1.实体类型

实体类型	中文含义	实体数量	举例
Disease	疾病	8792	感冒
Department	科室	54	内科
Check	检查项目	3342	血常规
Drug	治疗药品	1204	布洛芬片
Food	食物	4854	蜂蜜
Symptom	症状	6556	腹腔积血
Total	总计	24802	约2.5万实体

2.实体关系三元组

- <Disease,belongs_to,Department> 疾病所属科室
- <Disease,inspection_item,Check> 疾病检查项目
- <Disease,common_drug,Drug> 疾病常用药物
- <Disease,has_symptom,Symptom> 疾病症状
- <Disease,good_food,Food> 疾病宜吃食物
- <Disease,avoid_food,Food> 疾病忌吃食物
- <Disease,recommand_recipes,Food> 推荐食谱
- <Disease,has_complication,Symptom> 疾病并发症

3.实体关系信息

实体关系类型	中文含义	关系数量	举例
belongs_to	属于	8784	<哮喘,belongs_to,内科>
common_drug	常用药物	13477	<小儿肺炎，common_drug，小儿肺热平胶囊>
good_food	宜吃食物	34221	<胸椎骨折,good_food,黑鱼>
avoid_food	忌吃食物	34215	<感冒,avoid_food,猪油>
check_item	检查项目	39098	<肾结石,check_item，尿液颜色>
recommand_recipes	推荐食谱	39663	<肝病,recommand_recipes,芝麻小米粥>
has_complication	并发症	19151	<痔疮,has_complication,直肠癌>
has_symptom	疾病症状	58398	<冠心病,has_symptom,心慌；呼吸困难；心力衰竭>
Total	总计	247,007	近25万实体关系

4.疾病（中心）节点介绍

属性类型	中文含义	举例

属性类型	中文含义	举例
name	疾病名称	感冒
desc	疾病描述	发热伴寒战；咽痛；流鼻涕
cause	疾病病因	当有受凉，淋雨，过度疲劳
prevent	预防措施	补充维生素E、维生素C
treat_cycle	治疗周期	7-14天
treat_way	治疗方式	感冒可以尝试如下治疗：对症治疗;中医治疗;支持性治疗
cure_prob	治愈概率	97%
susceptible_people	易感人群	无特定人群
medical_insurance	是否医保疾病	未知
transmission_way	传染方式	呼吸道传播
treat_cost	治疗费用	根据不同医院，收费标准不一致，市三甲医院约（500-1000元）
nursing	护理方法	日常护理xxx

5.可解决的问题类型

question_type	问题类型	举例
disease_symptom	已知疾病查看症状	小儿肺炎有什么症状
symptom_disease	已知症状查看可能的疾病	最近老是流鼻涕怎么办
disease_cause	疾病原因	总是失眠是什么原因
disease_complication	疾病并发症	感冒有哪些并发症
disease_drug	疾病常用药物	白内障一般吃什么药
drug_disease	药物能治疗啥疾病	阿莫西林胶囊能治疗啥
disease_avoid_food	疾病忌口	肝病不能吃什么
disease_good_food	疾病宜吃	肺结核吃什么好
food_avoid_disease	什么疾病不能吃的食物	什么人最好不要吃蜂蜜
food_good_disease	食物适合哪些人吃	腰果适合哪些人吃
disease_check	疾病检查项目	怎么查出来是不是脑膜炎
check_disease	已知检查找疾病	血常规能查出来啥病
disease_prevent	疾病预防方法	怎么样才能防止肾虚
disease_treat_way	疾病治疗方法	高血压要怎么治
disease_cure_prob	疾病治愈概率	肺结核能治好吗

question_type	问题类型	举例
disease_susceptible_people	疾病易感人群	什么人容易得高血压？
disease_department	疾病去哪个科室	痔疮属于哪个科室的
disease_treat_cost	疾病治疗费用	治疗肾结石要多少钱
disease_medical_insurance	某病是医保疾病吗	肾结石是医保疾病吗
disease_treat_cycle	某疾病的治疗周期	感冒要多久才能好
disease_desc	疾病概述	抑郁症

6.实现思路

- 图谱构建
    - 数据爬取
    - 数据预处理
    - 实体类型构建
    - 关系类型构建
    - 创建neo4j数据库
    - 知识图谱可视化
- 问答系统
    - 自然语言查询（Question Query）
    - 意图识别(Intention Recognition)
    - 实体识别（Entity Recognition）
    - 实体链接(Entity Linking)[目前未实现]
    - 查询语句构建(Query Construction)
    - 返回查询结果(Return Answering)

7.关键技术方法

- 数据获取
    - scrapy-spider ： 普通的页面解析或分布式爬虫（防止中断数据丢失）
    - 数据存储：redis、mongodb等
- 实体识别
    - 目前采用的方法
        - 根据提取的领域关键词，基于trie树构建快速查询AC Tree。
        - 用AC Tree对输入的自然语言问句,匹配潜在的关键词，作为候选实体
    - 其他方法弊端（针对此项目）
        - 因为缺乏大量领域监督数据，无法基于现有的热门方法BiLSTM+CRF训练模型进行识别。
        - HanLP等平台的工具基本都不提供实体识别方法，并且这些工具的训练语料很少涉及垂直领域知识。

- **意图识别:**
    - **我们采用的方法**
        - **基于词典模板的规则分类方法，意图对应项目中的question_type**
        - **该方法需要领域专家构建模板**
        - **缺乏大量的监督数据，相比以下其它方法，该方法是较好的选择**
    - **其他常用方法**
        - 基于词典模板的规则分类
        - 基于过往日志匹配（适用于搜索引擎）
        - 基于分类模型进行意图识别(CNN,RNN,MachineLearning)
- **实体链接:**
    - 缺乏大量监督数据，暂时未做实体连接、知识融合等操作
    - 可作为改进方向

8.实现工具

- 开发环境: Windows10、Python3.6
- 前端 ：jQuery、AJAX、Boottrap4、CSS3、HTML5
- 后端 : Flask、、Scrapy
- 数据库：Neo4j、MongoDB

9.已发现未解决的问题类型

- 老是头疼怎么办 (常用俗语，不能完全匹配实体)
- 失眠用什么药（缺少数据.数据源没有数据的问题）
- 感冒用什么药物 （缺少数据.数据源没有数据的问题）
- 什么样的人容易感冒？

10.项目改进方向：

- 实体链接：如何把俗语、常用于准确的链接到知识库中的实体？
- 知识库知识完善：现有的知识不完善，可以从其他数据源获取？
- 知识融合：从其他数据源获取的知识如何与已有知识库的知识进行有效的融合？