医药问答系统

本项目是一个基于python和neo4j图数据库的医药类知识图谱问答系统。

运行环境说明

本项目在以下环境中进行过测试：

```
系统：window 10
python版本：python 3.8
IDE：PyCharm
数据库：neo4j
```

```
系统：Ubuntu 20.04LTS - WSL2
python版本：python 3.8
数据库：neo4j
```

所需依赖

```
ahocorasick 1.4.2 (python中包名为pyahocorasick)
py2neo 2021.1.5 (py2neo旧版本可能导致neo4j连接失败)
```

运行方式

1、下载neo4j相关组件，新建数据库，修改`build_medicalgraph.py`和`answer_search.py`中以下语句的账号密码（初始账号和密码均为neo4j）；

```
self.g = Graph("http://localhost:7474", auth=("neo4j","123456"))
```

2、运行程序`build_medicalgraph.py`，构建图数据库

注：原始数据保存在`data/medical.json`中，构建时若是使用原始数据则这一步耗时为几个小时，可以使用从原始数据中截取的部分数据`data_medical1.json`，耗时几分钟。更改`build_medicalgraph.py`中的以下语句即可更改数据源：

```
self.data_path = os.path.join(cur_dir, 'data/medical.json')
```

3、运行主程序`chat_graph.py`

项目说明

项目结构图

```
C:.
|  answer_search.py
|  build_medicalgraph.py
|  chatbot_graph.py
|  question_classifier.py
|  question_parser.py
|  README.md
|
└─data
        medical.json
        medical1.json
```

`build_medicalgraph.py` 连接数据库，构建知识图谱

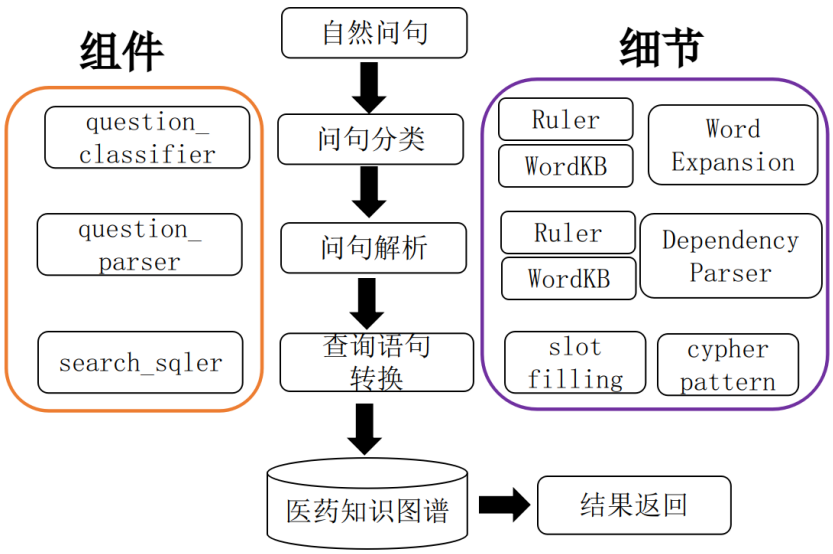
`question_parser.py` 构建实体节点，解析主函数

`question_classifier.py` 构建特征词词典

`answer_search.py` 执行查询，调用相应模板

`chatbot_graph.py` 主程序

程序结构图：



基于知识图谱的问答框架

数据处理

原始数据格式为`json`格式，数据样式如下：

```
{ "_id" : { "$oid" : "5bb578b6831b973a137e3ee6" }, "name" : "肺泡蛋白质沉积症", "desc" : "肺泡蛋白质沉积症(简称PAP)，又称Rosen-Castle-man-Liebow综合征，是一种罕见疾病。该病以肺泡和细
{ "_id" : { "$oid" : "5bb578b6831b973a137e3ee7" }, "name" : "百日咳", "desc" : "百日咳(pertussis，whoopingcough)是由百日咳杆菌所致的急性呼吸道传染病。其特征为阵发性痉挛性咳嗽，咳嗽末
{ "_id" : { "$oid" : "5bb578b6831b973a137e3ee8" }, "name" : "苯中毒", "desc" : "苯(benzene)是从煤焦油分馏及石油裂解所得的一种芳香烃化合物，系无色有芳香气味的油状液体。挥发基速，易燃易
{ "_id" : { "$oid" : "5bb578b6831b973a137e3ee9" }, "name" : "喘息样支气管炎", "desc" : "喘息样支气管炎(asthmatoidbronchitis)又称哮喘性支气管炎，泛指一组喘息表现的婴幼儿急性支气管炎，
{ "_id" : { "$oid" : "5bb578b6831b973a137e3eea" }, "name" : "成人呼吸窘迫综合征", "desc" : "成人呼吸窘迫综合征简称ARDS，是一种继发的，以急性呼吸窘迫和低氧血症为特征的综合征。又称休克
{ "_id" : { "$oid" : "5bb578b6831b973a137e3eeb" }, "name" : "大量羊水吸入", "desc" : "胎儿在宫内或分娩过程中吸入较大量羊水称大量羊水吸入(massiveaminoticfluidaspiration)，又称羊水吸入
{ "_id" : { "$oid" : "5bb578b6831b973a137e3eec" }, "name" : "单纯性肺嗜酸粒细胞浸润症", "desc" : "单纯性肺嗜酸粒细胞浸润症，又名吕弗琉综合征，是吕弗琉于1932年首先描述本病。单纯性肺嗜
{ "_id" : { "$oid" : "5bb578b6831b973a137e3eed" }, "name" : "大叶性肺炎", "desc" : "大叶性肺炎(lobarpneumonia)，又名肺炎球菌肺炎，是由肺炎双球菌等细菌感染引起的呈大叶性分布的急性肺实
{ "_id" : { "$oid" : "5bb578b6831b973a137e3eee" }, "name" : "大楼病综合征", "desc" : "大楼病综合征有多种表现，均因接触各种有害物质所致。好发于办公室内工作的人群，或人员密集的大楼内工
{ "_id" : { "$oid" : "5bb578b6831b973a137e3eef" }, "name" : "二硫化碳中毒", "desc" : "二硫化碳(carbondisulfide，CS2)是工业上应用广泛的化学溶剂，也用于粘胶纤维、四氯化碳，农药生产等。
{ "_id" : { "$oid" : "5bb578b6831b973a137e3ef0" }, "name" : "肺-胸膜阿米巴病", "desc" : "肺-胸膜阿米巴病是溶组织阿米巴原虫感染所致的肺及胸膜化脓性炎症，肝原性病变多发生在右下肺，血源
{ "_id" : { "$oid" : "5bb578b6831b973a137e3ef1" }, "name" : "肺出血-肾炎综合征", "desc" : "肺出血-肾炎综合征，又称抗基膜性肾小球肾炎，Goodpasture综合征或Goodpasture病，可能系病毒感染
{ "_id" : { "$oid" : "5bb578b6831b973a137e3ef2" }, "name" : "肺放线菌病", "desc" : "肺放线菌病(pulmonaryactinomycosis)是由厌氧的以色列放线菌感染肺部引起的慢性化脓性肉芽肿性疾病，病变
{ "_id" : { "$oid" : "5bb578b6831b973a137e3ef3" }, "name" : "肺泡蛋白沉着症", "desc" : "肺泡蛋白沉着症(Pulmonaryalveolarproteinosis，PAP)是一种原因未明的少见疾病。其特点是肺泡内或细
{ "_id" : { "$oid" : "5bb578b6831b973a137e3ef4" }, "name" : "肺曲菌病", "desc" : "肺曲菌病(pulmonaryaspergillosis)致病菌主要为烟曲菌，少数为黄曲菌、土曲菌、黑曲菌，棒状曲菌、构巢曲菌
{ "_id" : { "$oid" : "5bb578b6831b973a137e3ef5" }, "name" : "放射性肺炎", "desc" : "放射性肺炎(radiationpneumonitis)系由于肺癌，乳腺癌，食管癌，恶性淋巴瘤或胸部其他恶性肿瘤经放射治疗
{ "_id" : { "$oid" : "5bb578b6831b973a137e3ef6" }, "name" : "肺念珠菌病", "desc" : "肺念珠菌病是一种常见的肺真菌病，由念球菌属(主要是白念珠菌)感染所致。本病多为继发性感染，在人体抵抗
{ "_id" : { "$oid" : "5bb578b6831b973a137e3ef7" }, "name" : "肺大疱", "desc" : "肺大疱是指由于各种原因导致肺泡腔内压力升高，肺泡壁破裂，互相融合，在肺组织形成的含气囊腔。直径超过1厘米
{ "_id" : { "$oid" : "5bb578b6831b973a137e3ef8" }, "name" : "肺炎球菌肺炎", "desc" : "肺炎球菌肺炎是由肺炎链球菌所引起，占社区获得性肺炎中的半数以上。起病急，有寒战、高热、胸痛、咳嗽
{ "_id" : { "$oid" : "5bb578b6831b973a137e3ef9" }, "name" : "肺气肿", "desc" : "肺气肿是指终末细支气管远端(呼吸细支气管，肺泡管、肺泡囊和肺泡)的气道弹性减退，过度膨胀，充气和肺容积增
{ "_id" : { "$oid" : "5bb578b6831b973a137e3efa" }, "name" : "肺炎杆菌肺炎", "desc" : "肺炎杆菌肺炎(克雷白杆菌肺炎)，克雷白杆菌肺炎是由肺炎克雷白杆菌引起的急性肺部炎症，多见于老年、营
{ "_id" : { "$oid" : "5bb578b6831b973a137e3efb" }, "name" : "肺脓肿", "desc" : "肺脓肿(lungabscess)是由于多种病因所引起的肺组织化脓性病变，早期为化脓性炎症，继而坏死形成脓肿。多发生于
{ "_id" : { "$oid" : "5bb578b6831b973a137e3efc" }, "name" : "肺栓塞", "desc" : "肺栓塞(pulmonaryembolism)是指栓塞物进入肺动脉及其分支，阻断组织血液供应所引起的病理和临床状态。常见的栓
{ "_id" : { "$oid" : "5bb578b6831b973a137e3efd" }, "name" : "肺炎", "desc" : "肺炎是指终末气道，肺泡和肺间质的炎症，可由疾病微生物、理化因素，免疫损伤、过敏及药物所致。细菌性肺炎是最
{ "_id" : { "$oid" : "5bb578b6831b973a137e3efe" }, "name" : "肺大泡", "desc" : "肺大泡是一个病理学名词，是指由于各种原因导致肺泡腔内压力升高，肺泡壁破裂，互相融合，在肺组织形成的含气
{ "_id" : { "$oid" : "5bb578b6831b973a137e3eff" }, "name" : "肺水肿", "desc" : "肺水肿(pulmonaryedema)是指由于某种原因引起肺内组织液的生成和回流平衡失调，使大量组织液在很短时间内不能
{ "_id" : { "$oid" : "5bb578b6831b973a137e3f00" }, "name" : "非典", "desc" : "传染性非典型肺炎(严重急性呼吸综合征，SARS)的简称，是由SARS冠状病毒(SARS-CoV)引起的一种具有明显传染性、可
{ "_id" : { "$oid" : "5bb578b6831b973a137e3f01" }, "name" : "肺转移瘤", "desc" : "肺转移瘤是指原发于身体其它部位的恶性肿瘤经血道或淋巴道转移到肺。据统计在死于恶性肿瘤的病例中约20～30
{ "_id" : { "$oid" : "5bb578b6831b973a137e3f02" }, "name" : "肺炎性假瘤", "desc" : "肺炎性假瘤是肺内良性肿块，是一种肺实质非特异性炎性增生性肿瘤样病变，是由肺内慢性炎症产生的肉芽肿、
{ "_id" : { "$oid" : "5bb578b6831b973a137e3f03" }, "name" : "肺隐球菌病", "desc" : "肺隐球菌病(pulmonarycryptococcosis)为新型隐球菌感染引起的亚急性或慢性内脏真菌病，此菌属腐生酵母菌，
{ "_id" : { "$oid" : "5bb578b6831b973a137e3f04" }, "name" : "肺癌", "desc" : "肺癌发生于支气管粘膜上皮，亦称支气管癌。近50年来许多国家都报道肺癌的发病率明显增高，在男性癌瘤病人中，肺
{ "_id" : { "$oid" : "5bb578b6831b973a137e3f05" }, "name" : "镉中毒", "desc" : "镉中毒(cadmiumpoisoning):金属镉微毒，硫化镉、硒磺酸镉属低毒，氧化镉、硫酸镉、氯化镉、硝酸镉等属中等毒性
{ "_id" : { "$oid" : "5bb578b6831b973a137e3f06" }, "name" : "感冒", "desc" : "感冒，总体上分为普通感冒和流行性感冒，在这里先讨论普通感冒。普通感冒，祖国医学称\"伤风\"，是由多种病毒引
{ "_id" : { "$oid" : "5bb578b6831b973a137e3f07" }, "name" : "感染性休克", "desc" : "感染性休克(septicshock)，亦称脓毒性休克，是指由微生物及其毒素等代谢产物所引起的脓毒性休克，感染灶中
{ "_id" : { "$oid" : "5bb578b6831b973a137e3f08" }, "name" : "过敏性休克", "desc" : "过敏性休克是由于一般对人体无害的特异性过敏源作用于过敏体质的病人，导致以急性周围循环灌注不足为主的
{ "_id" : { "$oid" : "5bb578b6831b973a137e3f09" }, "name" : "Goodpasture综合征", "desc" : "Goodpasture综合征，又称抗基膜性肾小球肾炎，肺出血-肾炎综合征或Goodpasture病，可能系病毒感染
{ "_id" : { "$oid" : "5bb578b6831b973a137e3f0a" }, "name" : "过敏性肺炎", "desc" : "过敏性肺炎(hypersensitivity\npneumonitis)是一组由不同致敏原引起的非哮喘性变应性肺疾患，以弥漫性间质
{ "_id" : { "$oid" : "5bb578b6831b973a137e3f0b" }, "name" : "汞中毒", "desc" : "汞为白色液态金属，常温下易蒸发，汞中毒(mercury\npoisoning)以慢性为多见，主要发生在生产活动中，主要以蒸
{ "_id" : { "$oid" : "5bb578b6831b973a137e3f0c" }, "name" : "呼吸道异物", "desc" : "呼吸道异物(foreignbodyinrespiratory\ntract)系指喉、气管和支气管异物，是耳鼻咽喉科急危症之一，直接威
{ "_id" : { "$oid" : "5bb578b6831b973a137e3f0d" }, "name" : "呼吸性细支气管相关的间质性肺疾病", "desc" : "呼吸性细支气管相关的间质性肺疾病(respiratorybronchiolitis-associatedinterstiti
{ "_id" : { "$oid" : "5bb578b6831b973a137e3f0e" }, "name" : "呼吸衰竭", "desc" : "呼吸衰竭是由各种原因导致严重呼吸功能障碍，肺通气和(或)换气功能的严重障碍，以致不能进行有效的气体交换
{ "_id" : { "$oid" : "5bb578b6831b973a137e3f0f" }, "name" : "呼吸道合胞病毒肺炎", "desc" : "呼吸道合胞病毒肺炎(respiratorysyncytialviruspneumonia)简称合胞病毒肺炎，主要由呼吸道合胞病
{ "_id" : { "$oid" : "5bb578b6831b973a137e3f10" }, "name" : "减压病", "desc" : "减压症(Decompressionsickness，简称DCS)，俗称潜水夫病或沉箱病，是由于高压环境作业后减压不当，体内原已溶
{ "_id" : { "$oid" : "5bb578b6831b973a137e3f11" }, "name" : "急性呼吸衰竭", "desc" : "急性呼吸衰竭是指患者由于某种原因在短期内呼吸功能迅速失去代偿，出现严重缺氧和(或)呼吸性酸中毒者。
{ "_id" : { "$oid" : "5bb578b6831b973a137e3f12" }, "name" : "急性肺脓肿", "desc" : "急性肺脓肿是指由于多种病原菌引起的肺部化脓性感染，早期为肺组织的感染性炎症，继而坏死、液化、外周有
{ "_id" : { "$oid" : "5bb578b6831b973a137e3f13" }, "name" : "军团病", "desc" : "军团病是军团杆菌(Legionellosis)所引起的以肺炎为主的感染，又称为军团病(Legionelladisease)，病原菌主要来
```


使用python内置的json包来读取json文件

构建知识图谱

创建节点

```
'''建立节点'''
 def create_node(self, label, nodes):
        count = 0
        for node_name in nodes:
            node = Node(label, name=node_name)
            self.g.create(node)
            count += 1
            print(count, len(nodes))
        return

 '''创建知识图谱中心疾病的节点'''
def create_diseases_nodes(self, disease_infos):
    count = 0
    for disease_dict in disease_infos:
        node = Node("Disease", name=disease_dict['name'], desc=disease_dict['desc'],
                    prevent=disease_dict['prevent'] ,cause=disease_dict['cause'],
                    easy_get=disease_dict['easy_get'],cure_lasttime=disease_dict['cure_la
                    cure_department=disease_dict['cure_department']
                    ,cure_way=disease_dict['cure_way'] , cured_prob=disease_dict['cured_p
        self.g.create(node)
        count += 1
        print(count)
    return

'''创建知识图谱实体节点类型schema'''
def create_graphnodes(self):
```

```
    Drugs, Foods, Checks, Departments, Producers, Symptoms, Diseases, disease_infos,rels_
    self.create_diseases_nodes(disease_infos)
    self.create_node('Drug', Drugs)
    print(len(Drugs))
    self.create_node('Food', Foods)
    print(len(Foods))
    self.create_node('Check', Checks)
    print(len(Checks))
    self.create_node('Department', Departments)
    print(len(Departments))
    self.create_node('Producer', Producers)
    print(len(Producers))
    self.create_node('Symptom', Symptoms)
    return
```

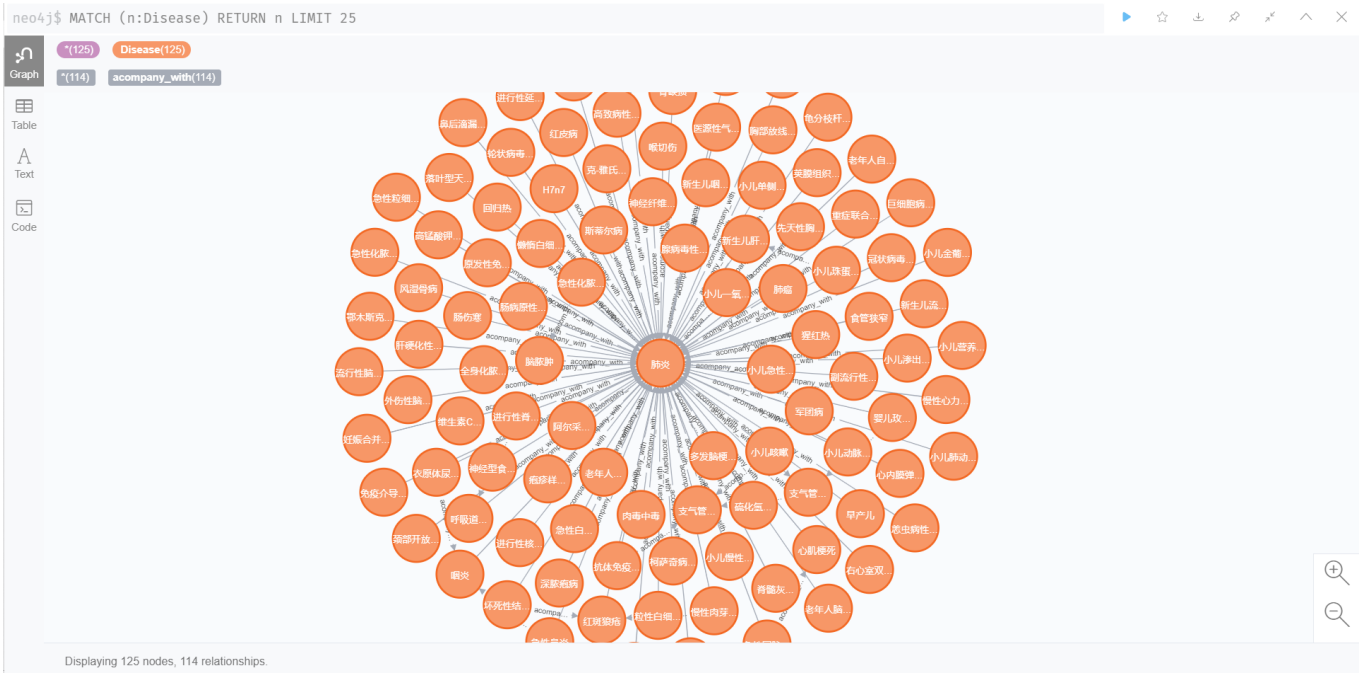
创建边

```
'''创建实体关联边'''
    def create_relationship(self, start_node, end_node, edges, rel_type, rel_name):
        count = 0
        # 去重处理
        set_edges = []
        for edge in edges:
            set_edges.append('###'.join(edge))
        all = len(set(set_edges))
        for edge in set(set_edges):
            edge = edge.split('###')
            p = edge[0]
            q = edge[1]
            query = "match(p:%s),(q:%s) where p.name='%s'and q.name='%s' create (p)-[rel
                start_node, end_node, p, q, rel_type, rel_name)
            try:
                self.g.run(query)
                count += 1
                print(rel_type, count, all)
            except Exception as e:
                print(e)
        return

 '''创建实体关系边'''
    def create_graphrels(self):
        Drugs, Foods, Checks, Departments, Producers, Symptoms, Diseases, disease_infos,
        self.create_relationship('Disease', 'Food', rels_recommandeat, 'recommand_eat',
        self.create_relationship('Disease', 'Food', rels_noteat, 'no_eat', '忌吃')
        self.create_relationship('Disease', 'Food', rels_doeat, 'do_eat', '宜吃')
        self.create_relationship('Department', 'Department', rels_department, 'belongs_to
        self.create_relationship('Disease', 'Drug', rels_commonddrug, 'common_drug', '常用
        self.create_relationship('Producer', 'Drug', rels_drug_producer, 'drugs_of', '生产
        self.create_relationship('Disease', 'Drug', rels_recommanddrug, 'recommand_drug',
        self.create_relationship('Disease', 'Check', rels_check, 'need_check', '诊断检查'
        self.create_relationship('Disease', 'Symptom', rels_symptom, 'has_symptom', '症状
```

```
        self.create_relationship('Disease', 'Disease', rels_acompany, 'acompany_with', '
        self.create_relationship('Disease', 'Department', rels_category, 'belongs_to', '
```

构建结果示例：



最终效果

```
model init finished ......
用户:你好
Hugo: 您好，我是Hugo医药智能助理，希望可以帮到您。祝您身体棒棒!
用户:我有病
Hugo: 您好，我是Hugo医药智能助理，希望可以帮到您。祝您身体棒棒!
用户:我感冒了
Hugo: 感冒,熟悉一下：感冒，总体上分为普通感冒和流行性感冒，在这里先讨论普通感冒。普通感冒，祖国医学称"伤风"，是由多种病毒引起的一种呼吸道常见病，
用户:肺炎是什么
Hugo: 肺炎,熟悉一下：肺炎是指终末气道，肺泡和肺间质的炎症，可由疾病微生物、理化因素，免疫损伤、过敏及药物所致。细菌性肺炎是最常见的肺炎，也是最常



用户:感冒怎么治疗
Hugo: 感冒可以尝试如下治疗：对症治疗;中医治疗;支持性治疗
用户:感冒吃什么药
Hugo: 感冒宜食的食物包括有：南瓜子仁;鹌鹑蛋;芝麻;鸡蛋
推荐食谱包括有：凉拌香椿;绿豆薏米饭;葱蒜粥;薏米莲子粥;赤小豆粥;姜丝萝卜汤;香椿芽粥;醋熘土豆丝
感冒通常的使用的药品包括：蒲公英颗粒；愈美胶囊；酚咖片；头孢丙烯分散片；伤风停胶囊；喉痛灵片；洛索洛芬钠胶囊；感冒灵颗粒；风油精；抗病毒口服液；利
用户:
```