迭代一任务说明

近年来大模型被广泛应用到各个领域，成为人们生活中必不可少的工具。业界针对大模型已探索出多种应用形式，如RAG（Retrieval Augmented Generation, 检索增强生成），Agent工具链等等。然而，尽管大模型在各种任务中展现出了强大的能力，其实际效果和表现却往往存在很大的不确定性。对自己的大模型应用的各个组件进行多维度的评估能够帮助我们更好地理解应用的实现完成度与不足，从而为后续的改进指明思路和方向。

2025年软工三课程分为三个前后依托的迭代阶段，以开源项目Ragas为切入点，系统学习大模型应用评估知识，最终搭建一个可靠的RAG应用以及配套的评估平台。迭代一的具体任务如下：

1. 阅读，标注开源软件，完成开源项目分析文档

1.1 开源项目 Ragas

项目介绍： Ragas是一个用于评估大模型应用的开源项目，特色是评估RAG系统，也支持评估大模型Agent，或者用于评估简单的大模型任务。Ragas库中内置了各种的评估指标，比如上下文精准度、上下文召回率、可信度、回答相关性、回答正确性等等。这些指标按照计算过程中是否使用大模型可以分为大模型驱动的指标和非大模型驱动的指标；按照评估场景可以分为RAG指标、Agent指标、自然语言比对指标、自然语言转SQL任务指标、通用指标等等。

仓库链接： explodinggradients/ragas: Supercharge Your LLM Application Evaluations 🚀

文档地址： Ragas-Doc

基本信息： Python项目 10957行代码

1.2 要求

阅读Ragas源码以及官方文档，完成开源项目分析文档：

1. 开源项目分析文档参考：*2025软工三-开源项目分析文档示例-开源软件泛读、标注和维护报告文档*
2. 不要求和参考文档内容完全一致，但至少包括：功能描述、开源软件的软件架构及各个包和类的作用、软件功能与类间的对应关系、阅读收获
3. 文档以可读性、有用性、可度量为基本前提

2. 复现指标，完成评估实验

2.1 数据集介绍

translation.csv：

某个大模型针对中译英任务获得的结果，其中`source_text`是中文原文，`ground_truth`是人工提供的高质量参考翻译，`llm_rsp`是大模型的翻译结果。

news_summary.csv：

某个大模型针对新闻总结任务获得的结果，其中`news_content`是新闻原文，`summary`是大模型对新闻的总结，`label`是新闻的标签。

2.2 要求

针对课程提供的两个数据集，自己实现若干指标（不是直接调Ragas库，至少实现一种大模型驱动的指标）对其进行评估，提交代码与评估报告，评估报告内容包括但不限于：指标选取原因，指标实现过程，评估结果分析；