估。

4.4 报告生成

```
output_file = os.path.join("examples", "rag_evaluation_report.html")
evaluator.generate_full_report(ragas_metrics, gpt_scores, output_file)
```

评估完成后，生成包含详细评估指标的报告。

---

5. 关键点说明

1. **系统扩展性**：
   - 可增加公式解析和处理模块以扩展 `PDFParser` 的功能。
   - 向量化模型可切换为更高精度的预训练模型。
2. **错误处理**：
   - 检查 `text_data` 是否为空，避免检索阶段索引越界。
   - 过滤无效索引，确保返回结果可靠。
3. **评估方法多样性**：
   - 结合 RAGAs 指标和 GPT 评分实现定量与定性评估相结合。

---

4. 示例代码分析

4.1 初始化 RAG 系统

```
pdf_path = "examples/power_textbook.pdf"
index_path = "examples/knowledge_base.index"
rag_system = RAGSystem(pdf_path, index_path)
rag_system.build_knowledge_base()
```

上述代码初始化 `RAGSystem`，并调用 `build_knowledge_base` 构建知识库。

4.2 文档检索

```
query = "What is the principle of transformer operation?"
retrieved_docs = rag_system.retrieve(query, top_k=1)
print("Retrieved Docs:", retrieved_docs)
```

系统根据用户输入的查询，检索最相关的文档内容。

4.3 测试集生成与评估

```
testcase_generator = TestCaseGenerator(api_key)
generated_testcases =
testcase_generator.generate_test_cases(sample_context, 5)

evaluator = RAGEvaluator(api_key)
ragas_metrics = evaluator.evaluate_rag_with_ragas(test_cases, rag_system)
```

测试案例由 `TestCaseGenerator` 自动生成，随后由 `RAGEvaluator` 对系统性能进行量化评估。

4.4 报告生成

```
output_file = os.path.join("examples", "rag_evaluation_report.html")
evaluator.generate_full_report(ragas_metrics, gpt_scores, output_file)
```

评估完成后，生成包含详细评估指标的报告。

---

5. 测试结果

测试一共用到了以下几个询问, 均根据给定提示词生成: Prompt: The grading theory is more of theoretical interest than practical for the following reasons. Capacitance grading is difficult of non-availability of materials with widely varying permittivities and secondly with time the permittivities of the materials may change as a result this may completely change the potential gradient distribution and may even lead to complete rupture of the cable dielectric material at normal working voltage.

Query 1: What is the practical relevance of capacitance grading? Capacitance grading has more theoretical interest than practical relevance due to the difficulty in obtaining materials with widely varying permittivities. Additionally, changes in the permittivities of the materials over time can alter the potential gradient distribution and potentially cause rupture of the cable dielectric material even at normal working voltage. Query 2: Why is capacitance grading challenging? Capacitance grading is challenging because it is difficult to obtain materials with widely varying permittivities, which are needed for effective grading. This makes the implementation of capacitance grading difficult in practice. Query 3: What are the potential consequences of changes in permittivity over time? Changes in permittivity over time can result in alterations to the potential gradient distribution, which can lead to rupture of the cable dielectric material even at normal working voltage. This can have serious implications for the safe and reliable operation of the system. Query 4: What is the impact of non-availability of materials with varying permittivities? The non-

availability of materials with widely varying permittivities makes capacitance grading challenging, as the permittivity of the materials plays a crucial role in determining the potential gradient distribution. This can result in less effective grading, which may compromise the reliability and safety of the system. Query 5: What happens if the cable dielectric material ruptures at normal working voltage? If the cable dielectric material ruptures at normal working voltage, it can lead to a complete failure of the insulation system, which can result in serious consequences, such as fire or electrical shock. This highlights the importance of proper grading and the selection of suitable materials for effective insulation. Would you like me to include the evaluation scores and reasons as well? I could add those under each Q&A pair if you'd find that helpful.
以下是未做任何优化时 RAGAS 系统返回的评分

RAGA Metrics

Metric	Value 1	Value 2	Value 3	Value 4	Value 5
Context Precision	NaN	NaN	NaN	NaN	NaN
Faithfulness	NaN	NaN	NaN	NaN	NaN
Answer Relevancy	0.1844	0.1054	-0.1076	-0.0882	0.1218
Context Recall	NaN	0.0	0.0	1.0	0.0

GROQ Scores

Query 1: {"score": 9, "reason": "The response provides a clear and accurate explanation of the practical relevance of capacitance grading in electrical power systems. The answer highlights the theoretical interest in capacitance grading, but also explains the challenges in obtaining materials with widely varying permittivities. Furthermore, it points out the potential issues caused by changes in the permittivities of the materials over time, which could lead to cable dielectric rupture. The response is well-structured, concise, and informative. However, it could have been improved by providing specific examples or applications of capacitance grading in practical scenarios, which would have helped to further illustrate its relevance."}

Query 2: {"score": 9, "reason": "The response gives a clear and accurate reason for why capacitance grading is challenging. However, it could be improved by providing a more detailed explanation or example of the difficulties in obtaining materials with widely varying permittivities."}

Query 3: {"score": 1, "reason": "The response from the RAG system is not related to the user's question. The provided link does not give any information about the potential consequences of changes in permittivity over time. A relevant and informative answer would have addressed the user's inquiry, discussing the impact of permittivity changes on the potential gradient distribution, and the resulting effects on the cable dielectric material and system operation."}

Query 4: {"score": 2, "reason": "The response from the RAG system does not address the user's question about the impact of non-availability of materials with varying permittivities. Instead, it provides a link to a website. A good response should have addressed the consequences of the absence of such materials, such as challenges in capacitance grading, potential compromise in system reliability, and safety issues, as mentioned in the reference answer."}

Query 5: {"score": 9, "reason": "The response is accurate, detailed, and provides a clear explanation of the consequences of cable dielectric material rupturing at normal working voltage. However, it could be

improved by providing a more direct answer at the beginning, before delving into the explanation."}

在引入了 chunking 和 re-rank 机制后

RAGAs Metrics

Metric	Value 1	Value 2	Value 3	Value 4	Value 5
Answer Relevancy	0.1169	0.0632	0.0809	0.1283	0.1074
Context Precision	0.0	0.0	0.0	NaN	NaN
Faithfulness	1.0	NaN	1.0	1.0	1.0
Context Recall	1.0	0.0	0.0	0.6667	0.0

GROQ Scores

Query 1: {"score": 9, "reason": "The response is quite accurate and complete, correctly identifying the difficulty in obtaining materials with widely varying permittivities and the potential for permittivity changes over time. However, it could provide a more detailed explanation of the consequences of permittivity changes, such as how it could lead to cable dielectric rupture at normal working voltage."}

Query 2: {"score": 6, "reason": "The response is partially correct but lacks detail and examples. The statement mentions two limitations of capacitance grading, but it could benefit from elaborating on the difficulties in obtaining materials with a wide range of permittivities and the potential consequences of changes in permittivity over time. Providing examples or specific scenarios could further strengthen the answer."}

Query 3: {"score": 9, "reason": "The response is correct and provides a reasonable explanation for why capacitance grading is not practical for real-world applications. However, it could have been more comprehensive by mentioning other factors such as the complexity of the manufacturing process and the cost of obtaining materials with the required permittivity values."}

Query 4: {"score": 8, "reason": "The response correctly identifies two factors that make capacitance grading more theoretical than practical. However, it could have provided more detailed explanations for each factor to improve understanding. For example, it could have mentioned specific challenges in obtaining materials and how permittivity changes over time can impact capacitance grading. Additionally, it could have mentioned any potential solutions or workarounds to these challenges."}

Query 5: {"score": 8, "reason": "The response is partially correct, but it could be more detailed and provide more examples or scenarios where capacitance grading in cables can lead to risks. The answer mentions the potential gradient distribution and the rupture of the cable dielectric material, but it could also discuss the impact of environmental factors, such as temperature and humidity, on the cable's performance and longevity. Additionally, the response could have included safety measures or best practices to mitigate these risks."}

关于RAGAs 指标中的 NaN 问题, 官网有以下解释 JSON 解析问题：

- RAGAs 要求模型输出可解析的 JSON 格式，因为其评估框架基于 Pydantic 结构化数据模型。

- 如果模型的输出格式不符合 JSON 标准（如包含多余的引号、非法字符或缺少闭合标记），解析将失败，导致评估结果出现 NaN。 非理想评分场景：
- 某些测试案例可能不适合评分，例如模型的输出为 "I don't know" 或其他与问题无关的回复。
- 在这种情况下，尝试对响应的可信度或相关性进行评分可能会导致异常，RAGAs 会用 NaN 表示无效的评分结果。

6. 关键点说明

1. **系统扩展性**：
    - 可增加公式解析和处理模块以扩展 `PDFParser` 的功能。
    - 向量化模型可切换为更高精度的预训练模型。
2. **错误处理**：
    - 检查 `text_data` 是否为空，避免检索阶段索引越界。
    - 过滤无效索引，确保返回结果可靠。
3. **评估方法多样性**：
    - 结合 RAGAs 指标和 GPT 评分实现定量与定性评估相结合。

---

7. 总结

本代码通过解析 PDF 文档构建知识库，并结合 RAG 检索技术实现高效问答。其核心在于向量化与检索机制的结合，同时通过 RAGAs 和 GPT 双重评估保证系统性能和可靠性。

如需进一步优化，可探索以下方向：

- 提升文档解析能力（如公式解析）， 已找到一个 pix2tex 的库支持将图片转化为 latex, 但代码本身有 bug
- 增加测试批量和次数
- 使用更先进的嵌入模型提升检索效果。
- 引入领域知识增强模型性能。
- FlashRank 请求 LLM 的次数过于频繁，极易触发too many requests, 考虑加个延时

8. 参考来源

参考来源

1. **RAG系统的7个检索指标：信息检索任务准确性评估指南**
   https://53ai.com/news/RAG/2024091193265.html
2. **Advanced RAG 07：在RAG系统中进行表格数据处理的新思路**
   https://blog.csdn.net/Baihai_IDP/article/details/138898779
3. **Langchain**
   https://python.langchain.com/

4. **Chat with your PDF: Using Langchain, F.A.I.S.S., and OpenAI to Query PDFs**
https://medium.com/@johnthuo/chat-with-your-pdf-using-langchain-f-a-i-s-s-and-openai-to-query-pdfs-e7bfde086155

5. **破解PDF解析难题：RAG中高效解析复杂PDF的最佳选择**
https://www.53ai.com/news/RAG/2024111591723.html