华南理工大学
South China University of Technology

《机器学习》 课 程 设 计 报 告

(2019-2020 学年第二学期)

基于 Elasticsearch 的 COVID-19 智能问答机器人

学生姓名： 何梵 李洁滢 陈奕凯

提交日期：2020 年 8 月 14 日　　学生签名：

学　　号	201730053261 何梵 201730730339 李洁滢 201730730124 陈奕凯	座位编号	
学　　院	经济与贸易学院	专业班级	汇丰金融科技精英班
课程名称	机器学习	任课教师	黄晓宇
教师评语：			
本论文成绩评定： _____分			

目录

基于 Elasticsearch 的 COVID-19 智能问答机器人

1. 功能介绍

1.1 实体检索

实体检索即输入实体名称，返回该实体的所有属性和属性值。

示例输入

- **新冠病毒**
- **新冠肺炎**
- **密切接触者**

1.2 实体的属性检索

输入实体名称和一个属性名称，如果该实体存在该属性值，则返回该属性值。

示例输入

- **新冠肺炎的传染源？**
- **购物的注意事项有哪些？**
- **疫情期间感到焦虑怎么办？**

1.3 多跳查询

多跳查询即形如"姚明的女儿的身高"的查询,即"姚明:女儿"查询得到的是实体"姚明"的一个属性，但同时这个属性值也作为一个实体存在于数据集中，那么就可以接着对该实体继续查询其属性。

示例输入

- **广东的省会的感染人数**

1.4 根据属性值查询实体

输入隐含多对［属性名 operator 属性值]的自然语言问句， 它们之间的关系可以是 AND，OR，NOT，同时属性值可以是等于，大于，小于一个输入值，返回满足这些属性限制的实体。

示例输入

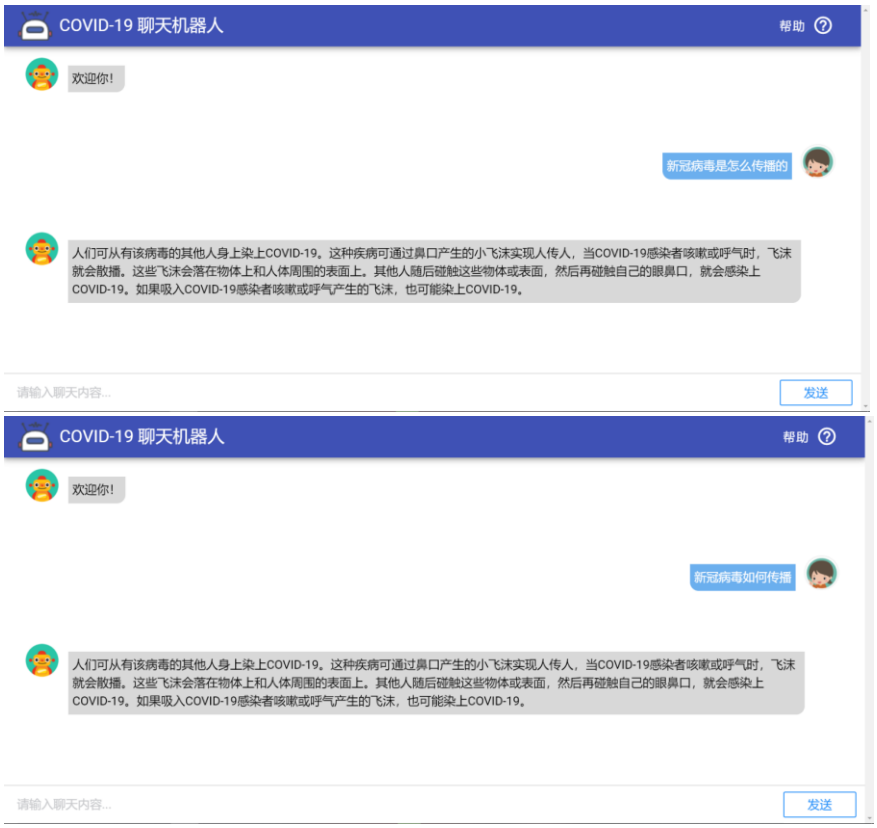
- **累计确诊超过 1500 的中国省份或美国州**

1.5 模糊匹配

同一个问题有多种问法，我们设置同义词库将问题映射成标准问法，从而可以在Elasticsearch搜索。

示例输入

- **新冠病毒是如何传播的？**
- **新冠病毒怎么传播？**
- **新冠病毒的传播途径？**
- **映射为标准问法：新冠病毒的传播方式**



2. 数据准备

2.1 爬取丁香园、世界卫生组织官网的关于新冠肺炎的问答集

```
QA_dict={}
try:
    response = requests.get(url,headers=headers)#请求网页内容
    if response.status_code == 200:#判断是否正确响应
        for i in range(7):
            page=response.json().get('data').get('items')[0]['article'][i]
            for j in range(len(page)):
                Q=re.sub("<(.*?)>",'',page['detail_search'][j]['title'].rep
                A=re.sub("<(.*?)>",'',page['detail_search'][j]['content'].r
                QA_dict[Q]=A
except requests.ConnectionError as e:
    print('Error', e.args)
```

	Question	Answer
0	什么是新型冠状病毒？	此次流行的冠状病毒为一种新发现的冠状病毒，国际病毒分
1	新型冠状病毒肺炎由什么	由 SARS-Cov-2 冠状病毒引起，WHO 将 SARS-Cov-2 感染导
2	冠状病毒的致病性如何？	冠状病毒主要感染成人或较大儿童，引起普通感冒和咽喉炎
3	新型冠状病毒与 SARS 病	新型冠状病毒与 SARS 病毒、MERS 病毒同属于冠状病毒这个
4	出现什么症状可能感染了	新型冠状病毒感染的一般症状有：发热、乏力、干咳，逐渐
5	如果出现发热、乏力、干	很多呼吸道疾病都会表现为发热、乏力、干咳等症状，是否
6	干咳是症状之一，那么干	干咳与咳嗽的主要区别在于是否有痰。干咳是指咳嗽时无痰
7	出现什么症状需要及时就	如果出现发热、乏力、肌肉酸痛、咳嗽、咳痰、气促、腹泻
8	新型冠状病毒肺炎的病原	新型冠状病毒属于 β 属的新型冠状病毒，有包膜，颗粒呈
9	新型冠状病毒肺炎如何传	经呼吸道飞沫、接触传播是主要的传播途径，相对封闭环境
10	什么是可疑暴露者？什么	可疑暴露者是指暴露于新型冠状病毒检测阳性的野生动物、
11	什么是接触传播？	接触传播，包括直接接触传播和间接接触传播。直接接触传
12	新型冠状病毒肺炎患者有	新型冠状病毒肺炎往往以发热作为主要起病的表现，可合并

2.2 将一问一答转化为三元组（实体名：属性名：属性值）

S(subject)	P(predicate)	O(object)
购物	注意事项	购物时，要与他人保持至少1米的
焦虑	做法	在大流行疫情一类情况下，感到焦
接触传播	定义	接触传播，包括直接接触传播和间
咳嗽和打喷嚏	注意事项	咳嗽和打喷嚏时，含有病毒的飞沫
口罩	佩戴方法	如果选择佩戴口罩：在拿口罩前，
口罩	选择	医用口罩（又叫手术专用口罩）：
老年人	防范措施	要在所在社区防范COVID-19，可
密切接触者	定义	密切接触者是指与可疑感染者或确
密切接触者	做法	对于密切接触者，需要在家进行医
危重病例	诊断方法	符合以下情况之一者：出现呼吸衰
心理创伤	做法	无论是什么原因，失去亲人总会很
新冠病毒	传播方式	人们可从有该病毒的其他人身上获

2.3 将三元组数据集转化为 json 格式

Elasticsearch 要求文档的输入格式为 json。将实验数据集转化为 json 格式后，每个实体对应一个 json 的 object，也即 Elasticsearch 中的一个文档。

{"po": [{"pred": "病因", "obj": "由SARS-Cov-2冠状病毒引起，因为人群缺少对新型病毒株SARS-Cov-2的免疫力，所以人群普遍易感。WHO将SARS-Cov-2感染导致的疾病命名为COVID-19，其中多数感染可以导致肺炎，就称之为新型冠状病毒肺炎/新冠肺炎。"}, {"pred": "病原学特点", "obj": "新型冠状病毒属于β属的新型冠状病毒，有包膜，颗粒呈圆形或椭圆形，常为多形性，直径60～140nm，其基因特征与SARSr-CoV和MERSr-CoV有明显区别。目前研究显示与蝙蝠SARSr样冠状病毒同源性高达85%以上，体外分离培养时，SARS-Cov-296小时左右即可在人呼吸道上皮细胞内发现，而在Vero_E6和Huh-7细胞系中分离培养约需6天。"}, {"pred": "出现症状的应对措施", "obj": "如果出现与COVID-19有关的症状，请立即就医。可能的话，先打电话进行咨询，并提供有关原先已有疾病和正在服用药物的情况。遵循卫生保健工作者的指示并定期监视你的症状。 如果呼吸困难，请立即与急救机构联系，因为这可能是呼吸道感染所致。如有可能，请先打电话联系，以了解下一步要怎么做。如果与他人合住，保证在怀疑自己感染后立即使用事先确定的空间进行自我隔离。你和其他住户成员也应尽可能佩戴医用口罩。点击这里，了解如何佩戴口罩。 如果与他人合住，并且卫生保健工作者建议采用COVID-19家庭护理方式，则其他住户成员应遵循现行的关于COVID-19轻症患者家庭护理和接触者管理的指导意见。如果独居，并且卫生保健工作者建议采用COVID-19家庭护理方式，则请家人、朋友、邻居、卫生保健工作者或当地志愿组织定期查看你的情况，并遵照现行的照护人员指导意见在需要时提供支持。"}], "subj": "新冠肺炎"}

2.4 属性同义词扩展

因为实验的数据集较小，包含的属性种类不多，因此可以人工增加一些同义的属性词。下面的文件中每一行的第一个词为数据中存在的属性，后面的为后来添加的同义的属性词。在解析查询语句的时候，如遇到同义的属性词，可将其映射到数据集中存在的属性上。

病因	由什么引起				
病原学特点	特点				
出现症状的应对措施	出现症状怎么办	出现症状如何应对			
传播方式	如何传播	怎么传播	传播途径		
传染源	从哪里传染	传染的起点			
存活时间	存活多久	活多久	多久失活		
定义	是什么				
感染后是否一定得肺炎	感染后一定会得肺炎	感染后会得肺炎	感染后一定得肺炎		
和流感病毒的异同	和流感病毒有什么不同	和流感病毒不同在哪里	和流感病毒一样吗	是不是流感病毒	

3. 导入 Elasticsearch

3.1 数据准备

导入 ElasticSearch 所使用的数据集为一个基于丁香园、世界卫生组织处理得到的三元组数据集，每个三元组描述了依据（实体名：属性名：属性值）处理的一问一答数据集。在将此数据集导入 elasticsearch 之前，需要考虑其在 elasticsearch 中存储的方式。

- 其中，所有属性除了"确诊人数"这个属性之外，都存在一个名为"po"的 list 对象中，每个属性及其属性值作为一个小的 object，分别用键"pred"和"obj"来标识属性名和属性值。
- 之所以要将"确诊人数"单独考虑，而不是和其它属性一样也存储在 list 中，是因为这个属性要支持范围搜索，即"确诊人数>1000"这样的搜索，因此要求它们在存储时的数据类型为 integer，而 list 中的所有属性的属性值的存储类型都为 keyword(不分词的 string，只支持全文匹配)。
- 之所以每一对(属性名，属性值)存储为一个 object，并放入一个 list 中，是因为这是 elasticsearch 定义的一种 nested object 的数据类型，这种数据类型能存储大量拥有相同的 key 的对象，并且可以对之进行有效的检索。这样，不论数据集中有多少种种类不同的属性，都可以以相同的格式存储。
- 之所以不是每一个三元组存储为一篇文档，而是一个实体相关的所有属性及属性值存储为一篇文档，是因为要支持通过多对(属性，属性值)联合检索满足要求的实体，以这种格式存储，能提高检索效率。

3.2 导入数据

我们通过命令行，使用 Elasticsearch 中的 Bulk 进行批量导入，将特定处理好的数据导入到 ElasticSearch 中，具体格式如下例。

{"index": {"_index":"demo"}}↵
{"po": [{"pred": "病因", "obj": "由 SARS-Cov-2 冠状病毒引起，因为人群缺少对新型病毒株 SARS-Cov-2 的免疫力，所以人群普遍易感。WHO 将 SARS-Cov-2 感染导致的疾病命名为 COVID-19，其中多数感染可以导致肺炎，就称之为新型冠状病毒肺炎/新冠肺炎。"}, {"pred": "病原学特点", "obj": "新型冠状病毒属于 β 属的新型冠状病毒，有包膜，颗粒呈圆形或椭圆形，常为多形性，直径 60～140nm，其基因特征与 SARSr-CoV 和 MERSr-CoV 有明显区别。目前研究显示与蝙蝠 SARSr 样冠状病毒同源性高达 85%以上，体外分离培养时，SARS-Cov-296 小时左右即可在人呼吸道上皮细胞内发现，而在 Vero_E6 和 Huh-7 细胞系中分离培养约需 6 天。"}, {"pred": "出现症状的应对措施", "obj": "如果出现与 COVID-19 有关的症状，请立即就医。可能的话，先打电话进行咨询，并提供有关原先已有疾病和正在服用药物的情况。遵循卫生保健工作者的指示并定期监视你的症状。 如果呼吸困难，请立即与急救机构联系，因为这可能是呼吸道感染所致。如有可能，请先打电话联系，以了解下一步要怎么做。如果与他人合住，保证在怀疑自己感染后立即使用事先确定的空间进行自我隔离。你和其他住户成员也应尽可能佩戴医用口罩。点击这里，了解如何佩戴口罩。 如果与他人合住，并且卫生保健工作者建议采用 COVID-19 家庭护理方式，则其他住户成员应遵循现行的关于 COVID-19 轻症患者家庭护理和接触者管理的指导意见。如果独居，并且卫生保健工作者建议采用 COVID-19 家庭护理方式，则请家人、朋友、邻居、卫生保健工作者或当地志愿组织定期查看你的情况，并遵照现行的照护人员指导意见在需要时提供支持。"}],
"subj": "新冠肺炎"}↵

4. 自然语言转化为 Logical form

4.1 解析自然语言

首先，对于输入的自然语言问句，识别出其中出现在知识库中的实体名，属性名以及属性值识别的流程如下：

（1） 分词。分词工具例如 jieba 分词支持自定义词典，可以将分词算法的词典替换成所有实体名，根据实体名对自然语言问句进行分词处理。

```
input="感染新冠肺炎后应该怎么办"
# 分词
q_sep=list(jieba.cut(input))
q_sep
```

（2） 从知识库中匹配属性。对于属性，由于种类较少，直接用字典记录知识库中的所有属性，用匹配的方法找出所有属性名。

```
# 识别属性名
rel_match=[]
for w in q_sep:
    if w in rel_set:
        rel_match.extend([w])
    rel_match.extend([k for k, v in attr_dict.items() if w in v])
rel_match
```

（3） 从知识库中匹配实体名。

```
# 识别实体名
entity_match=[]
for w in q_sep:
    if w in entity_set:
        entity_match.append(w)
entity_match
```

（4） 从知识库中匹配属性值。

```
# 识别属性值
attr_value_match=[]
for w in q_sep:
    if w in attr_value_set:
        attr_value_match.append(w)
attr_value_match
```

4.2 生成 Logical form

在识别出查询中所有的实体名，属性名和属性值后，依据它们的数目及位置，确定查询的类型，以便映射到的对应的 logical form。

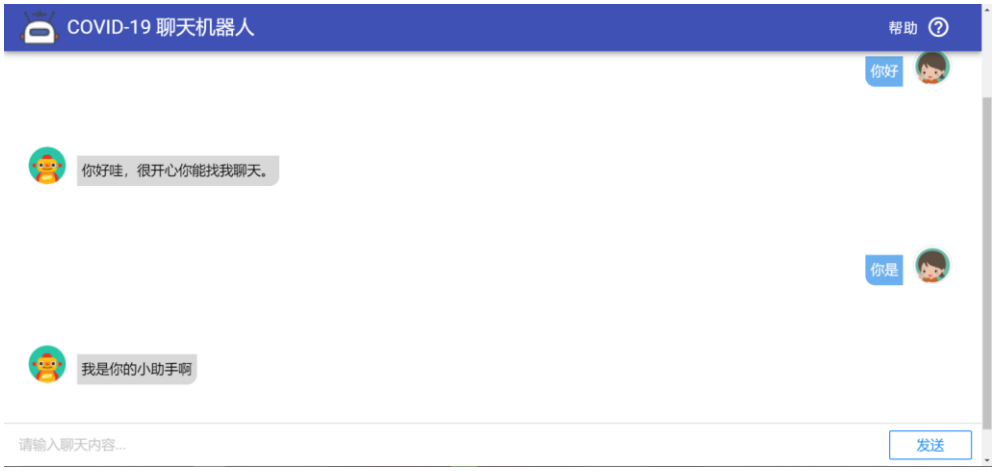
```
# logical form
# 实体检索 e.g.:新冠病毒
if len(rel_match)==0: # 如果没有属性名，则查询的是实体
    lf=entity_match[0]
# 实体的属性检索 e.g.:新冠病毒的定义
# 实体属性的多跳检索 e.g.:广东的省会的感染人数
elif len(entity_match)!=0: # 如果含有属性名和实体名，则是属性查询
    lf=entity_match[0]+':'+':'.join(rel_match)
# 多种属性条件检索实体 e.g.:累计确诊人数超过1000人的中国省份或美国州
# 如果含有属性名、不含实体名，则是根据属性值检索实体
elif len(entity_match)==0:
    op=None
    if '超过' in q_sep or '大于' in q_sep or '多于' in q_sep or '以上' in q_sep:
        op='>'
    elif '小于' in q_sep or '少于' in q_sep or '以下' in q_sep:
        op='<'
    elif '不超过' in q_sep or '不大于' in q_sep or '不多于' in q_sep:
        op='<='
    elif '不小于' in q_sep or '不少于' in q_sep:
        op='>='
    num=re.findall(r"\d+",input)[0]
    if len(attr_value_match)!=0:
        attr_value_match=['所属:'+i for i in attr_value_match]
        attr_value_match[1:]=['OR'+i for i in attr_value_match[1:]]
        lf=rel_match[0]+op+num+'AND'+''.join(attr_value_match)
```

5. Logical form 翻译成 ES 查询语句

```
    #自然语句转换成logical Form
    def ChatBot(self, sentence):
        q_sep=list(jieba.cut(sentence))
        # 识别属性名
        rel_match=[]
        es = Elasticsearch()
        for w in q_sep:
            if w in self.rel_set:
                rel_match.extend([w])
            rel_match.extend([k for k, v in self.attr_dict.items() if w in v])
        # 识别实体名
        entity_match=[]
        for w in q_sep:
            if w in self.entity_set:
                entity_match.append(w)
        # 识别属性值
        attr_value_match=[]
        for w in q_sep:
            if w in self.attr_value_set:
                attr_value_match.append(w)
        lf=None
        # logical form
        # 实体检索 e.g.:新冠病毒
        if len(rel_match)==0: # 如果没有属性名，则查询的是实体
            query = []
            lf=entity_match[0]
            query = json.dumps({"query": { "bool":{"filter":{"term" :{"subj" : lf}}}}},ensure_ascii=False)
            result = es.search(index='demo', body=query)
            #解析Json
            ans = ''
            for e in result['hits']['hits']:
                subj = e['_source']['subj']
                pred = e['_source']['po'][0]['pred']
                name = subj + pred
                ans = ans + name + ":" + e['_source']['po'][0]['obj'] + '\n'
            ans = ans[:-1]
        # 实体的属性检索 e.g.:新冠病毒的定义
        # 实体属性的多跳检索 e.g.:广东的省会的感染人数
        elif len(entity_match)!=0: # 如果含有属性名和实体名，则是属性查询
            lf=entity_match[0]+':'+':'.join(rel_match)
            divide = lf.replace(" ","").split(':')
            for i in range(len(divide)-1):
                if i == 0:
                    query = []
                    entity = divide[i]
                    pred = divide[i+1]
```

6. 添加腾讯智能闲聊机器人

为了保证问答机器人对答的流畅性，我们加入了腾讯智能闲聊机器人的 API。当搜索不到问题的答案时，则调用闲聊机器人 API。



7. 可视化

我们使用 Vue 框架编写前端。用户将问题输入聊天框，点击“发送”按钮或回车键，将问题发送给机器人，显示在对话框右侧。机器人把问题的回答返回，显示在对话框左侧。

8. 人员分工

项目整体规划(何梵)

项目进度与方向把控

算法代码交互与整合

数据处理(何梵、李洁滢)

数据爬取

数据清洗

属性同义词扩展

Neo4j 格式转换

数据导入 ElasticSearch

自然语言处理(李洁滢)

分词

实体与属性提取

模糊匹配

逻辑语句转换(陈奕凯)

将 Logical Form 转换为 DSL

实体检索

属性检索

多跳查询

根据属性值查询实体

前端界面(李洁滢)

机器人整体界面编写

后端接口(陈奕凯、李洁滢)

知识问答机器人接口制作

腾讯闲聊机器人接口

前后端交互

9. 项目点评与展望

李洁滢

我们小组的智能问答机器人利用 Elasticsearch 实现了实体检索、实体的属性检索、多跳查询、根据属性值查询实体四种查询逻辑，基本可以覆盖所有关于新型冠状病毒的提问形式。同时采用了同义词库实现不同问法到标准问法的映射（模糊匹配），调用腾讯智能闲聊机器人接口使机器人在匹配不到问题答案的时候也能作出合理回应，提升了聊天的流畅性。然而，我们的项目也存在以下几点有待改进的地方：

1. 数据量过小。由于我们只从丁香园、世界卫生组织的官方网站上爬取了一百多对问答，经人工和实体、属性标注后只剩下几十个三元组，数据量十分稀少，只能回答若干的人们关于新冠肺炎最常问的问题。关于新冠疫情感染人数的统计数据也是静态的，不能准确地回答实时的疫情数据。如果我们采用自动的数据采集方法，采集实时、动态的疫情知识，能够提升智能问答机器人的全面性和准确度。

2. 模糊匹配能力有限。采用同义词词库映射的方法只适用于数据量不大、问法变式不多的情况。当我们的知识数据库变得更庞大时，采用 N-Gram 等算法计算字符串举例进行问句的模糊匹配将是更合理的选择。

3. 只能实现单轮对话。机器人只能回答“如果感染了新冠肺炎怎么办？"这种单轮对话而不能回答“感染新冠肺炎有什么症状？""我该怎么办？"这一系列有递进关系的问题，不够贴近人类的真实对话。我们可以引入对话跟踪技术根据多轮的对话逐步确定用户当前的目标，从而确保对话的连贯性。

陈奕凯

我们小组的实现的聊天机器人功能是比较完善的，目前主要能进行答复的问题为新冠相关的科普性知识以及中美两国主要城市的感染人数，同时语料库所没有的答案由闲聊借口也能回复一二。缺点在于目前可回答问题数量较少，以及暂时不能够进行多轮对话。进一步的改进措施有两：在做语句 NLP 时，使用 text similarity 的方法来进行相近词意的匹配。扩展语料库。

何梵

我们小组在开展聊天机器人这个项目的过程中，方向和分工都比较明确，我们先从最基础的功能、最简单的问答开始实现，然后再不断地扩展新功能。知识问答机器人基本上可以覆盖我们日常的一些关于新冠肺炎疫情的基本问题，较好地作出合理的答案。

在项目的实现中，我们着重考虑的是机器人背后的问答逻辑，争取做到机器人内部逻辑完善，我们最终通过实体检索、实体的属性检索、多跳查询、根据属性值查询实现对内部逻辑进行实现，基本上完成了问答逻辑的全覆盖，但由于数据量的原因，导致我们知识问答机器人看起来并不出色。但如果背后支持的数据量非常庞大，我认为我们的问答机器人工作也可以做得非常出色。

对于内部逻辑实现，徐博士在知识图谱小组的答辩环节提到是否能够进行多次关系的跳转，即类似于“小明的母亲的身高”这种查询方式；在上一组知识问答机器人的答辩环节提到是否能够进行多属性查询，即类似于“钟南山和李兰娟的职业”这种查询方式。当时这两个小组分别对这两种提到的功能都没有进行实现，但是我们通过多条查询和多属性查询，实现了这两种功能，还是非常开心的。

10. 项目代码

https://github.com/Hefan-scut/COVID-19-Chatbot/