

基于GraphRAG的公司信息收集

项目目标：本项目旨在通过构建基于Neo4j的知识图谱，开发一个可扩展的公司信息收集系统。项目将利用Cypher进行知识图谱查询，并结合大语言模型(LLM)增强数据提取与分析能力。该系统将有效组织和检索与公司相关的数据，助力商业决策。

模块概述：

1. 数据输入与预处理：
 - 对公司原始数据进行预处理，以适应图数据库的结构化要求。
 - 搭建Neo4j环境并进行数据导入。
2. 知识图谱构建：
 - 定义实体(如公司、行业、人物、产品)的图谱架构。
 - 建立实体之间的关系。
 - 将数据插入Neo4j。
3. Cypher图谱查询：
 - 编写从基础到高级的Cypher查询以提取公司信息。
 - 查询实体关系、发现模式、检索公司结构化信息。
 - 通过实际例子展示复杂查询如何获得公司之间的关联信息。
4. **RAG与LLM集成**：
 - 介绍GraphRAG(结合知识图谱和RAG进行动态信息检索)。
 - 利用LLM查询知识图谱，提升信息提取的准确性。
5. 项目案例：基于公司信息收集的商业决策支持：
 - 通过收集并分析目标公司的信息，解决实际商业问题。
 - 展示知识图谱和LLM生成的洞察如何为商业决策提供支持。

关键技术：

- 图数据库概念与Neo4j操作。
- Cypher图数据库查询。
- LLM集成提升数据检索效果。
- 知识图谱可视化技术。
- 数据分析与总结展示的技能。

前置知识：

- 具备基础的Python编程能力。
- 了解数据库概念。
- 对LLM有初步了解。

工具与软件：

- **Neo4j**(用于构建知识图谱与查询)
- **Python**(Numpy, Pandas, Neo4j驱动, 可视化库)
- **Cypher**(Neo4j的查询语言)
- **OpenAI API**(用于调用大语言模型, 注意:会产生token usage费用, 需自己承担)

- **Langchain** (大模型框架)
- **GitHub** (用于版本控制与代码管理)

项目成果：

- 完整的Neo4j知识图谱。
- 一套用于检索公司信息的Cypher查询脚本。
- 与LLM集成的数据提取与处理Python脚本。
- 总结知识图谱和LLM分析结果的展示文档(PPT)。

学习收获：

- 掌握从零构建知识图谱的完整流程。
- 精通Cypher查询以检索复杂数据模式。
- 学会将LLM与GraphRAG集成, 提升信息检索效率。
- 提升数据可视化与商业决策展示技巧。
- 获得项目管理、编程及数据科学相关实践经验。