哈利波特人物关系知识图谱项目报告

参与成员：

- 董博宇
- 沈子衿
- 闫徐天任

1. 知识抽取

由哈利波特爱好者维护的哈利波特中文维基包含对哈利波特世界各个方面的介绍，本项目的知识主要来自此网站。在该维基的人物介绍界面中，包含对该人物的表格概述和大段文本介绍。

1.1 从半结构化数据提取知识

维基人物介绍页面中包含大量我们需要的关系信息，因此我们编写了一个爬虫程序，对页面中半结构化数据进行解析（从中抽取人物的属性信息和关系信息）。

我们的爬虫基于 Scrapy 框架，它是一个为了爬取网站数据，提取结构性数据而编写的应用框架，分为引擎、调度器、下载器、下载器中间件、管道等组件。我们实现的爬虫代码位于 `./crawler/spiders/wiki_spider.py` ，爬取的信息经过 `process_data.py` 初步预处理后（更改异型字、对同一概念统一命名）存储于 `./harry_potter_property.json` ，实例如下：


```
"莉莉·伊万斯": {
  "出生": "1960年1月30日英格兰",
  "逝世": "1981年10月31日（21岁)戈德里克山谷，英格兰",
  "血统": "麻瓜出身",
  "婚姻状况": "已婚（被谋杀前短暂丧偶）",
  "物种": "人类",
  "性别": "女",
  "头发颜色": "深红色",
  "眼睛颜色": "亮绿色",
  "皮肤颜色": "浅色",
  "家庭信息": {
   "父亲": "伊万斯先生",
   "母亲": "伊万斯夫人",
   "姐姐": "佩妮·德思礼",
   "丈夫": "詹姆·波特",
```

```
    "儿子": "哈利·波特",
    "儿媳": "金妮·韦斯莱",
    "孙子": [
     "詹姆·小天狼星·波特",
     "阿不思·西弗勒斯·波特"
    ],
    "孙女": "莉莉·卢娜·波特",
    "姐夫": "弗农·德思礼",
    "外甥": "达力·德思礼",
    "公公": "弗利蒙·波特",
    "婆婆": "尤菲米娅·波特",
    "亲家": "波特家族"
   },
   "学院": "格兰芬多学院",
   "从属": [
    "凤凰社",
    "霍格沃茨魔法学校",
    "鼻涕虫俱乐部",
    "格兰芬多学院",
    "波特家族",
    "伊万斯家庭"
   ]
  }
```


1.2 从纯文本数据中抽取关系

我们尝试使用 DeepKE（浙江大学基于深度学习的开源中文关系抽取工具）从人物介绍纯文本中进行关系抽取。

由于缺乏训练数据，我们首先搜寻了开源的人物关系抽取数据集，再编写程序将数据转换为 DeepKE 接受的格式，将其按8:1:1的比例分为训练集、测试集和验证集，并对数据进行清洗。 最终，该数据集包含3097个训练数据、306条测试数据和387条验证数据，共有unknown、夫妻、父母、兄弟姐妹、上下级、师生、好友、同学、合作、同人、情侣、祖孙、同门、亲戚等 14 个类别的关系。

我们使用 CNN 模型、默认训练配置进行训练（最优验证集准确率 50%，测试集准确率 48%）。

训练好模型后，使用 `predict.py` 对哈利波特维基百科中的纯文本人物介绍进行人物关系抽取，下面给出三个实例：

```
请输入句子：德拉科·马尔福（Draco Malfoy），是个纯血统巫师，卢修斯·马尔福和纳西莎·马尔福（原姓布莱克）的独生子。
请输入句中需要预测关系的头实体：德拉科·马尔福
请输入头实体类型（可以为空，按enter跳过）：人物
请输入句中需要预测关系的尾实体：卢修斯·马尔福
请输入尾实体类型（可以为空，按enter跳过）：人物
[__main__][INFO] - "德拉科·马尔福" 和 "卢修斯·马尔福" 在句中关系为："父母"，置信度为0.20。

请输入句子：德拉科最终与阿斯托利亚·格林格拉斯结婚，并至少有了一个孩子—斯科皮·马尔福。
请输入句中需要预测关系的头实体：德拉科
请输入头实体类型（可以为空，按enter跳过）：人物
请输入句中需要预测关系的尾实体：阿斯托利亚·格林格拉斯
请输入尾实体类型（可以为空，按enter跳过）：人物
[__main__][INFO] - "德拉科" 和 "阿斯托利亚·格林格拉斯" 在句中关系为："夫妻"，置信度为0.34。

请输入句子：赫敏也很照顾她的孩子。当罗恩开玩笑说如果谁没分进格兰芬多就解除谁的继承权时，赫敏安慰了女儿罗丝和侄子阿不思。
请输入句中需要预测关系的头实体：赫敏
请输入头实体类型（可以为空，按enter跳过）：人物
请输入句中需要预测关系的尾实体：罗丝
请输入尾实体类型（可以为空，按enter跳过）：人物
[__main__][INFO] - "赫敏" 和 "罗丝" 在句中关系为："父母"，置信度为0.41。
```

由于从半结构化数据中抽取的关系粒度更细、准确率更高，且已经覆盖比较全面，后续构建图谱时仅采用 1.1 节获取的信息。

2. 知识存储

我们最初通过编写程序对 `./harry_potter_property.json` 进行结构变换，形成关系三元组文件 `harry_potter.json` 进行存储，截取实例如下：

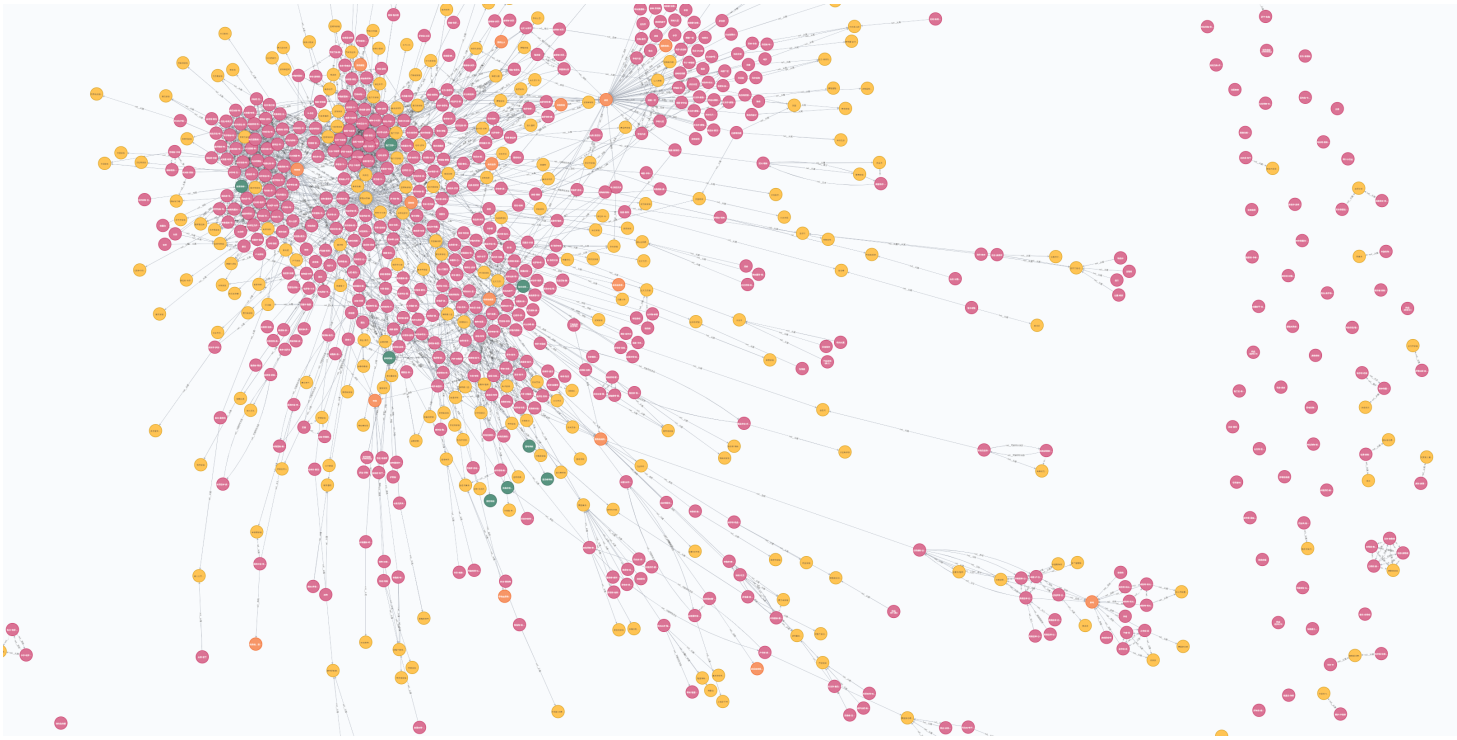
```
{"entity1": "亚瑟·韦斯莱", "entity2": "罗恩·韦斯莱", "relation": "父亲"}
{"entity1": "莫丽·韦斯莱，原姓普威特", "entity2": "罗恩·韦斯莱", "relation": "母亲"}
{"entity1": "比尔·韦斯莱", "entity2": "罗恩·韦斯莱", "relation": "哥哥"}
{"entity1": "查理·韦斯莱", "entity2": "罗恩·韦斯莱", "relation": "哥哥"}
{"entity1": "珀西·韦斯莱", "entity2": "罗恩·韦斯莱", "relation": "哥哥"}
{"entity1": "弗雷德·韦斯莱", "entity2": "罗恩·韦斯莱", "relation": "哥哥"}
{"entity1": "乔治·韦斯莱", "entity2": "罗恩·韦斯莱", "relation": "哥哥"}
{"entity1": "金妮·韦斯莱", "entity2": "罗恩·韦斯莱", "relation": "妹妹"}
{"entity1": "赫敏·格兰杰", "entity2": "罗恩·韦斯莱", "relation": "妻子"}
```

随着对课程进一步学习，我们希望使用 Neo4j 对知识图谱进行存储和应用。Neo4j 是一个图数据库管理系统，符合 ACID 标准，与关系型数据库不同，它原生支持图的存储和处理。为了支持数据的存储，我们编写了 `json2rdf.py` 将爬取的半结构化数据转化为 RDF 语言表述的知识格式（使用 Turtle 格式）。RDF（Resource Description Framework，资源描述框架）是一种资源描述语言，它可以用三元组集合的方式来表示事物之间的语义关系。

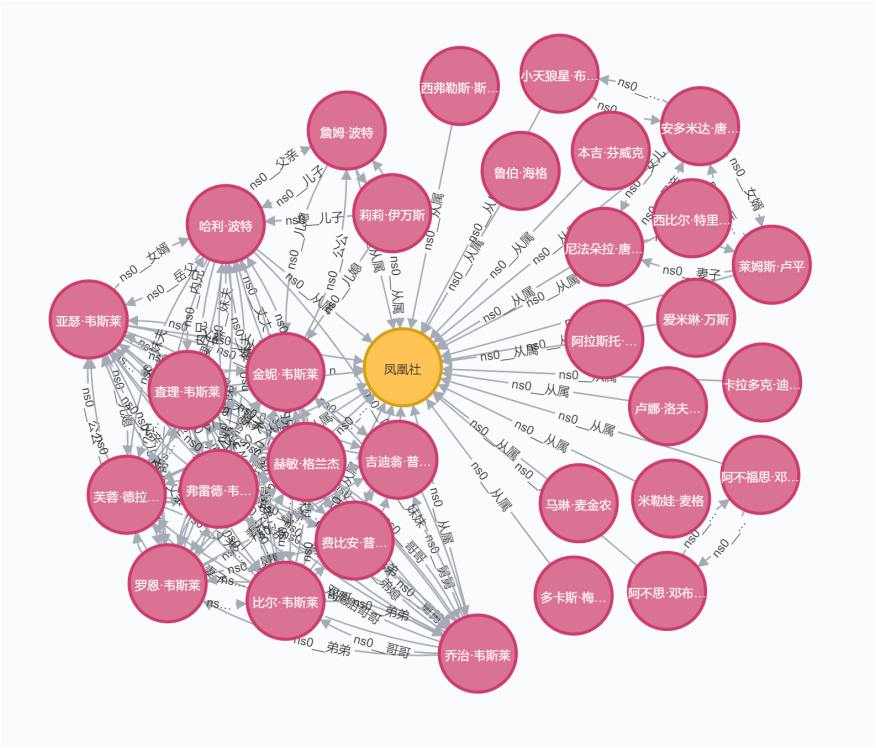
我们转换生成的rdf文件中的资源实体包括人物、组织、学院等，示例如下：

```
character:卢娜·洛夫古德
        relation:name "卢娜·洛夫古德" ;
        a relation:Character ;
        relation:出生 "1981年2月13日英国" ;
        relation:血统 blood:纯血统或混血统 ;
        relation:婚姻状况 "已婚" ;
        relation:物种 "人类" ;
        relation:性别 "女" ;
        relation:头发颜色 "脏金色" ;
        relation:眼睛颜色 "浅银白色" ;
        relation:皮肤颜色 "苍白" ;
        relation:母亲 character:潘多拉·洛夫古德 ;
        relation:丈夫 character:罗夫·斯卡曼德 ;
        relation:儿子 character:洛肯·斯卡曼德 ;
        relation:儿子 character:莱桑德·斯卡曼德 ;
        relation:祖翁 character:纽特·斯卡曼德 ;
        relation:职业 "神奇动物学家" ;
        relation:学院 house:拉文克劳学院 ;
        relation:从属 group:洛夫古德家庭 ;
        relation:从属 group:斯卡曼德家族 ;
        relation:从属 group:霍格沃茨魔法学校 ;
        relation:从属 group:拉文克劳学院 ;
        relation:从属 group:邓布利多军 ;
        relation:从属 group:凤凰社 ;
        relation:从属 group:《唱唱反调》 .
```

接下来，我们使用 Neo4j Labs 中的 Neosemantics 插件，将 RDF 文件导入到 Neo4j 数据库中，利用neo4j的可视化工具，我们可以看到整张图的结构如下：



其中节点的类型包括人物、组织、学院等。我们可以通过Cypher语句在数据库中对数据进行查询，Cypher是一种声明性图查询语言，它允许在属性图中进行表达性和高效的数据查询。与关系型数据库不同，Cypher对于图数据库操作进行了特化，数据被构造为节点和关系连接，以关注数据中的实体如何相互连接和相互关联。我们可以通过cypher语句在数据库中查询人物相关信息，如查询所有凤凰社的成员：



3. 知识计算

在将获得的数据存入neo4j数据库之后，我们使用了Cypher查询语言执行了一系列主流知识计算工作。

3.1 图谱基本信息

我们使用基本的可视化手段，统计了本知识图谱的一些基本信息，包括节点总数、节点度数、最短路径等。现列举如下。

图谱的节点总数：

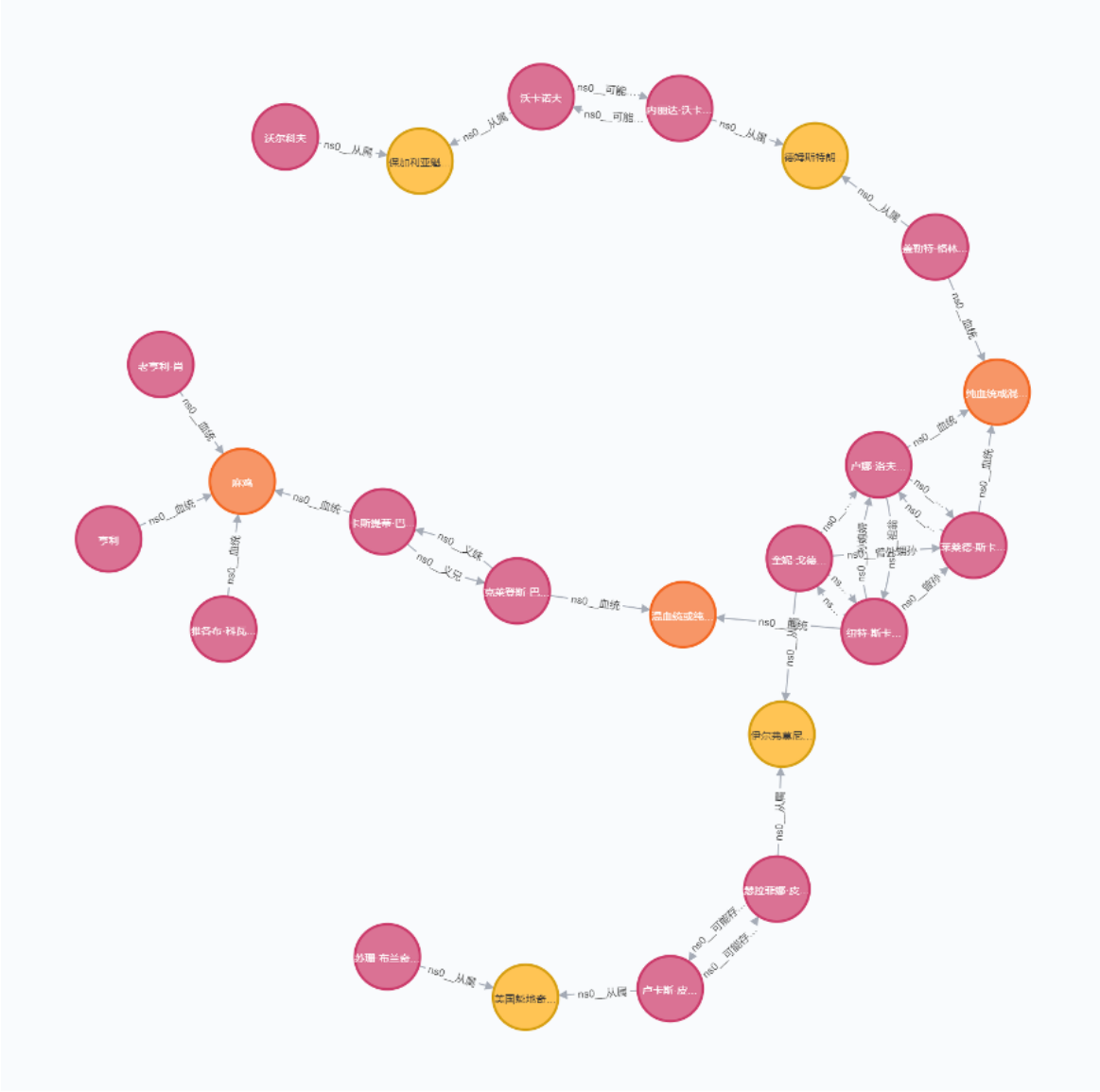
	count(c)
1	669

部分高度数节点度数：

	name	degree
1	"哈利·波特"	70
2	"亚瑟·韦斯莱"	67
3	"金妮·韦斯莱"	58
4	"罗恩·韦斯莱"	58
5	"比尔·韦斯莱"	51
6	"阿不思·西弗勒斯·波特"	50
7	"查理·韦斯莱"	49
8	"赫敏·格兰杰"	47
9	"弗雷德·韦斯莱"	45
10	"乔治·韦斯莱"	44

图谱直径（最长的最短路径）与路径可视化：

	length	path
1	13	["角色#劳珊·布兰奇弗劳尔", "组织#美国魁地奇国家队", "角色#卢卡斯·皮奎利", "角色#瑟拉菲娜·皮奎利", "组织#伊尔弗莫尼魔法学校", "角色#奎妮·戈德斯坦", "角色#卢娜·洛夫古德", "血统#纯血统或混血统", "角色#盖勒特·格林德沃", "组织#德姆斯特朗学院", "角色#内丽达·沃卡诺娃", "角色#沃卡诺夫", "组织#保加利亚魁地奇国家队", "角色#沃尔科夫"]
2	13	["角色#沃尔科夫", "组织#保加利亚魁地奇国家队", "角色#沃卡诺夫", "角色#内丽达·沃卡诺娃", "组织#德姆斯特朗学院", "角色#盖勒特·格林德沃", "血统#纯血统或混血统", "角色#莱桑德·斯卡曼德", "角色#纽特·斯卡曼德", "血统#混血统或纯血统", "角色#克莱登斯·巴瑞波恩", "角色#卡斯提蒂·巴瑞波恩", "血统#麻鸡", "角色#老亨利·肖"]
3	13	["角色#沃尔科夫", "组织#保加利亚魁地奇国家队", "角色#沃卡诺夫", "角色#内丽达·沃卡诺娃", "组织#德姆斯特朗学院", "角色#盖勒特·格林德沃", "血统#纯血统或混血统", "角色#莱桑德·斯卡曼德", "角色#纽特·斯卡曼德", "血统#混血统或纯血统", "角色#克莱登斯·巴瑞波恩", "角色#卡斯提蒂·巴瑞波恩", "血统#麻鸡", "角色#雅各布·科瓦尔斯基"]
4	13	["角色#亨利", "血统#麻鸡", "角色#卡斯提蒂·巴瑞波恩", "角色#克莱登斯·巴瑞波恩", "血统#混血统或纯血统", "角色#纽特·斯卡曼德", "角色#莱桑德·斯卡曼德", "血统#纯血统或混血统", "角色#盖勒特·格林德沃", "组织#德姆斯特朗学院", "角色#内丽达·沃卡诺娃", "角色#沃卡诺夫", "组织#保加利亚魁地奇国家队", "角色#沃尔科夫"]



3.2 PageRank

PageRank，又称网页排名、谷歌左侧排名、PR，是Google公司所使用的对其搜索引擎搜索结果中的网页进行排名的一种算法，本质上是一种以网页之间的超链接个数和质量作为主要因素粗略地分析网页的重要性的算法。其基本假设是：更重要的页面往往更多地被其他页面引用（或称其他页面中会更多地加入通向该页面的超链接）。 其将从A页面到B页面的链接解释为“A页面给B页面投票”，并根据投票来源（甚至来源的来源，即链接到A页面的页面）和投票对象的等级来决定被投票页面的等级。简单的说，一个高等级的页面可以提升其他低等级的页面。

在本工作中，我们使用了neo4j官方提供的Graph Data Science（GDS）工具包进行了 PageRank 的生成和后续的社区检测，结果（分数最高的几个人物）如下：

	name	score
1	"塞蒂娜·沃贝克"	3.193918639623846
2	"厄文·瓦布尔"	3.0560807875054623
3	"亚瑟·韦斯莱"	2.6437935175161056
4	"威廉一世"	2.3281605859046812
5	"玛丽·卢·巴瑞波恩"	2.0415618714538484
6	"西格纳斯·布莱克三世"	2.030287133838563
7	"菲尼亚斯·奈杰勒斯·布莱克"	2.0268379391985234
8	"罗夫·斯卡曼德"	1.9477174026001696
9	"安多米达·唐克斯"	1.9090889391558457
10	"纳西莎·马尔福"	1.9022708447796122

3.3 社区检测

在图论中，内部连接比较紧密的节点子集合对应的子图叫做社区（community）。其中，各社区节点集合彼此没有交集的称为非重叠型（disjoint）社区，有交集的称为重叠型（overlapping）社区。网络图中包含一个个社区的现象称为社区结构，它是网络中的一个普遍特征。给定一个网络图，找出其社区结构的过程叫做社区检测（community detection）。社区检测算法的一个关键作用在于可用于从网络中提取有用的信息。

在本工作中，我们使用GDS提供的社区检测算法中的Louvain算法对数据进行提取和处理，结果如下。可以发现，属于同一家族的实体基本都被聚合在了同一个社区中，这说明算法是有效的。

	communityCount	modularity	modularities
1	463	0.7092729693284414	[0.661820572271665, 0.7084898839721376, 0.7092729693284414]

	name	communityId	intermediateCommunityIds
1	"乔治·韦斯莱"	540	*null*
2	"亚瑟·韦斯莱"	540	*null*
3	"塞普蒂默斯·韦斯莱"	390	*null*
4	"多米尼克·韦斯莱"	540	*null*
5	"奥黛丽·韦斯莱"	540	*null*
6	"弗雷德·韦斯莱"	540	*null*
7	"弗雷德·韦斯莱二世"	540	*null*
8	"查理·韦斯莱"	540	*null*
9	"比尔·韦斯莱"	540	*null*

4. 知识应用——基于规则的知识问答

基于上述工作，我们在知识图谱上尝试了基于规则的知识问答这一应用场景。在实际交互过程中，访问者输入一段特殊或一般疑问的文本，问答算法提取其中的关键字后执行数据库查找，最终返回对应的结果。

我们目前支持的问句形式有：

- 查询人物的属性，包括出生/逝世/血统/物种/性别/身高/婚姻/职业/学院
- 查询人物关系对象，如哈利·波特的父亲是谁，谁的父亲是哈利·波特
- 查询组织的从属关系，包括人物从属于哪些组织，某组织有哪些成员

具体效果如下：

```
问题：哈利·波特是什么血统？
回答： 哈利·波特的血统是混血统

问题：哈利·波特的父亲是谁？
回答： 哈利·波特的父亲是詹姆·波特

问题：谁的父亲是哈利·波特？
回答： 莉莉·卢娜·波特  阿不思·西弗勒斯·波特  詹姆·小天狼星·波特  的父亲是哈利·波特

问题：请问赫敏·格兰杰的出生信息？
回答： 赫敏·格兰杰的出生是1979年9月19日英国

问题：阿不思·邓布利多是什么职业？
回答： 阿不思·邓布利多的职业是霍格沃茨校长 （早于1971年-1997年）

问题：西弗勒斯·斯内普是哪个学院的？
回答： 西弗勒斯·斯内普的学院是斯莱特林学院

问题：小天狼星·布莱克属于哪些组织？
回答： 小天狼星·布莱克从属的组织有：凤凰社  霍格沃茨魔法学校  劫猎者  波特家族  布莱克家族


问题：食死徒有哪些成员？
回答： 从属于食死徒的成员有：特拉弗斯  西弗勒斯·斯内普  雷古勒斯·布莱克  德拉科·马尔福  贝拉特里克斯·莱斯特兰奇
安东宁·多洛霍夫  威尔克斯  芬里尔·格雷伯克  奥古斯特·卢克伍德  斯坦·桑帕克
多洛雷斯·乌姆里奇  卢修斯·马尔福  马法尔达·霍普柯克  汤姆·里德尔
```

在处理时，获得疑问文本后，算法首先对其进行分词，从中识别出如下实体：属性词、专有名词、组织词、关系词和从属介词。随后利用这些实体作为关键词，我们使用py2neo连接neo4j数据库并在数据库中使用cypher语句进行查询，再对返回结果进行组装，就得到了回答。