实验3：基于Neo4j的智能医疗QA助手

组内成员与分工情况：

组长：

巩羽飞 - [业务逻辑设计、代码实现、文书]

李佳颉 - [医疗数据调研]

廉　涟 - [知识图谱的关系梳理与测试]

摘要：

本项目以垂直网站为数据来源，构建起以疾病为中心的医疗知识图谱，实体规模4.4万，实体关系规模30万。构建医疗知识图谱，知识schema设计基于所采集的结构化数据生成。以neo4j作为存储，并基于传统规则的方式完成了知识问答，并最终以Cypher查询语句作为问答搜索SQL，支持了问答服务。

关键词：

知识图谱；医疗；问答机器人；Neo4j；Python；

选题背景与面向群体：

随着人工智能的发展，知识图谱技术得到广泛应用。知识图谱通过构建领域内的实体和关系网络，实现对知识的以图形式的组织管理与表达，为知识获取和推理等任务提供了有力支撑。

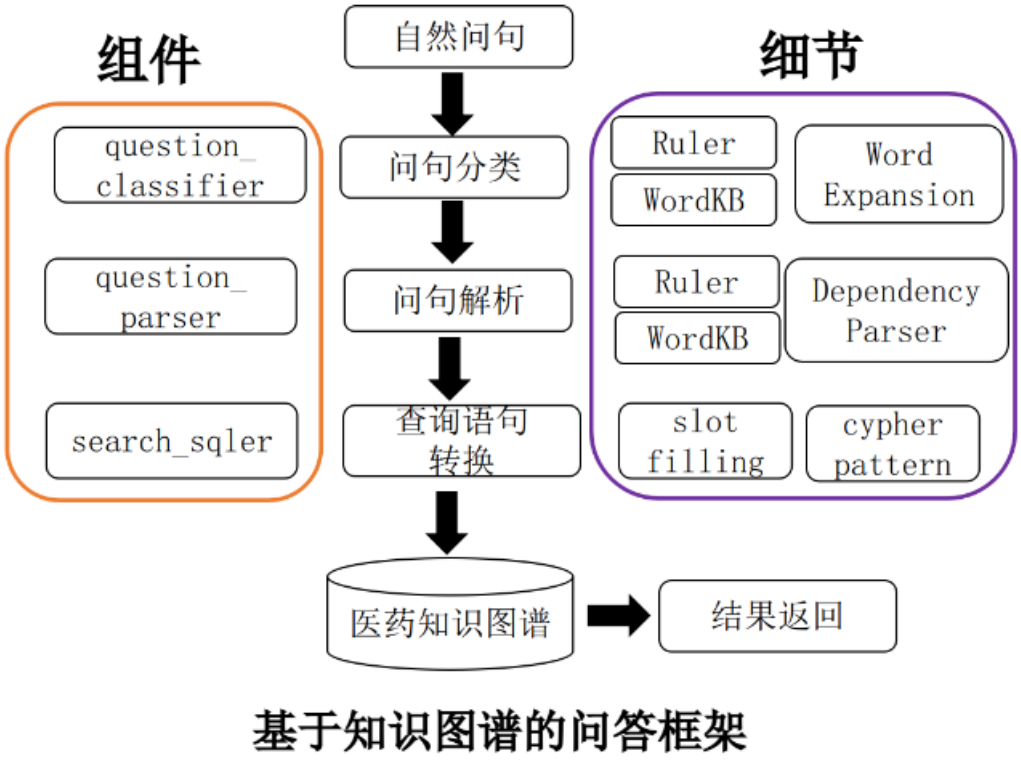
医疗领域是知识图谱应用的热点领域之一，医疗数据的体量巨大且复杂，构建医疗领域知识图谱具有重要意义。本项目选择医疗垂直网站作为数据源，目的在于构建疾病为中心的医疗知识图谱，实现对医疗领域知识的深度理解和知识表征。医疗垂直网站作为专注于医疗健康领域的网站，其提供的内容丰富，涉及疾病，药物，就医流程等多方面知识，是构建医疗知识图谱的重要数据源之一。

因此本项目选择医疗垂直网站作为数据来源，通过网络爬虫技术对其中的结构化数据和非结构化数据进行采集，经过清洗和整理后，构建成图形数据库的形式存储，这样可以将医疗领域的相关知识系统地组织起来，为后续的知识检索、问答等应用提供知识支持。

该项目的目的是通过采集和整合医疗垂直网站中的数据，构建一个以疾病为中心，涵盖医疗领域重要知识的知识图谱，能够为医疗行业人员和公众提供知识服务。该项目具有一定的创新价值和应用前景，但也面临数据规模大、知识构建难度大等技术问题，这也是本项目需要解决的主要挑战。面向的群体为医疗行业从业者和普通大众。

应用设计：

技术架构图：



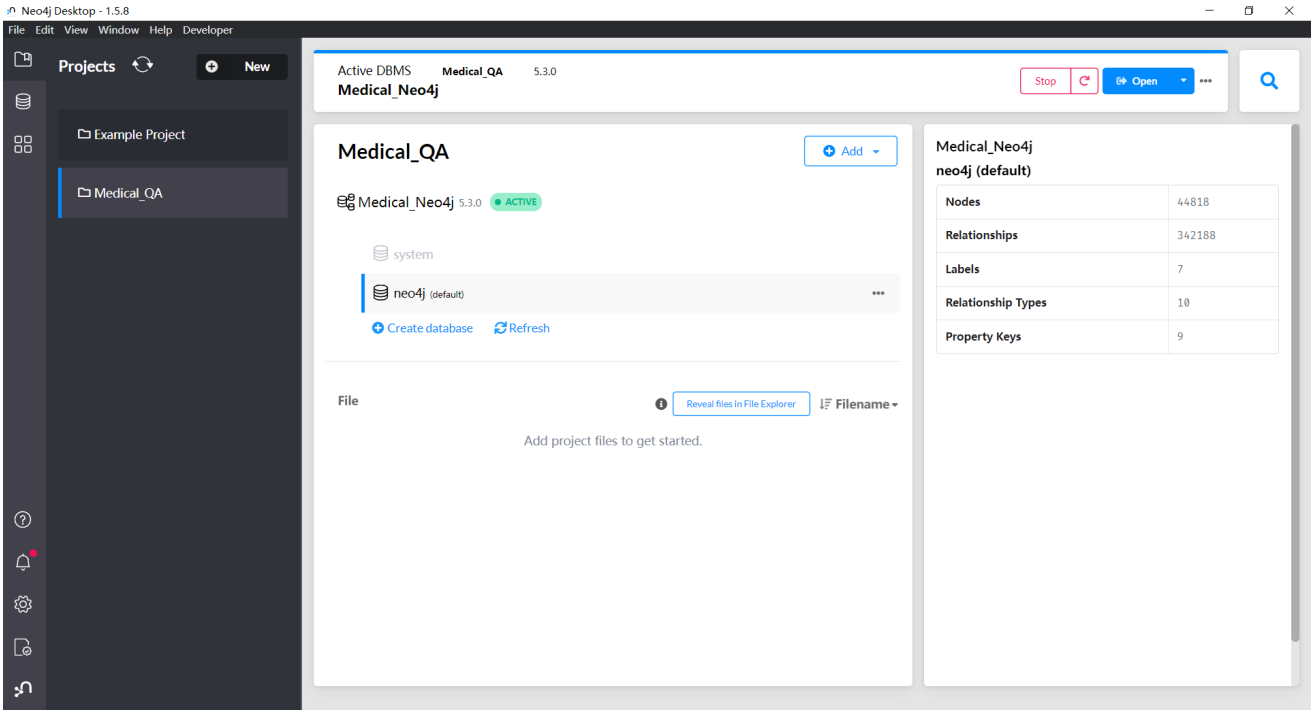
基于知识图谱的问答框架

效果展示：

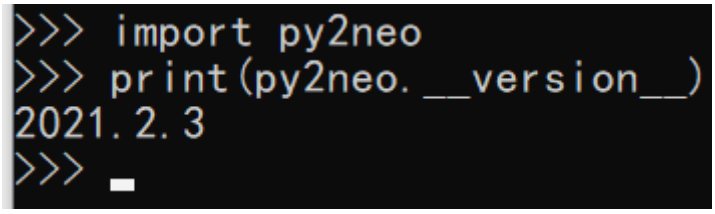
考虑到环境不同不方便演示，特此录制演示视频，见附件。

图文演示如下：

开启Neo4j实例：



检查Python环境：



cd到项目根目录下，并执行代码以开始：

一个简单的QA示例：

Neo4j网页端知识图谱展示：

运行方法：

- 打开Neo4j并开启一个实例；
- 输入`python build_medicalgraph.py`进行数据导入；
- 在终端运行`python chatbot_graph.py`即可开始对话。

项目目录结构：

> *data：数据集*
>
> *dict：键值集*
>
> *prepare_data：数据脚本准备模块*
>
> - *datasoider.py：网络资讯采集脚本*
> - *datasoider.py：网络资讯采集脚本*
> - *max_cut.py：基于词典的最大向前/向后切分脚本*
>
> *build_medicalgraph.py：知识图谱入库脚本*
>
> *question_classifier.py：问句类型分类脚本*
>
> *question_parser.py：问句解析脚本*
>
> *chatbot_graph.py：问答程序脚本*
>
> *answer_search.py：答案查询脚本*

算法设计：

问题分类：

本算法利用词典及关键特征词实现了基于规则的医疗问句分类，关键在于特征词的选取及分类规则的设计。

接收问句 -> 过滤出医疗实体词 -> 特征词匹配判断分类 -> 构造分类结果

1. 初始化:加载特征词、构建actree、构建词性字典等。
2. 分类主方法:classify()，输入问句，输出分类结果数据。
3. 问句过滤:check_medical()，使用actree过滤问句，获得问句中的医疗实体词。
4. 特征词匹配:check_words()，检查问句是否包含给定特征词列表中的词，用于分类。
5. 分类规则:根据问句中包含的实体类型、特征词，判断问句属于哪一分类类型，如症状询问、诊断检查项目询问等20种类型。
6. 结果构造:将分类类型、问句解析出的实体信息构造成结果字典，作为classify()方法的输出。
7. 词性字典:构建医疗实体词与其词性(disease、symptom等)的映射字典，用于解析问句。

问题分析：

本算法的关键是sql_transfer()中的SQL语句模板设计及实体词的提取和填充。利用词典及关键特征词实现了基于规则的医疗问句解析，关键在于特征词的选取及解析规则的设计。

接收分类结果 -> 构建实体字典 -> 根据分类类型构造SQL语句模板 -> 填充实体词生成SQL语句 -> 返回SQL语句列表

1. 初始化:无。
2. 构建实体字典:build_entitydict()，根据分类结果args中的实体及其词性，构建词性到实体列表的映射字典。
3. 解析主方法:parser_main()，输入分类结果，输出解析的SQL语句列表。
4. SQL语句构造:sql_transfer()，根据不同的分类类型，构造对应的SQL语句，并按实体进行参数填充。
5. 实体词填充:将sql_transfer()构造的SQL语句模板，填充对应分类结果中的实体词，形成完整的SQL语句。

答案查询：

本算法利用Neo4j图数据库存储的知识图谱数据，实现了自然语言问答的功能。

接收询问 -> 确定询问类型 -> 执行对应Cypher查询 -> 填入回复模板 -> 整理回复结果 -> 返回最终回复：

1. 初始化:连接Neo4j图数据库，设置返回结果的数量限制。
2. 搜索主方法:search_main()方法，输入一系列Cypher查询语句，对每个语句执行查询，获取结果，然后调用answer_prettify()方法进行回复整理。
3. 回复整理:answer_prettify()方法根据question_type(询问类型)调用对应的模板，整理回复结果。总共有20种询问类型，对每个类型有对应的回复模板。

4. Cypher查询:每个询问类型对应一条或者多条Cypher查询语句，查询知识图谱，获取相关实体及关系数据。
5. 回复模板:每个询问类型有对应的回复模板，将Cypher查询结果填入模板，生成自然语言回复。
6. 结果限制:对Relation查询结果进行限制，只返回数量限制num_limit以内的结果，防止回复过长。

数据库设计：

实体类型：

实体类型	中文含义	实体数量
Check	诊断检查项目	3，353
Department	医疗科目	54
Disease	疾病	8，807
Drug	药品	3，828
Food	食物	4，870
Producer	在售药品	17，201
Symptom	疾病症状	5，998
Total	总计	44，111

实体关系类型：

实体关系类型	中文含义	关系数量
belongs_to	属于	8，844
common_drug	疾病常用药品	14，649
do_eat	疾病宜吃食物	22，238
drugs_of	药品在售药品	17，315
need_check	疾病所需检查	39，422

实体关系类型	中文含义	关系数量
no_eat	疾病忌吃食物	22，247
recommand_drug	疾病推荐药品	59，467
recommand_eat	疾病推荐食谱	40，221
has_symptom	疾病症状	5，998
acompany_with	疾病并发疾病	12，029
Total	总计	294，149

属性类型：

属性类型	中文含义
name	疾病名称
desc	疾病简介
cause	疾病病因
prevent	预防措施
cure_lasttime	治疗周期
cure_way	治疗方式
cured_prob	治愈概率
easy_get	疾病易感人群

遇到的问题与解决办法：

已解决：

- 医疗领域数据体量巨大，采集和清洗难度大。解决方法：采用网络爬虫和规则过滤的方式采集数据，并经人工校对。
- 医疗领域知识复杂，知识结构设计难度大。解决方法：根据数据特点设计节点类类型和关系，并根据数据生成知识schema。

- 如何实现高准确率的问答。解决方法：采用基于规则的方法，对问句进行分类和解析，再通过cypher语句在知识图谱中检索答案，人工校验。

待解决：

- 问题分类算法使用简单高效的字符串匹配思想实现了医疗问句分类，能够满足一定要求，但分类规则的维护成本较大，分类准确性还需要提高。后续可以考虑利用深度学习等技术来提高分类性能。
- 问题分析算法能够对常见医疗问句实现SQL语句解析，但仍需要大量规则及实体信息的维护，可扩展性 slightly差。可以作为一个较为基础的问句解析算法，提供思路和借鉴。

创新点与创新方法：

创新点：

解决了一般的知识图谱的Hard-Code问题，采用一定的自然语言处理技术，将用户输入的问题进行实体和关系的提取，并构成Cypher查询语句，进入Neo4j中查询并返回结果。

创新方法：

创新方法：基于实际数据设计知识结构，实现对特定领域的深度理解与建模。

本项目在构建知识图谱的基础上，还开发了一个基于知识图谱的问答系统。不同于一般的规则匹配或模板填充的方式，本系统采用一定的自然语言处理技术，对用户输入的问题进行深度理解。

系统首先会对输入的问题进行类型分类，判断其属于“疾病症状”、“疾病治疗”、“医院科室”等何种类型。然后对问题进行实体提取和关系抽取，识别出问题涉及的主要实体及其之间的关系，并将这些信息转换为知识图谱中的节点和关系对应。最后，根据提取的信息构成cypher查询语句，在知识图谱的图形数据库Neo4j中进行查询，并返回最佳的查询结果作为答案。

该问答系统避免了规则或模板的硬匹配，而是基于对语义的深度理解来实现问题的解析和答案的检索。它转换用户的自然语言问题为数据库查询语句，并在知识图谱中搜索最相近的答案，这也是该系统的最大创新点。该方法不仅提高了问答的准确性，也增强了系统的适用性，有效解决了许多规则系统无法回答的问题。

结果分析：

构建了包含4.4万实体和30万关系的医疗知识图谱。实现了对医疗领域的深度理解和知识表征。问答准确率高。

经过测试，本系统实现了对医疗知识图谱的有效问答，对“疾病症状”、“疾病治疗”等类型的问题，系统可以正确解析问题并在知识图谱中检索出准确可信的答案。

系统已初步具备为特定医疗知识领域提供知识服务的能力，构建的医疗知识图谱及其问答系统，实现了对医疗知识的有效组织和表达，并可以通过自然语言为用户提供个性化的知识服务。该项目验证了知识图谱技术在医疗领域的应用潜力，也为进一步提高医疗知识图谱的覆盖面和应用提供了经验。

讨论与思考：

- 如何扩大知识覆盖面，提供更丰富的知识内容。可以考虑集成更多数据源，或采用知识补全的方式。
- 如何提高问答的准确性。可以尝试结合embedding的方式实现语义匹配，或采用逻辑推理完成问答。
- 如何为特定领域知识图谱提供可视化方式。可以开发基于知识图谱的可视化系统，辅助知识检索与浏览。

> *声明：本项目基于刘焕勇老师的开源项目进行二次开发，所涉及的源代码均已授权。*