一、项目简介

众所周知，民以食为天。因此，本组设计并实现了菜谱领域的知识图谱，经过数据采集、清洗、存储构造知识图谱，并基于知识图谱实现了可视化与问答系统。

小组成员及分工如下：

姓名	学号	分工
施方丹	22021064	数据存储与可视化
邱志林	22021061	数据采集与整理
贺婷婷	3170104341	查询设计，模板匹配
杨佳妮	3170103462	查询设计，问句解析

二、技术实现

2.1 图谱结构设计

本知识图谱包含两类实体（菜谱、食材），两种实体间关系（主食材、辅料）。

- 菜谱实体的属性包含菜谱名称、类型、耗时、口味、工艺、做法；
- 食材实体的属性包含食材名称、分类、简介、营养价值、食用功效。
- 主食材/辅料关系的属性包含用量。

样例如下：

- 菜谱：

```
{
"identity": 1,
"labels": [
 "recipe"
```

```
],
"properties": {
"耗时": "廿分钟",
"做法": "1:土豆去皮。2:土豆切丝。3:浸水。4:烧肉切片。5:茼蒿切段。6:锅子烧热放入烧肉煸出油。7:放入土豆丝。8:土豆丝炒熟后放入茼蒿段。9:调味即可享受。10:开动啦。",
"name": "土豆丝炒烧肉",
"口味": "咸鲜",
"工艺": "炒",
"类型": "热菜"
}
}
```


- 食材:


```
{
"identity": 12766,
"labels": [
 "material"
],
"properties": {
"name": "香草",
"食材简介": "香草，正规的叫法为芳香植物，是具有药用植物和香料植物共有属性的植物类群，全世界有3000多种，而薰衣草、迷迭香、百里香、藿香、香茅、薄荷、九层塔等为著名的品种。通常也有调味、制作香料或萃取精油等功用，其中很多也具备药用价值。虽然一般所谓的香草主要是指取自绿色植物的叶的部份，但包括花、果实、种子、树皮、根等，植物的各个部位都有可能入药。",
"食用功效": "具有安神 、催眠 、提神 、宁神 、止咳嗽 、美容 、保健的作用。",
"营养价值": "香草通常有调味、制作香料或萃取精油等功用，其中很多也具备药用价值。",
"类别": "调味品"
}
}
```


- 边:


```
{
"start": {
"identity": 25596,
...
},
"end": {
"identity": 21,
...
},
"segments": [
 {
   "start": {
"identity": 25596,
...
   },
   "relationship": {
"identity": 99920,
"start": 25596,
"end": 21,
"type": "主食材",
"properties": {
"用量": "500克"
     }
```

```
    },
    "end": {
"identity": 21,
...
    }
  }
],
"length": 1.0
}
```


2.2 数据采集

本组经过调研，选择了美食天下作为数据来源。

在数据采集方面，我们使用Python的urllib3库来获取网页的html文件，并使用BeautifulSoup(bs)库来对DOM进行解析并提取出我们需要的数据。具体方式如下：

```
http = urllib3.PoolManager() #开启一个连接池
response = http.request('GET',"https://home.meishichina.com/recipe-type.html") #按网址访问对应网页
soup = bs(response.data.decode(), 'lxml') #利用bs将获取的数据进行整理
```

在获取soup之后可以利用`item=soup.find('div')`或`item=soup.find_all('li')`之类的形式去找到相应的项，并用`name=item.string`把相应的文字取出。比如说对耗时/口味/工艺信息的抽取就如下所示：

```
#把信息表里的耗时/口味/工艺信息抽取出来
box = soup.find(name = 'div', attrs = {'class' : 'recipeCategory_sub_R mt30 clear'}).find_all('li')
    info = {}
    for it in box:
        spans = it.find_all('span')
        info[spans[1].string] = spans[0].a.string
cook_time = info['耗时'] if '耗时' in info else ''
flavor = info['口味'] if '口味' in info else ''
method = info['工艺'] if '工艺' in info else ''
```

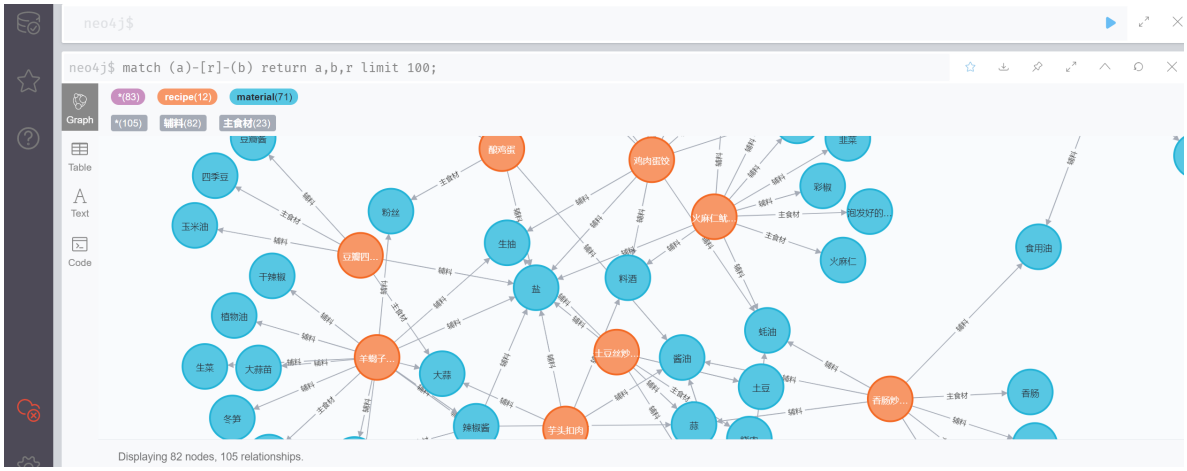
用Python的dict对菜肴的各项信息进行分类存储，并最终存在json文件中供后续可视化和QA的使用。

2.3 数据可视化

本项目实现了两种不同的可视化方式。

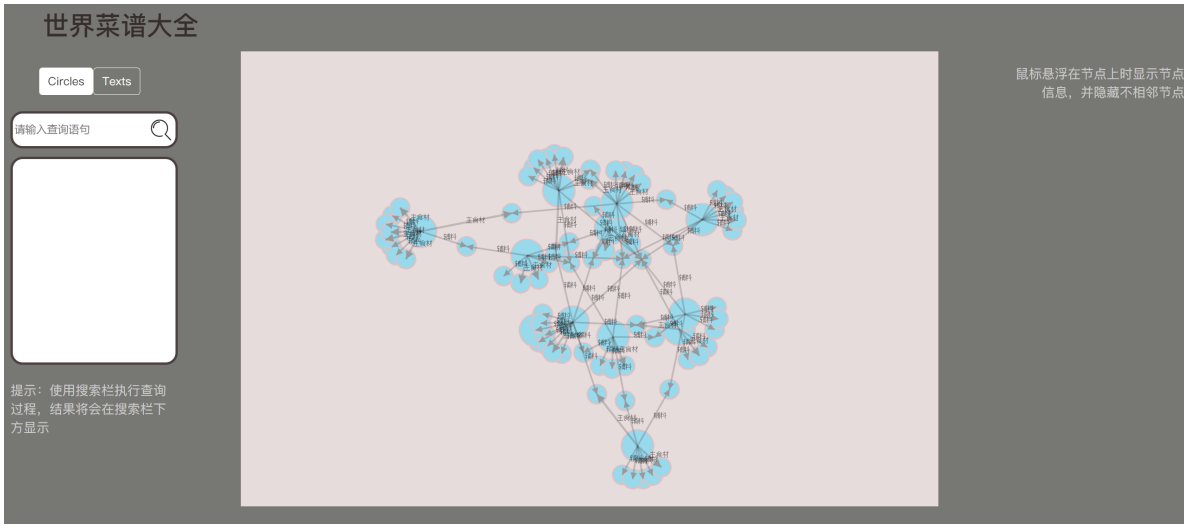
基于`Neo4j`

在项目中我们选择的底层存储数据库为常用的图数据库`Neo4j`。而它自带了图的可视化功能，可直接访问我们的数据库页面访问我们的图谱。



手动实现网页

为了方便与美观，我们还手动设计了相应的图谱可视化网页，并搭建了对应的后端以支持可视化模块。主要使用了`JavaScript`中的`D3`库来实现图谱的绘制，并添加相应的动作以提升图谱可视化效果。



由于图谱过大，完全显示在页面会显得非常凌乱，因此这里只选择了完整图谱的一个子图进行显示。

2.4 KBQA

基于知识图谱的问答系统需要根据用户输入的自然语言问句，返回知识图谱中的相关信息。

尽管课上介绍了众多有趣的方法，出于时间的考虑本组仍选择了基于模板的方法实现系统，该部分由问答设计、问句解析、模板匹配两部分组成。

问答设计

经过讨论，我们认为以下四种问题有较高的出现可能性：

问题类型	说明	举例	回答
property_query	查询实体的属性	烤肠怎么做/烤肠的做法	1:准备烤肠。2:高火，五分钟。3:切口，中低火再来烤2分钟。
property_constraint	查询符合约束的实体	咸香的炒菜有哪些	蚝油四季豆、爆炒海螺片...
relationship_query	查询实体间的关系	红烧肉要用多少五花肉?	一块
relationship_constraint	查询满足关系的实体	红烧肉的主食材是?	五花肉
unknown	无法解析出相应问句模板	我好吃吗	听不懂欸（/´д`）

> 对于未能解析出相应问句模板的问句，回答"听不懂"；对于查询无结果的问句，回答"我不知道"。

问句解析

问句解析部分通过分析用户的输入，得到问题的类型以及相应的参数，结构设计为：

```
question_type:"property_query"
args:{
    nodes:["烤肠":"material","香菇鸡":"recipe"]
    constraints:["口味":"酱香", "工艺":"炒"]]
    properties:["做法","耗时"]
    relationships:["主食材"]
}
```

解析过程如下：

1. 匹配关键词：由于领域内有较多专有名词，传统分词工具分词效果较差，因此使用匹配外部字典的方式检测关键词，并存储至相应参数。
2. 解析疑问词：集思广益，记录常用疑问词及相应语义，如"怎么做"->"做法"，"用多久"->"耗时"。
3. 根据疑问词及关键词匹配问题类型，并传递相应参数。

模板匹配

由于本组使用了Neo4j图数据库存储知识图谱，因此使用了Cypher Query Language进行数据查询。

根据上一步中得到的问题类型及问题参数翻译成相应的CQL查询，连接数据库得到查询结果后经过一些小小的处理，最终输出给用户。

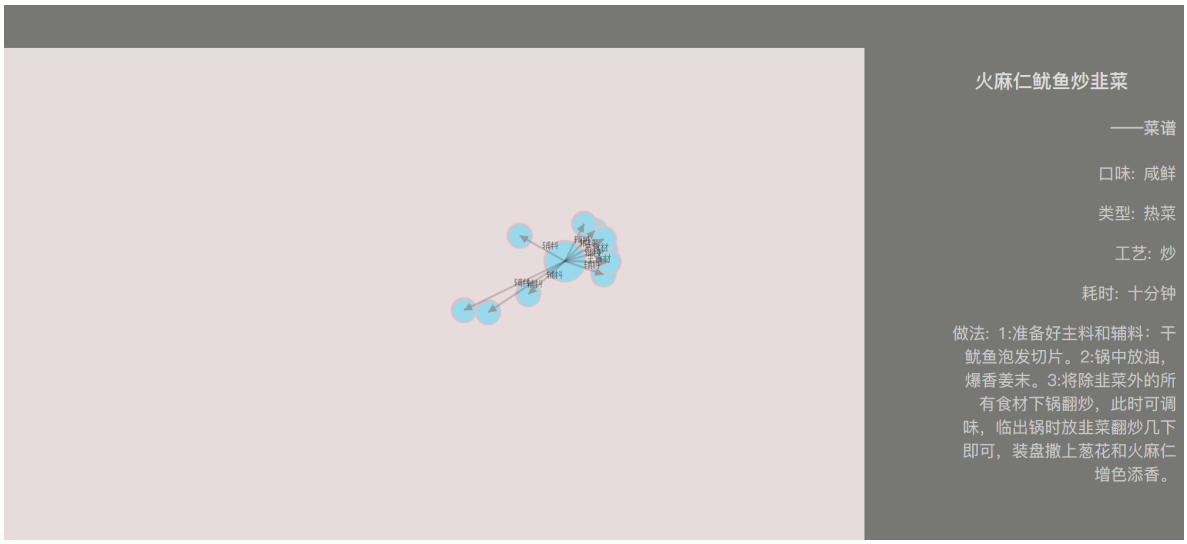
举例如下：

问句	解析结果	
红烧肉要用多少五花肉	question_type:"relationship_query", args:{ nodes:["烤肠":"recipe","猪肉":"material"],relationship:["主食材","辅料"]}	MATCH(n:recipe{name:"红烧肉"})-[r:主食材]->(m:material{name:"五花肉"}) return r.用量 UNION MATCH(n:recipe{name:"红烧肉"})-[r:辅料]->(m:material{name:"五花肉"}) return r.用量

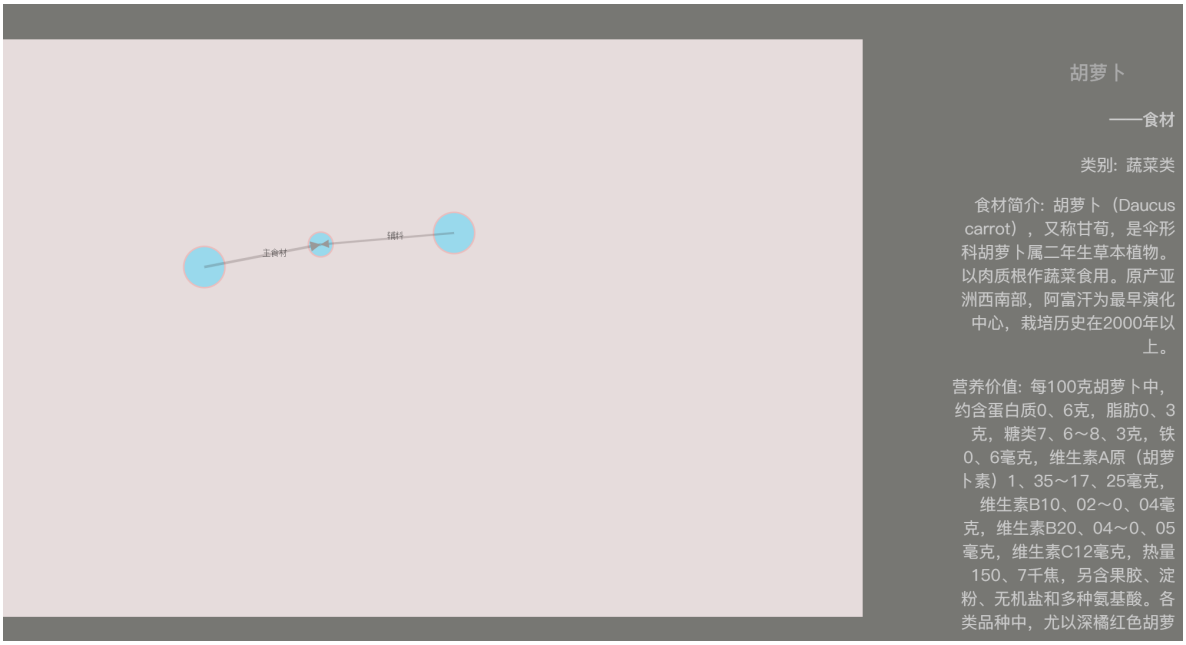
三、效果展示

3.1 可视化

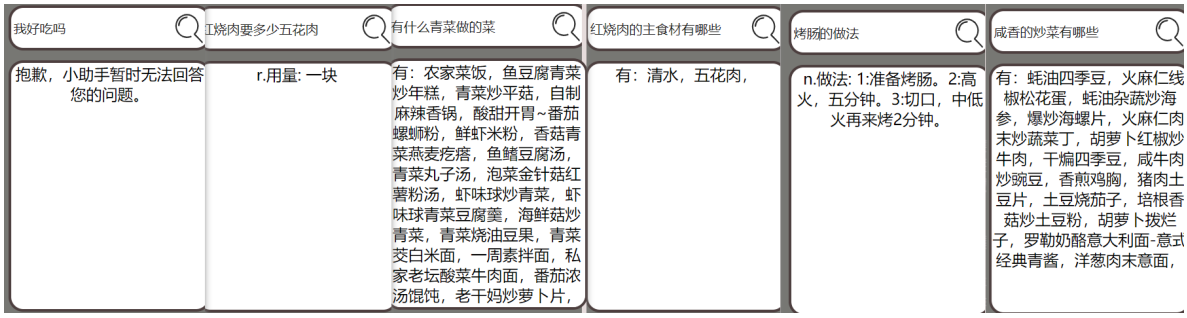
菜谱



食材


主食材
胡萝卜
——食材
类别：蔬菜类
食材简介：胡萝卜（Daucus carrot），又称甘荀，是伞形科胡萝卜属二年生草本植物。以肉质根作蔬菜食用。原产亚洲西南部，阿富汗为最早演化中心，栽培历史在2000年以上。
营养价值：每100克胡萝卜中，约含蛋白质0、6克，脂肪0、3克，糖类7、6～8、3克，铁0、6毫克，维生素A原（胡萝卜素）1、35～17、25毫克，维生素B10、02～0、04毫克，维生素B20、04～0、05毫克，维生素C12毫克，热量150、7千焦，另含果胶、淀粉、无机盐和多种氨基酸。各类品种中，尤以深橘红色胡萝

3.2 KBQA


我好吃吗
抱歉，小助手暂时无法回答您的问题。
红烧肉要多少五花肉
r.用量: 一块
有什么青菜做的菜
有：农家菜饭，鱼豆腐青菜炒年糕，青菜炒平菇，自制麻辣香锅，酸甜开胃~番茄螺蛳粉，鲜虾米粉，香菇青菜燕麦疙瘩，鱼鳍豆腐汤，青菜丸子汤，泡菜金针菇红薯粉汤，虾味球炒青菜，虾味球青菜豆腐羹，海鲜菇炒青菜，青菜烧油豆果，青菜茭白米面，一周素拌面，私家老坛酸菜牛肉面，番茄浓汤馄饨，老干妈炒萝卜片，
红烧肉的主食材有哪些
有：清水，五花肉，
烤肠的做法
n.做法: 1:准备烤肠。2:高火，五分钟。3:切口，中低火再来烤2分钟。
咸香的炒菜有哪些
有：蚝油四季豆，火麻仁线椒松花蛋，蚝油杂蔬炒海参，爆炒海螺片，火麻仁肉末炒蔬菜丁，胡萝卜红椒炒牛肉，干煸四季豆，咸牛肉炒豌豆，香煎鸡胸，猪肉土豆片，土豆烧茄子，培根香菇炒土豆粉，胡萝卜拨烂子，罗勒奶酪意大利面-意式经典青酱，洋葱肉末意面，