基于豆瓣电影知识图谱的问答系统源码解析

基于豆瓣电影top250的知识图谱问答系统，因为其他原因，只设置了小部分数据来测试结果，主要实体是 ：电影名字，电影评分，电影排名，电影链接；关系有：电影名字的另种叫法，电影排名关系，电影链接对应关系，电影评分关系。

现在项目地址为：https://github.com/wuhq1/DoubanKG

问答结果展示：



收获：

熟悉一个基本知识图谱问答系统的构建框架，1、从豆瓣上爬取所需的网页数据，对半结构化数据进行数据清洗；2、将处理后的数据导入到neo4j的服务器中；3、构建问答框架：对输入的问题抽取实体和关系、进而对关系进行分类，对分类后的数据进行解析、匹配到对应的查询cypher语句， 最后搜索输出答案。

遇到的问题：

1、neo4j服务器自身存在bug，一旦误导入了数据，数据的标签不能修改，只能卸载后换个版本再下载。

2、在数据清洗中会有噪点数据，自身在数据分析方面能力有待提高，只能人工进行剔除。

3、问题不能单步回答，通过设置全局变量实现多步问答。

对于应用到农业水稻上的思考：

1. 对水稻进行知识图谱知识设计，语义类型设计、语义关系设计，本体对象设计等。（key point）
2. 存入知识库中的数据来源？一是结构化的数据 二是从官方下载文本，利用性能较好的NLP模型接口进行实体关系抽取，是不是还可以根据垂直领域中文字体的独特性，另外设计一种实体识别模型？（之前看过这类文献）
3. 知识图谱的构建，一般都使用neo4j来构建（其他protégé等教程很少，不好上手）
4. 问答系统的框架，要加强问题的灵活性，对问题中的实体进行相似度匹配。（还要去研究一下问答系统的经典框架）

下一步计划：

对水稻进行知识图谱的知识设计。多关注并学习以上谈及的方向。