## Recipe QAKB 基于知识图谱的菜谱问答系统

张佳雯 2018E8015061003

本项目主要参考了刘焕勇老师基于医药领域知识图谱的问答系统的工作（<https://github.com/liuhuanyong/QASystemOnMedicalKG>），在下厨房爬取的数据的基础上重新在neo4j上构建了一个新的知识图谱，并且实现了一个简单的自动问答系统。

#### 项目最终效果

以下是实际问答运行过程中的截图：





#### 性能说明

由于本项目的数据规模十分小，且应用领域非常明确（针对菜谱进行问答），在图谱和问答系统完成后我们对该系统只进行了简单的测试，其回答的准确率大约在65%左右。该系统无法回答的问题举例：我想吃火锅；水煮鱼怎么做；有甜口味的菜吗；有没有胡萝卜的素菜吗；黑椒牛柳可以用白胡椒吗 等。

#### 项目运行方式

1、配置要求：要求配置neo4j数据库及相应的python依赖包。neo4j数据库用户名密码记住，并修改相应文件（build\_recipegraph.py和answer\_search.py）

2、知识图谱数据导入：python build\_recipegraph.py

3、启动问答：python chatbot\_graph.py

#### 脚本目录

xml2json.py：将爬取到的xml文件进行数据清洗并转为json格式

build\_recipegraph.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

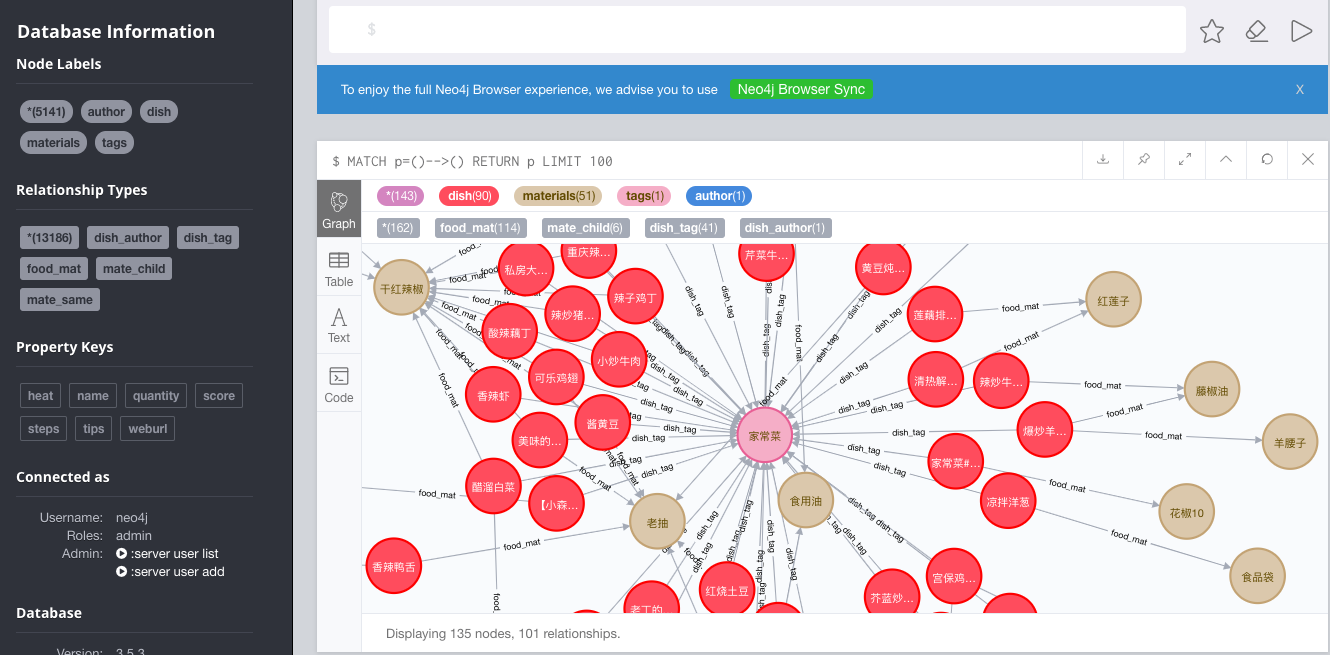
answer\_search.py：从图数据库中搜索相关实体

chatbot\_graph.py：问答程序脚本

#### 知识图谱的构建

数据来源：

从下厨房（<http://www.xiachufang.com/>）的‘家常菜’、‘快手菜’等类目中一共爬取1127个菜谱，经过数据筛选、去重、清洗过后，共留下1124条有效菜谱



知识图谱实体类型:

|  |  |  |  |
| --- | --- | --- | --- |
| 实体类型 | 中文含义 | 实体数量 | 举例 |
| dish | 菜名 | 1124 | 可乐鸡翅；家常豆腐 |
| materials | 食材 | 1823 | 土豆；鸡胸肉 |
| author | 作者 | 640 | 琉璃爱美食；老丁的私房菜 |
| tags | 标签 | 1554 | 冰箱；减肥 |
| total | 总计 | 5141 | 约五千个实体 |

知识图谱实体关系类型:

|  |  |  |  |
| --- | --- | --- | --- |
| 实体关系类型 | 中文含义 | 关系数量 | 举例 |
| food\_mat | 所需食材 | 7044 | <家常豆腐，需要，豆腐> |
| dish\_author | 菜品作者 | 1124 | <老丁的私房菜，做了，咕咾肉> |
| dish\_tag | 菜的标签 | 4743 | <家常豆腐，标签，素菜> |
| mate\_same | 同种食材 | 41 | <土豆，同义于，马铃薯> |
| mate\_child | 下属食材 | 234 | <鱼，包含了，鲫鱼> |
| total | 总计 | 13186 | 约一万三千条关系 |

知识图谱属性类型：

|  |  |  |
| --- | --- | --- |
| 属性类型 | 中文含义 | 举例 |
| name | 菜品名称 | 可乐鸡翅 |
| score | 菜品打分 | 8.6 |
| heat | 菜品热度 | 6741 |
| steps | 制作步骤 | 将胡萝卜洗干净，切段。苹果洗净后先将胡萝卜榨汁再将苹果榨汁，去渣，混合，搅拌。 |
| tips | 小贴士 | 建议使用的酱油是金标生抽 |
| weburl | 网页链接 | http://www.xiachufang.com/recipe/100451099/ |
| qiantity | 食材数量 | 50克 |

#### 自动问答

在自动问答系统中我们先将问句进行分词，然后通过关键词匹配进行问句分类、问句解析，最后转换成对应的cypher语句在neo4j中进行查询。

支持问答类型：

|  |  |  |
| --- | --- | --- |
| 问句类型 | 中文含义 | 问句举例 |
| dish\_desc | 菜名描述 | 可乐鸡翅怎么做 |
| mate\_dish | 食材描述 | 鸡肉可以做什么菜？ |
| no\_mate\_dish | 否定食材 | 没有胡萝卜的菜有哪些？ |
| tag\_dish | 标签描述 | 辣味的菜有哪些？ |
| no\_tag\_dish | 否定标签 | 不辣的菜有哪些 |
| cate\_desc | 种类表述 | 鱼可以做哪些菜 |
| dish\_mate | 罗列食材 | 做可乐鸡翅需要哪些材料？ |
| func\_dish | 食物功效 | 减肥吃什么好？ |
| author\_dish | 作者搜索 | 君之做过哪些菜？ |

#### 总结

1. 本次作业仅完成了一个简单且小规模的图谱，并且有很多前人的工作可以参考，但是由于是第一次自己从数据爬取开始搭建，也没有使用图数据库的经验，花费了大量的时间
2. 由于本次爬取的数据量比较小，很多菜谱没有包含在内，因此可能像水煮鱼之类的热门菜没有包含在内，无法获得相应结果，若是扩大数据量会好得多
3. 在本项目中菜的口味是根据菜谱网站上的标签而检索的，如果一道菜被打上了‘辣’的标签，那么它就可以通过关键词检索到，搜索‘家常菜’等菜的种类同理。
4. 对于‘番茄’和‘西红柿’这样的同义食材以及‘鱼’和‘鲫鱼’这样的包含性食材，我手动制作了一份同义食材和包含性食材的文档，并在图谱中将它们相关联，这样通过输入‘番茄’也可以获得含有‘西红柿’关键词的菜品，输入‘鱼’就可以获得鱼类目下所有种类的鱼制作的菜。
5. 由于应用场景非常简单，因此本次作业没有使用由问句自动生成表达式等技术，而是基于传统规则。