1 项目介绍

《三国演义》描写了从东汉末年到西晋初年之间近百年的历史风云，诉说了东汉末年的群雄割据混战和魏、蜀、吴三国之间的政治和军事斗争，塑造了一群叱咤风云的三国英雄人物。

《三国演义》里面有名有姓的人物不下千余人。因此，本组构建了《三国演义》中的实体与其关系的知识图谱。其中，实体主要包括人物、国家；关系主要包括人物与人物之间的关系（如父子、母子、兄弟、配偶等）、人物与国家之间的关系（如效力、主效力）。

我们首先通过爬虫从网站上完成数据采集。由于网站上为文本信息，因此需要对采集到的数据进行预处理。然后采用预训练的关系抽取模型抽取现有实体之间的关系，以构建知识图谱的三元组。完成知识图谱的构建后，我们将《三国演义》人物关系图谱导入Neo4j图数据库，完成了该图谱的存储与可视化。基于此Neo4j中的图谱存储，我们实现了一个小型的QA系统应用。

综上，我们的项目工作可以分为四个部分：
1）数据采集与预处理
2）关系抽取与图谱构建
3）Neo4j存储与可视化
4）QA系统应用的实现

小组成员：

姓名	学号
王嘉琪	22121067
谢谨蔓	22121062
丁宇佳	22121100
王鑫澳	22121106

2 实现过程

2.1 数据采集与预处理

我们经过调研，选择了三国人物谱 http://www.yanghuagai88.com/ 这个网站作为数据来源。

获取数据时，我们选择使用Python的scrapy框架Spider类来爬取网页数据，主要使用xpath来定位网页的特定位置进行解析，然后提取与人物相关的初始数据。将具体方法部分展示如下：

```
class QuotesSpider(scrapy.Spider):
    name = "quotes"

    def start_requests(self):
        urls = [
            'http://www.yanghuagai88.com/shuguorw'
        ]
        for url in urls:
            yield scrapy.Request(url=url, callback=self.parse)
```

```
    def parse(self, response):
        urls = response.xpath('//div[contains(@class,"post istop cate1
auth1")]//a/@href').getall()
        for url in urls:
            print('----------------------------')
            print(url)
            yield scrapy.Request(url=url, callback=self.haha)

    def haha(self,response):
        text = response.xpath('//div[contains(@class,"entry")]
[1]/h2[1]/preceding-sibling::*').xpath('string(.)').getall()
        text1 = "".join(text)
        text1 = text1.replace("\r", "").replace("\t", "").replace("\n",
"").replace(" ","")
        name = response.xpath('//div[contains(@class,"entry")]//a/text()')
[0].get()
        yield {'text': text1}
```

得到初始数据后，我们首先对不规则、错误的异常数据进行清洗。例如，数据中出现实体名称错误的情况，由个别不规则的数据导致，我们将这种数据直接删除处理；数据中的关系包含特殊字符，我们将这种关系也进行删除处理。

2.2 关系抽取与图谱构建

我们选择使用哈工大BERT-wwm模型，在人物关系数据上训练关系抽取模型。训练完成后，使用该模型对爬取到的数据进行关系抽取，得到各种实体间的关系。

得到各种实体间的关系后，仍然需要进行清洗。首先从中筛选出我们需要的关系，包括：父子、母子、兄弟、配偶、效力、主效力。第二，对于抽取到发生冲突的数据，我们选择进行删除处理。第三，对于同一个实体对存在多个关系的情况，我们选择概率最高的一组作为他们之间的关系。最后，对已抽取到的关系进行简单推理，例如存在A、兄弟、B的三元组和B、兄弟、C的三元组，可以做出推理：得到A、兄弟、C的三元组。

最终，我们得到我们需要的、干净的知识图谱数据，体现在renwu_relate.csv文件中，部分数据展示如下。

	A	B	C	D
1	编号	实体1	关系	实体2
65	1729	卞隆	brother	卞晖
66	1728	卞琳	brother	卞兰
67	1729	卞隆	father	卞兰
68	1726	卞晖	father	卞兰
69	6074	广阳乡君	mate	卞琳
70	1727	卞兰	brother	卞琳
71	1730	卞皇后[曹	father	卞琳
72	5606	显阳乡君	mate	卞隆
73	3845	顺阳乡君	mate	卞隆
74	1726	卞晖	brother	卞隆
75	5663	卞皇后[曹	father	卞隆
76	47	曹植	mother	卞氏
77	225	曹丕	mother	卞氏
78	205	曹熊	mother	卞氏
79	2925	敬侯夫人	mate	卞远
80	1412	卞秉	father	卞远
81	914	武宣皇后	father	卞远
82	1683	拔奇	father	伯固
83	4867	伊夷模	father	伯固
84	182	步协	brother	步阐
85	17	孙权	mate	步夫人

2.3 Neo4j存储与可视化

在本项目中，完成知识图谱的构建后，我们选择将《三国演义》人物关系图谱导入最为常用的Neo4j图数据库，作为底层存储数据库，完成了该图谱的存储与可视化。

具体来说，采用python的py2neo库作为Neo4j的接口，将知识图谱csv数据文件中的三元组存储至Neo4j数据库，部分代码如下。

```
def add_relation(data):
    relations = []
    rel_map = {
        'mate': '配偶',
        'cata': '主效',
        'cated': '曾效力',
        'father': '父亲',
        'mother': '母亲',
        'brother': '兄弟'
    }
    tx = graph.begin()

    def process_relation_(value):
        name1 = value['实体1']
        name2 = value['实体2']
        relation = value['关系']
        if name2 == '未知':
            return
        if relation == 'cata' or relation == 'cated':
            b = graph.nodes.match("Force", name=name2).first()
        else:
            b = graph.nodes.match("Person", name=name2).first()
        a = graph.nodes.match("Person", name=name1).first()
        if a == None or b == None:
            return
        rel = Relationship(a, rel_map[relation], b)
        relations.append(rel)

    data.apply(process_relation_, axis=1)
    relations = Subgraph(relationships=relations)
    tx.create(relations)
    graph.commit(tx)
```

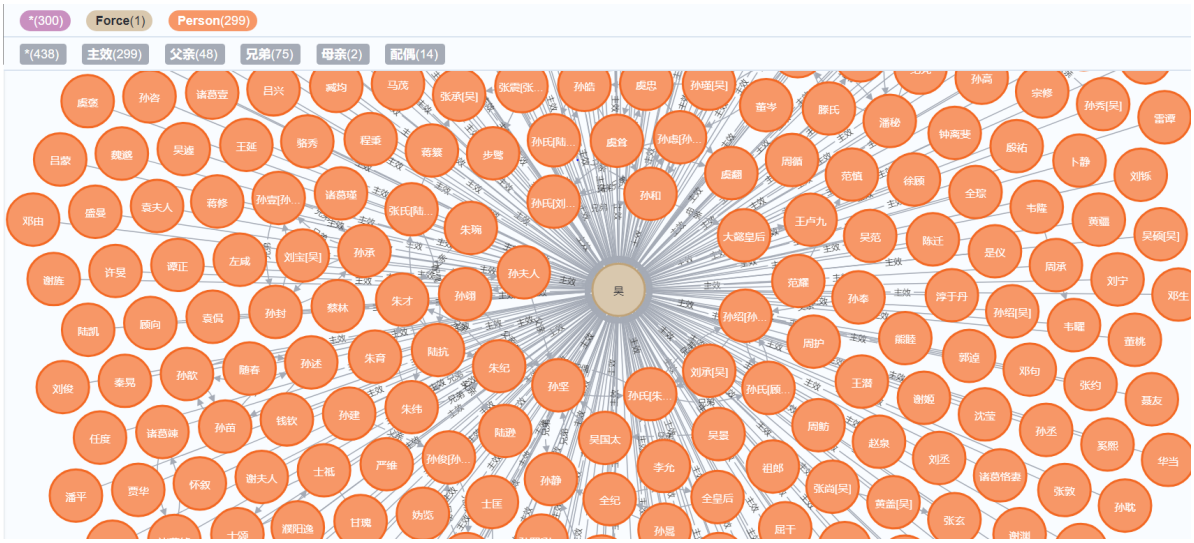
导入存储完成后，在Neo4j的本地网页端，可以进行可视化，也可以用相应的命令来进行知识查询操作，结果同样以可视化形式呈现。下面是部分展示。

1. 查询主效力吴国的所有人物以及他们之间的关系。

命令：

```
MATCH (a:Force{name:"吴"})-[:`主效`]-(b:Person) RETURN [(a)-[]-(b)] ORDER BY b
```

结果：

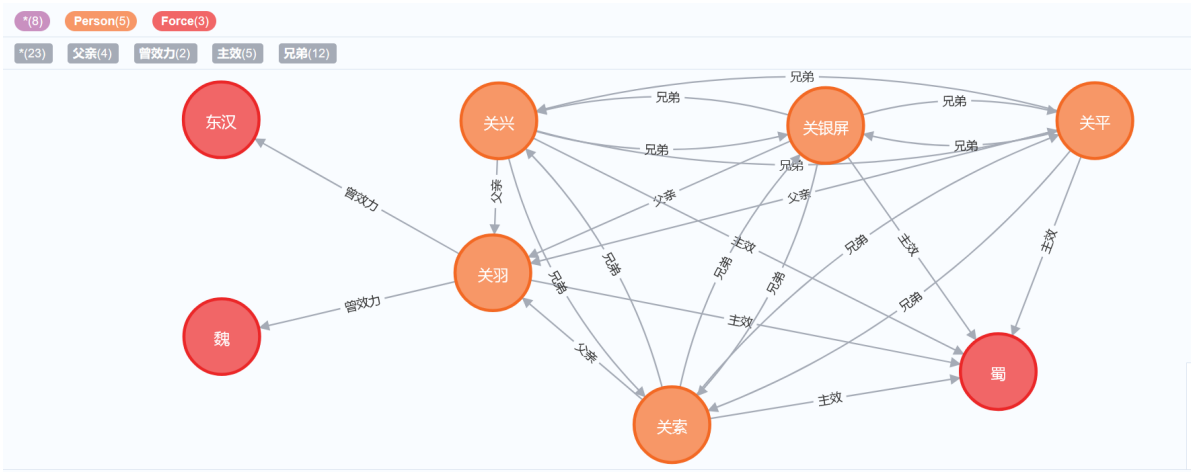


2. 查询关羽相关的所有关系。

命令：

```
MATCH (a:Person{name:"关羽"})-() RETURN a
```

结果：

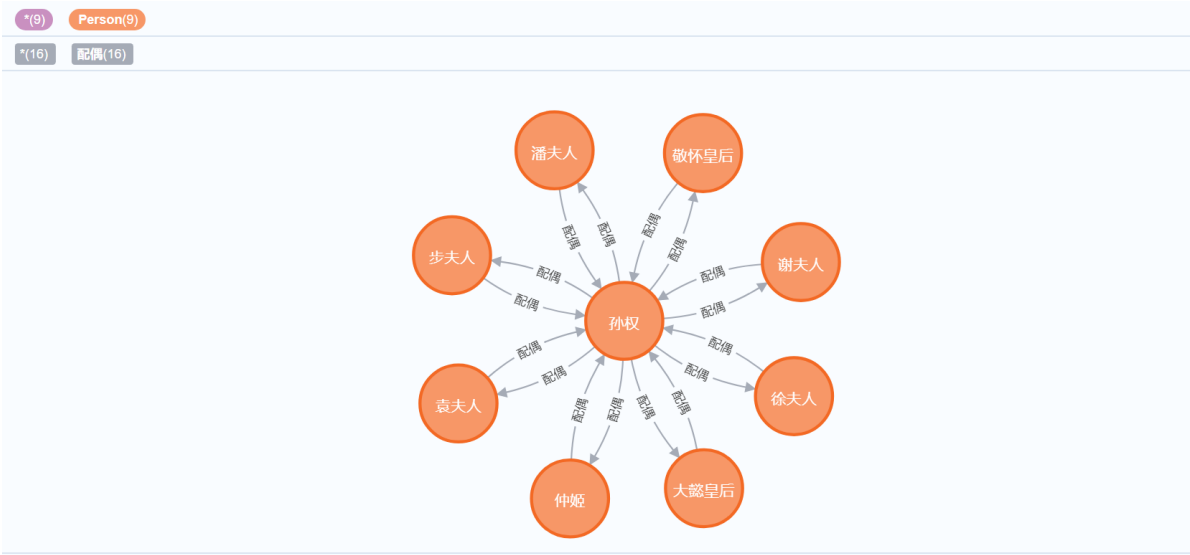


3. 查询孙权的所有配偶。

命令：

```
MATCH (a:Person{name:"孙权"})-[:`配偶`]-(b:Person) RETURN [(a)-[]-(b)]
```

结果：

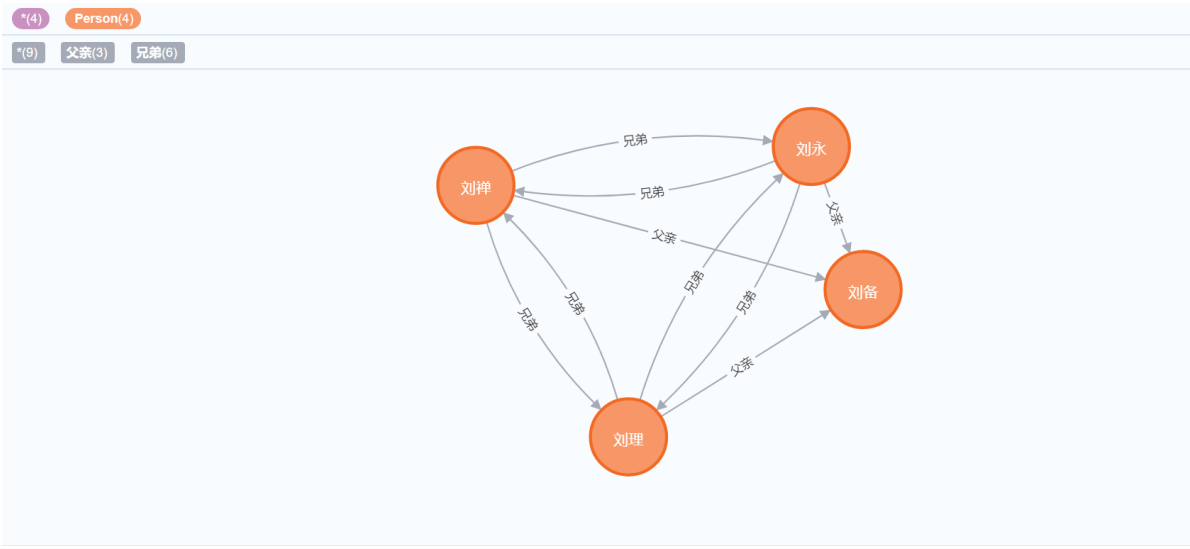


4. 查询刘备的所有儿子及其关系。

命令：

```
MATCH (a:Person{name:"刘备"})-[:`父亲`]-(b:Person) RETURN [(b)-[]->(a)]
```

结果：



2.4 QA系统应用的实现

1. 连接Neo4j

首先通过以下代码连接Neo4j中已存储的知识图谱的相关配置。

```
neo4j_url = "http://localhost:7474"
user = "neo4j"
pwd = "4399"
```

2. 问题解析

首先，尝试将输入转换为整型数据，如果不能转化为整型数据，证明问的是问题，开始进行问答操作。然后，查看询问的是否是人物关系问题，若是人物关系问题，且证明分割出来的是人名，则进入查询人物关系的问答程序；若询问的是效力国家的问题，则进入国家问题的问答程序。部分代码如下：

```
try:
    choose = int(question)
except:
    list1 = question.split("的")
    name = list1[0]
    if name != "none" and len(name) < 5:
        find_person(question, graph)
    else:
        find_work(question, graph)
```

3. 答案查询

然后，通过与Neo4j的链接来查询知识图谱，返回问题答案。其中做人物关系查询的代码如下：

```
def find_person(question, graph):
    list1 = question.split("的")
    name = list1[0]
    list2 = list1[1].split("是")
    relation = list2[0]
    cur_rel = rel_trans(relation)
    if cur_rel == "none":
        print("没有看懂您说的什么意思哦~")
        return 0
    else:
        print("为您找到以下答案：")
        print(name + "的" + relation + "为：")
        # 查找关系
        node_matcher = NodeMatcher(graph)
        relationship_matcher = RelationshipMatcher(graph)
        node = node_matcher.match("Person").where(name=name).first()
        relationship = list(relationship_matcher.match([node], r_type=cur_rel))
        if relationship == "none":
            print("抱歉，这件事我也不知道~")
        else:
            for i in relationship:
                print(i.end_node["name"])
```

下面是QA系统的问答展示。
1）小乔的姐妹

```
欢迎使用三国演义问答系统!
--------------------------------------------------
请输入您的问题:(输入0退出系统)小乔的姐妹
为您找到以下答案:
小乔的姐妹为:
大乔
--------------------------------------------------
```

2）孙权的父亲是

```
--------------------------------------------------
请输入您的问题:(输入0退出系统)孙权的父亲是
为您找到以下答案:
孙权的父亲为:
孙坚
--------------------------------------------------
```

3）赵云是哪国的

```
--------------------------------------------------
请输入您的问题:(输入0退出系统)赵云是哪国的
为您找到以下答案:
赵云曾经效力为:
袁绍
赵云主要效力为:
蜀
--------------------------------------------------
```

4）刘备的妻子是谁

```
----------------------------------------------------------------
请输入您的问题：（输入0退出系统）刘备的妻子是谁
为您找到以下答案：
刘备的妻子为：
穆皇后
孙夫人
昭烈皇后
麋夫人
----------------------------------------------------------------
```