

《人工智能程序设计实践》实验大作业报告

大作业题目	基于寻医问药网站的知识图谱构建和可视化		类型	数据爬取和处理
班 号	2103601	学 号	2021112845	
所在院系	计算学部	学 期	22 年秋季学期	任课教师 关毅
实验类型	综合设计型			

实验目的：

- 掌握程序设计的基本算法和简单数据结构基础，能够综合运用基本控制语句、算法和数据结构，以及自顶向下、逐步求精的模块化设计方法，能够设计具有一定规模的系统级python语言程序，提高系统编程能力；
- 掌握常见的XPath方法从网页HTML代码中检索、爬取信息
- 针对复杂的数据类型以及较为庞大的数据，能够使用恰当的算法和数据结构，完成计算、统计分类、检索、匹配等相关的软件系统的构造、测试与实现；
- 掌握常用的程序调试和测试方法。

实验要求：

- 掌握网络爬虫的爬取数据方法；
- 掌握三元组数据格式的存储方法，以及txt、csv、json文件类型的读写操作
- 掌握图数据库（Neo4J）的数据导入以及可视化展示；

实验内容：

- 利用网络爬虫，按照疾病部位爬取寻医问药网站中的疾病相关介绍，以及相关应用知识。
- 将爬取结果存储为三元组格式
- 用图数据库（Neo4J）存储并展示知识图谱

实验环境：

操作系统： Windows11

集成开发环境： Pycharm

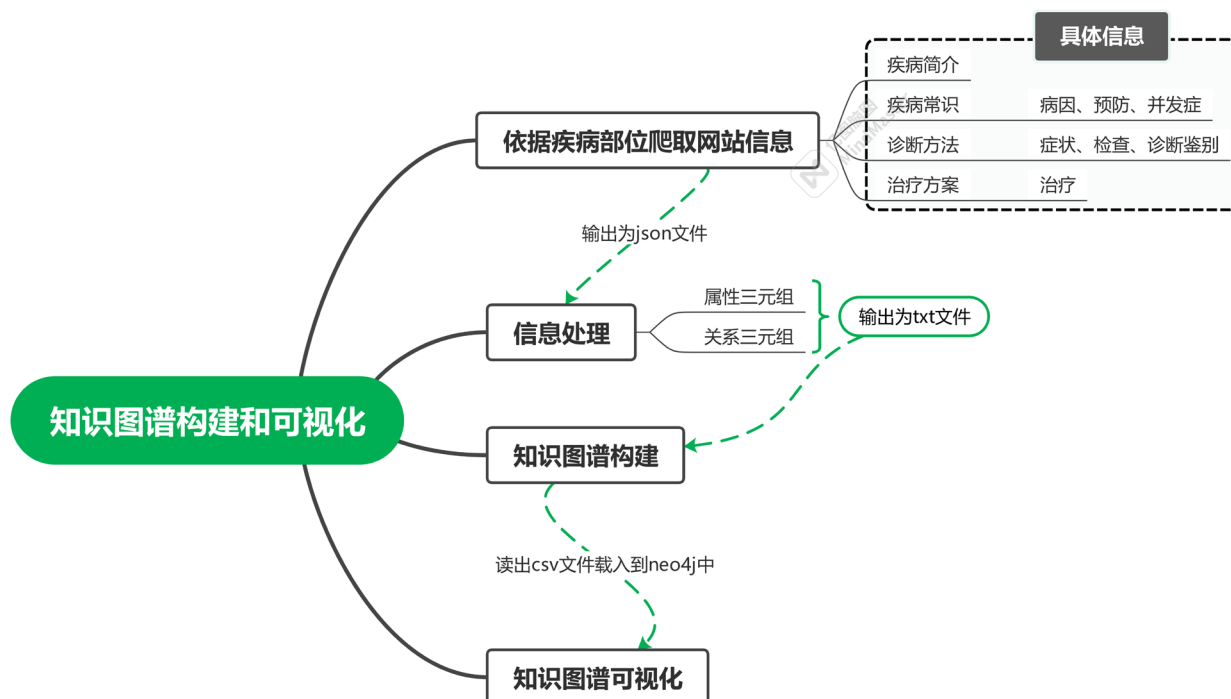
外部库： json、requests、re、lxml、pandas、csv

输入输出设计：

- get_json.py
无输入，输出为记录网站上具有索引结构的疾病信息json文件
- test_关系.py
输入get_json输出的json文件，输出关系三元组txt文件
- test_属性.py
输入get_json输出的json文件，输出属性三元组txt文件
- get_node_relation.py
输入关系三元组txt文件，输出为已搭建好的知识图谱的结点信息的csv文件

系统设计与实现:

1. 系统功能模块划分

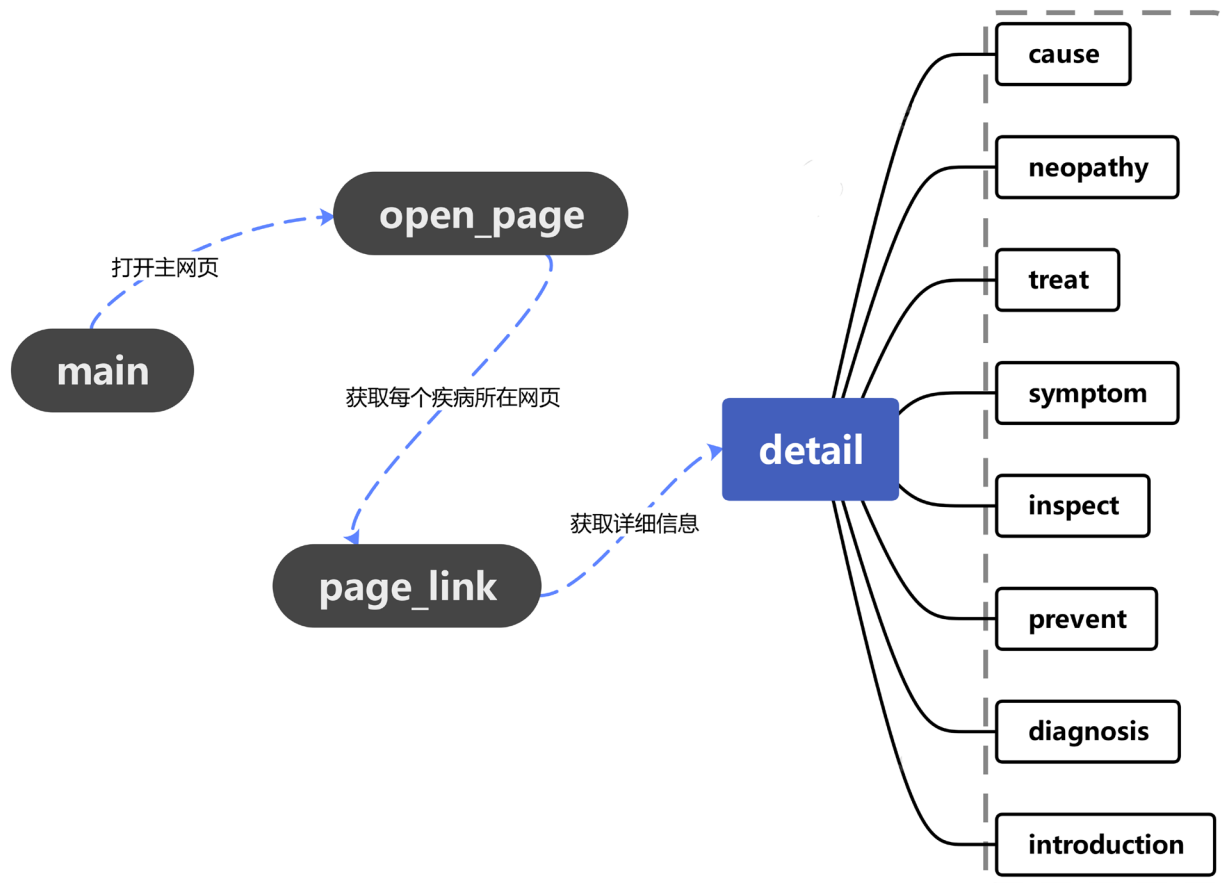


2. 函数功能和外部接口设计

◆ get_json.py ——爬取数据并记录为json文件

序号	函数名	函数功能	函数参数	函数返回值
1	open_page(url)	打开网页	网址url	无
2	introduction(url)	爬取疾病简介	网址url	简介(list)introduction
3	symptom(url)	爬取症状页面	网址url	常见症状(str)sum
4	inspect(url)	爬取检查页面	网址url	检查方式(str)sum
5	diagnosis(url)	爬取诊断鉴别页面	网址url	诊断手段(str)sum
6	treat(url)	爬取治疗页面	网址url	治疗手段(str)sum
7	cause(url)	爬取病因页面	网址url	病因(str)sum
8	prevent(url)	爬取预防页面	网址url	预防方法(str)sum
9	neopathy(url)	爬取并发症页面	网址url	并发症介绍(str)sum
10	extract(res,ser,modi)	正则化提取信息	目标字段 res,ser,modi	提取字段列表(list)
11	page_link(res,menu)	提取, 拼接网址	提取字段res, 目录menu	网址url
12	detail(link)	封装提取所有信息	疾病的特征数字	信息列表(list)content

各个函数之间的调用关系如下所示:



◆ **test_关系.py** ——读取数据为关系三元组，记录为txt文件

序号	函数名	函数功能	函数参数	函数返回值
1	my_split(data)	以空格分割文本	字符串data	分割后的列表list
2	name(data)	以冒号分割文本， 同时提取信息	字符串data	分割后的第一个字符串

◆ **test_属性.py** ——读取数据为属性三元组，记录为txt文件

◆ **get_node_relation.py** ——读取txt文件，记录为三元组形式的csv文件

3. 数据结构

列表：利用列表存储网址、疾病关键词等信息。

字典：按照层级关系以及对应关系将相关疾病的关键词键和对应的值储存在多层嵌套字典中。

三元组：依据关系和属性建立知识图谱所需要的三元组数据类型。

4. 算法

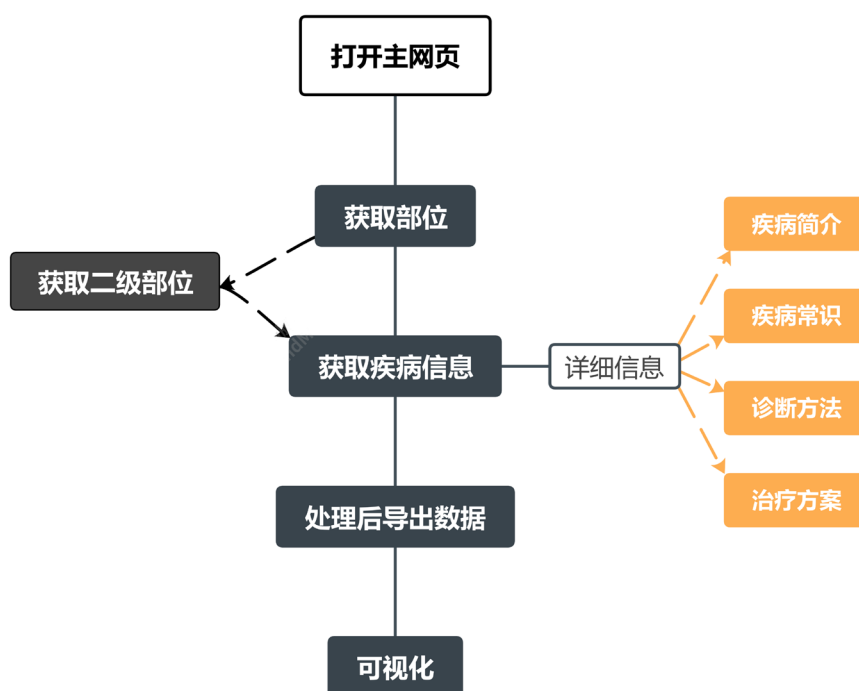
网络爬虫，使用 XPath 方法从网页 HTML 代码中检索、爬取信息。

循环遍历，多重循环遍历，读取、处理、统计爬取得到信息。

文件操作，读写 json、txt、csv 文件，搭建知识图谱。

可视化操作，利用 neo4j 导入文件达到可视化的效果。

5. 程序流程图



实验过程中遇到的问题及解决方法与思路:

问题 1: 爬取网站信息时出现乱码

原因: 编码方式不符合网站所使用的编码方式

解决方法: 将默认的 utf-8 编码方式更改为网站使用的 gb2312。

问题 2: 爬取信息时出现遗漏、信息不全的情况。

原因: 网站 HTML 代码目标目录下子标签不同。

解决方法: 使用代码 `content=ht.xpath('//div[@class="....."]').xpath('string().').strip()`, 得到某个标签下所有文本, 可以忽略 url。

问题 3: 进行字典键、值匹配对应读取时出现不对应的情况

原因: 某些关键词对应值为空。

解决方法: 增加是否为空的判断。

问题 4: 安装 neo4j 时配置不成功

原因: JDK 与 neo4j 版本不适配

解决方法: 下载 **JDK-15.0.2** 版本, 重新配置环境变量。

测试结果：

Json 文件：

🔍 疾病_relationship.json ×

C: > Users > ETERNITY > AppData > Local > Temp > BNZ.6311a9391357d3f > 🔍 疾病_relationship.json > 🔍 头部

📄 本文档包含许多非基本 ASCII unicode 字符 禁用非 ASCII 突出显示

1

{ "头部": { "鼻": { "鼻咽喉炎": { "医保疾病": "：", "否", "患病比例": "：", "0.01%-0.018%", "就诊科室": "：", "内科 呼吸内科", "治疗方式": "：", "药物治疗", "治疗周期": "：", "1-2月", "同医院, 收费标准不一致, 市三甲医院约 (500—1000元)", "头部脂溢性皮炎": { "医", "无传染性", "并发症": "：", "痤疮", "鳞皮炎", "酒渣鼻", "就诊科室": "：", "皮肤科性病", "药品": "：", "克林霉素甲硝唑搽剂", "曲咪新乳膏", "治疗费用": "：", "根据不同医院, 约", "0.012%", "易感人群": "：", "无特殊人群", "传染方式": "：", "无传染性", "并发症": "：", "期": "：", "根据不同医院, 收费标准不一致, 市三甲医院约 (1000—4000元)", "治愈率": { "医保疾病": "：", "否", "患病比例": "：", "0.001%", "易感人群": "：", "无特殊人群", "传", "：", "五官科 耳鼻喉科", "治疗方式": "：", "药物治疗 支持性治疗", "治疗周期": "：", "2", "同医院, 收费标准不一致, 市三甲医院约 (5000 — 10000元)", "鼻窦炎": { "医保疾", "并发症": "：", "头痛", "就诊科室": "：", "五官科 耳鼻喉科", "治疗方式": "：", "急性鼻", "品": "：", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": "：", "药物治疗 3000元左右, ", "：", "无特殊人群", "传染方式": "：", "无传染性", "并发症": "：", "肿胀", "就诊科室": "：", "常用药品": "：", "鼻渊丸", "通窍鼻炎片", "治疗费用": "：", "根据不同医院, 收费标", "：", "鼻", "无特定的人群", "传染方式": "：", "无传染性", "并发症": "：", "外耳道疖肿", "往来寒热", "就诊科室": "：", "五官科 耳鼻喉科", "治疗方式": "：", "药物治疗 康复治疗", "治疗周期": "：", "2-4", "周", "治愈率": "：", "95%", "常用药品": "：", "三黄片", "清热化毒丸", "治疗费用": "：", "根据不同医院, 收费标准不一致, 市三甲医院约 (1000—2000元)", "小儿鼻炎": { "医保疾病": "：", "否", "患病比例": "：", "临床常见病", "易感人群": "：", "儿童", "传染方式": "：", "飞沫传播, 有传染性", "并发症": "：", "急性化脓性中耳炎", "急性咽炎", "急性喉炎", "就诊科室": "：", "五官科 耳鼻喉科", "治疗方式": "：", "对症治疗 药物治疗 支持性治疗", "治疗周期": "：", "1-3个月", "治愈率": "：", "90%", "常用药品": "：", "注射用头孢唑林钠", "诺氟沙星胶囊", "治疗费用": "：", "根据不同医院, 收费标准不一致, 市三甲医院约 (3000—8000元)", "慢性鼻窦炎": { "医保疾病": "：", "否", "患病比例": "：", "0.01%-0.018%", "易感人群": "：", "无特定人群", "传染方式": "：", "与慢性刺激有关, 无传染性", "并发症": "：", "鼻囊炎", "咽喉炎", "就诊科室": "：", "五官科 耳鼻喉科", "治疗方式": "：", "药物治疗 对症治疗 支持性治疗", "治疗周期": "：", "1月", "治愈率": "：", "80%", "常用药品": "：", "慢严舒柠 咽炎片", "慢严舒柠 复方青橄榄利咽含片", "治疗费用": "：", "针对不同病因花费不同, 药物治疗约需3000元", "上颌窦炎": { "医保疾病": "：", "否", "患病比例": "：", "0.01%", "易感人群": "：", "无特定人群", "传染方式": "：", "无传染性", "并发症": "：", "头痛", "就诊科室": "：", "五官科 耳鼻喉科", "治疗方式": "：", "药物治疗 支持性治疗 物理治疗 手术", "治疗周期": "：", "2-6周", "治愈率": "：", "90%", "常用药品": "：", "头孢呋辛酯干混悬剂", "头孢呋辛酯胶囊", "治疗费用": "：", "根据不同医院, 收费标准不一致, 市三甲医院约 (2000—6000元)", "过敏性鼻炎": { "医保疾病": "：", "否", "患病比例": "：", "0.001%", "易感人群": "：", "无特定人群", "传染方式": "：", "鼻炎", "就诊科室": "：", "五官科 耳鼻喉科", "治疗方式": "：", "药物治疗 支持性治疗", "治疗周期": "：", "2-4周", "治愈率": "：", "98%", "常用药品": "：", "霍胆丸", "治疗费用": "：", "根据不同医院, 收费标准不一致, 市三甲医院约 (1000—5000元)", "鼻", "鼻囊炎": { "医保疾病": "：", "否", "患病比例": "：", "0.0001%", "易感人群": "：", "无特殊人群", "传染方式": "：", "无传染性", "并发症": "：", "头痛", "就诊科室": "：", "五官科 耳鼻喉科", "治疗方式": "：", "手术治疗 支持性治疗", "治疗周期": "：", "一周", "治愈率": "：", "95%", "常用药品": "：", "普菊片", "治疗费用": "：", "根据不同医院不同, 收费标准也不一致, 三甲甲等医院约 (1000—5000元)", "鼻", "血管瘤": { "医保疾病": "：", "否", "患病比例": "：", "0.025%", "易感人群": "：", "无特定的人群", "传染方式": "：", "无传染性", "并发症": "：", "鼻出血", "就诊科室": "：", "五官科 耳鼻喉科", "治疗方式": "：", "药物治疗 手术治疗", "治疗周期": "：", "2-4周", "治愈率": "：", "75%", "常用药品": "：", "注射用盐酸平阳霉素", "吡拉西坦葡萄糖注射液", "治疗费用": "：", "根据不同医院, 收费标准不一致", "鼻腔", "癌": { "医保疾病": "：", "否", "患病比例": "：", "0.003%", "易感人群": "：", "无特定的人群", "传染方式": "：", "无传染性", "并发症": "：", "鼻出血", "就诊科室": "：", "五官科 耳鼻喉科", "治疗方式": "：", "药物治疗 支持性治疗 手术治疗", "治疗周期": "：", "1-3个月", "治愈率": "：", "12%", "常用药品": "：", "鼻咽清毒颗粒", "转移因子口服溶液", "治疗费用": "：", "根据不同医院, 收费标准不一致, 市三甲医院约 (5000—30000元)", "鼻结核": { "医保疾病": "：", "否", "患病比例": "：", "0.005%", "易感人群": "：", "无特定人群", "传染方式": "：", "呼吸道传播", "并发症": "：", "鼻炎", "鼻出血", "就诊科室": "：", "传染科", "治疗方式": "：", "抗结核治疗, 对症治疗, 支持性治疗", "治疗周期": "：", "12个月以上", "治愈率": "：", "90%", "常用药品": "：", "勒马回囊膏", "白及颗粒", "治疗

致, 市三甲医院约 (1000 — 3000元) }}, {"颈部": { "反流性咽炎": { "医保疾病": " ", "否", "患病比例": " ", "0.01%-0.018%", "就诊科室": " ", "内科 呼吸内科", "治疗方式": " ", "药物治疗", "治疗周期": " ", "1-2月", "同医院, 收费标准不一致, 市三甲医院约 (500—1000元)", "头部脂溢性皮炎": { "医", "无传染性", "并发症": " ", "痤疮", "鳞皮炎", "酒渣鼻", "就诊科室": " ", "皮肤科性病", "药品": " ", "克林霉素甲硝唑搽剂", "曲咪新乳膏", "治疗费用": " ", "根据不同医院, 约", "0.012%", "易感人群": " ", "无特殊人群", "传", " ", "五官科 耳鼻喉科", "治疗方式": " ", "药物治疗 支持性治疗", "治疗周期": " ", "2", "同医院, 收费标准不一致, 市三甲医院约 (5000 — 10000元)", "鼻窦炎": { "医保疾", "并发症": " ", "头痛", "就诊科室": " ", "五官科 耳鼻喉科", "治疗方式": " ", "急性鼻", "品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 3000元左右, ", " ", "无特殊人群", "传染方式": " ", "无传染性", "并发症": " ", "肿胀", "就诊科室": " ", "常用药品": " ", "鼻渊丸", "通窍鼻炎片", "治疗费用": " ", "根据不同医院, 收费标", " ", "鼻", "无特定的人群", "传染方式": " ", "无传染性", "并发症": " ", "外耳道疖肿", "往来寒热", "就诊科室": " ", "五官科 耳鼻喉科", "治疗方式": " ", "药物治疗 康复治疗", "治疗周期": " ", "2-4", "周", "治愈率": " ", "95%", "常用药品": " ", "三黄片", "清热化毒丸", "治疗费用": " ", "根据不同医院, 收费标准不一致, 市三甲医院约 (1000—2000元)", "小儿鼻炎": { "医保疾病": " ", "否", "患病比例": " ", "临床常见病", "易感人群": " ", "儿童", "传染方式": " ", "飞沫传播, 有传染性", "并发症": " ", "急性化脓性中耳炎", "急性咽炎", "急性喉炎", "就诊科室": " ", "五官科 耳鼻喉科", "治疗方式": " ", "对症治疗 药物治疗 支持性治疗", "治疗周期": " ", "1-3个月", "治愈率": " ", "90%", "常用药品": " ", "注射用头孢唑林钠", "诺氟沙星胶囊", "治疗费用": " ", "根据不同医院, 收费标准不一致, 市三甲医院约 (3000—8000元)", "慢性鼻窦炎": { "医保疾病": " ", "否", "患病比例": " ", "0.01%-0.018%", "易感人群": " ", "无特定人群", "传染方式": " ", "与慢性刺激有关, 无传染性", "并发症": " ", "鼻囊炎", "咽喉炎", "就诊科室": " ", "五官科 耳鼻喉科", "治疗方式": " ", "药物治疗 对症治疗 支持性治疗", "治疗周期": " ", "1月", "治愈率": " ", "80%", "常用药品": " ", "慢严舒柠 咽炎片", "慢严舒柠 复方青橄榄利咽含片", "治疗费用": " ", "针对不同病因花费不同, 药物治疗约需3000元", "上颌窦炎": { "医保疾病": " ", "否", "患病比例": " ", "0.005%", "易感人群": " ", "无特定人群", "传染方式": " ", "呼吸道传播", "并发症": " ", "鼻炎", "鼻出血", "就诊科室": " ", "传染科", "治疗方式": " ", "抗结核治疗, 对症治疗, 支持性治疗", "治疗周期": " ", "12个月以上", "治愈率": " ", "90%", "常用药品": " ", "勒马回囊膏", "白及颗粒", "治疗			> {} 头部	特殊人群", "传染方式": " ", "无传染性", "并发症": " ", "治愈率": " ", "90%", "常用药品": " ", "复方片仔癀含漱液", "鼻病": " ", "鼻病比例": " ", "0.03%", "易感人群": " ", "无特殊人群", "治疗 手术治疗 中医治疗", "治疗周期": " ", "1-3个月", "否", "患病比例": " ", "1%", "易感人群": " ", "无特殊人群", "治疗 支持性治疗", "治疗周期": " ", "10-15天", "治愈率": " ", " ", "会厌囊肿": " ", "医保疾病": " ", "否", "患病比", "并发症": " ", "舌腺炎", "休克", "就诊科室": " ", "五官科", "根据不同医院, 收费标准不一致, 市三甲医院约 "无传染性", "并发症": " ", "急性中耳炎", "败血", "常用药品": " ", "头孢丙酮分散片", "新诺片", "治疗", "0.021%", "易感人群": " ", "无特定的人群", "传染方, "持续性治疗", "治疗周期": " ", "3-8周", "治愈率": " ", " ", "喉肿痛": " ", "医保疾病": " ", "否", "患病比例": " ", " ", "无特定的人群", "传染方式": " ", "无传染性", "并发症": " ", "外耳道疖肿", "往来寒热", "就诊科室": " ", "五官科 耳鼻喉科", "治疗方式": " ", "药物治疗 放射治疗 药物治疗 支持治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "治愈率": " ", "90%", "常用药品": " ", "阿奇霉素片", "阿奇霉素分散片", "治疗费用": " ", "药物治疗 支持性治疗 手术治疗", "治疗周期": " ", "2-3个月", "
--	--	--	---------	---

Txt 文件：

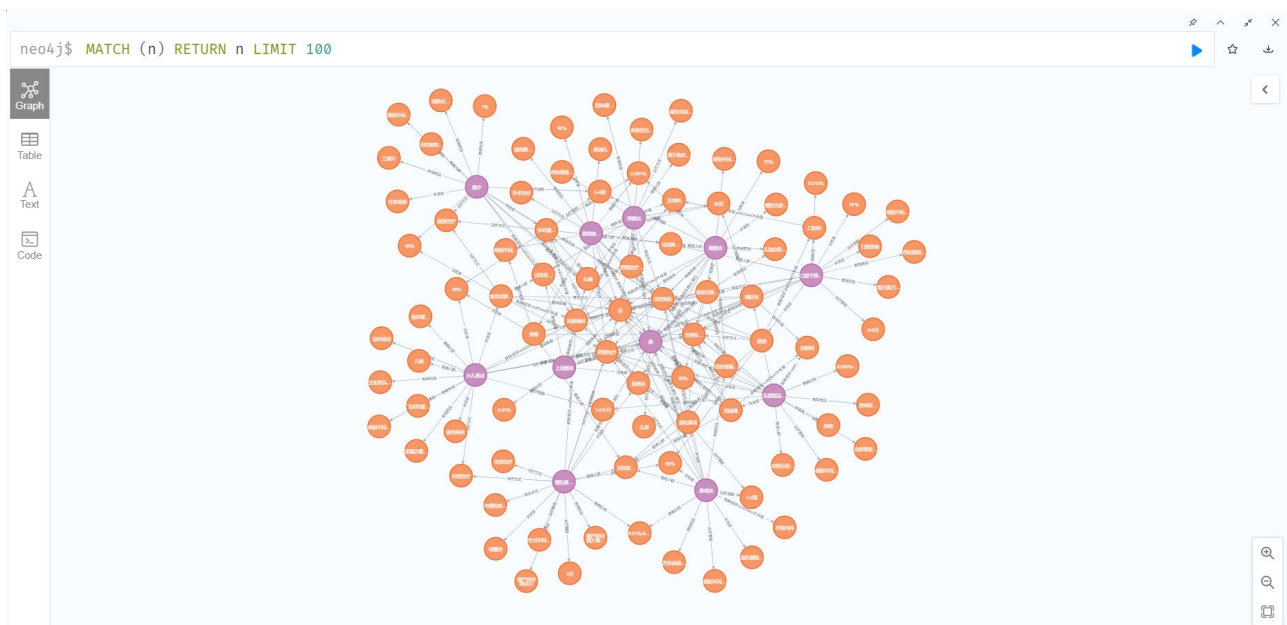
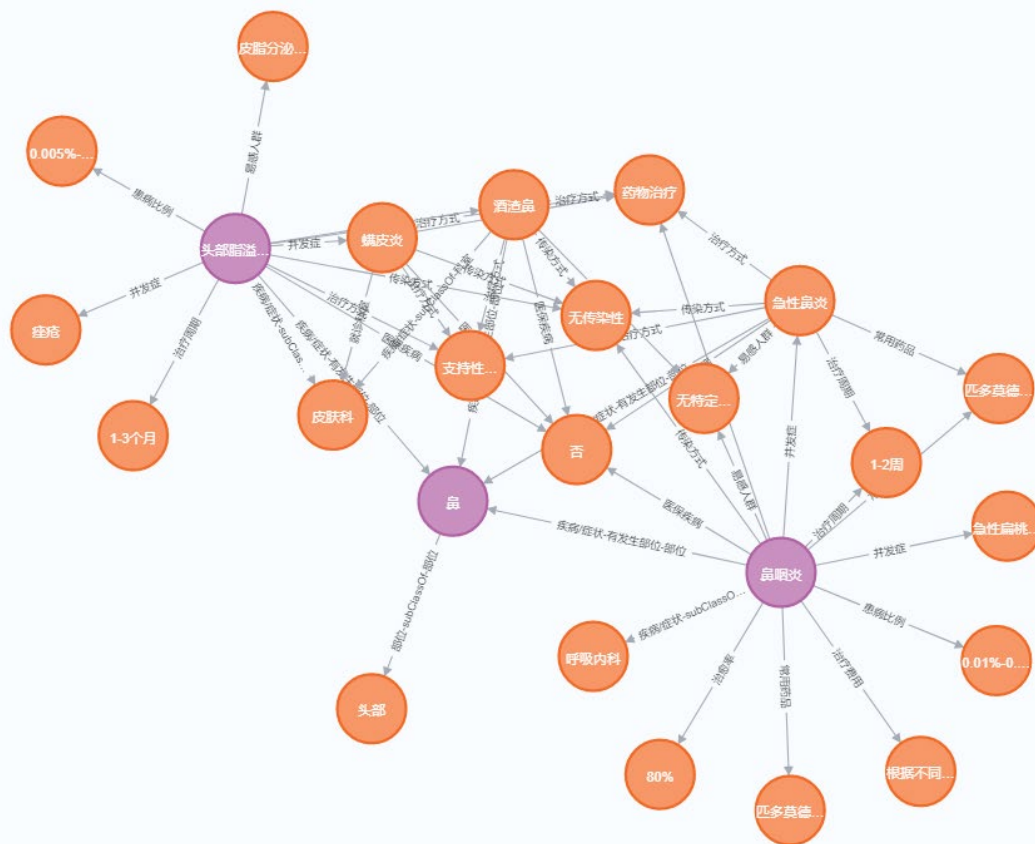


Csv 文件:

```
1 entity:ID,name,:LABEL
2 e0,鼻,疾病
3 e1,头部,tail-type
4 e2,鼻咽炎,疾病
5 e3,否,tail-type
6 e4,0.01%-0.018%,tail-type
7 e5,无特定人群,tail-type
8 e6,无传染性,tail-type
9 e7,急性扁桃体炎,tail-type
10 e8,急性鼻炎,tail-type
11 e9,呼吸内科,tail-type
12 e10,药物治疗,tail-type
13 e11,1-2周,tail-type
14 e12,80%,tail-type
15 e13,匹多莫德口服液,tail-type
16 e14,匹多莫德颗粒剂,tail-type
17 e15,根据不同医院,收费标准不一致,市三甲医院约(500--1000元),tail-type
18 e16,头部脂溢性皮炎,疾病
19 e17,0.005%-0.008%,tail-type
20 e18,皮脂分泌旺盛的人群,tail-type
21 e19,痤疮,tail-type
22 e20,蛲皮炎,tail-type
23 e21,酒渣鼻,tail-type
24 e22,皮肤科,tail-type
25 e23,支持性治疗,tail-type
26 e24,1-3个月,tail-type
27 e25,90%,tail-type
28 e26,克林霉素甲硝唑搽剂,tail-type
```

```
1 :START_ID,:END_ID,:TYPE
2 e0,e1,部位-subClassOf-部位
3 e2,e0,疾病/症状-有发生部位-部位
4 e2,e3,医保疾病
5 e2,e4,患病比例
6 e2,e5,易感人群
7 e2,e6,传染方式
8 e2,e7,并发症
9 e2,e8,并发症
10 e2,e9,疾病/症状-subClassOf-科室
11 e2,e10,治疗方式
12 e2,e11,治疗周期
13 e2,e12,治愈率
14 e2,e13,常用药品
15 e2,e14,常用药品
16 e2,e15,治疗费用
17 e16,e0,疾病/症状-有发生部位-部位
18 e16,e3,医保疾病
19 e16,e17,患病比例
20 e16,e18,易感人群
21 e16,e6,传染方式
22 e16,e19,并发症
23 e16,e20,并发症
24 e16,e21,并发症
25 e16,e22,疾病/症状-subClassOf-科室
26 e16,e10,治疗方式
27 e16,e23,治疗方式
28 e16,e24,治疗周期
```

Neo4j 可视化:



程序的全部源代码：见附录		
分析总结、收获和体会：		
<p>数据统计：（使用图或表的方法分析关系三元组、属性三元组数据，包括总数、分布情况等）</p> <ul style="list-style-type: none"> ✚ 结点共计 34897 个，统计信息包括传染方式、是否医保疾病、就诊科室、常用药品、并发症、患病比例、易感人群、治愈率、治疗周期、治疗方式、治疗费用。 ✚ 数据之间索引类型包括疾病与其所属科室、所属部位的关系，以及疾病属性之间的相互连接。可依据某个特定属性提取子图分析。 <p>优点：</p> <ul style="list-style-type: none"> ✚ 功能全面，几乎涵盖上千种疾病的所有信息，覆盖知识面较广 ✚ 封装好，代码的可读性较高，各个变量以及函数见名知意，注释较为详细。 ✚ 结果利用 neo4j 实现可视化，更加清晰明了。 ✚ 使用文件操作及时保存写出数据，提高程序调式的运行速度以及效率。 ✚ 函数变量少，减少了参数的传递，使程序较为简洁。 <p>创新之处：</p> <ul style="list-style-type: none"> ✚ 使用网络爬虫，在读取时就加以处理，减少了程序的空间及时间复杂度。 ✚ 利用 neo4j 实现知识图谱的可视化，在索引以及分类上更为直观清晰。 <p>不足之处：</p> <ul style="list-style-type: none"> ✚ 未区分部位与疾病，一定程度上分类不明确，且部分函数重复度较高，程序比较冗杂 <p>需要改进的地方：</p> <ul style="list-style-type: none"> ✚ 增加分类以及查询算法，后期还可基于人工智能生成自助问答系统等交互性更强的软件 <p>收获与学习体会：</p> <p>通过两周的程序设计实践，对人工智能基本分类、聚类算法有基本了解与实践，学习了网络爬虫以及知识图谱构建，对专业知识以及数据结构算法有了基本了解，同时代码能力也有所锤炼</p>		
自我评价：	是	否
程序运行是否无 bug？	√	
是否在撰写报告之前观看了 spoc 里的代码规范视频？	√	
程序代码是否符合代码规范(对齐与缩进，有必要的注释)？	√	
是否按模块化要求进行了程序设计，系统功能是否完善？	√	
是否是独立完成？		√
<p>自我评语：</p> <p>熟悉了 python 语言的使用，以及了解的代码的基本格式及基本操作，尝试去锤炼程序的健壮性和可读性。知识面层面，对人工智能常见的分类聚类思想有了了解。</p> <p style="text-align: right;">报告完成日期： 2022/9/5</p>		

附件:

1.get_json.py

```
1. import json
2. import requests
3. import re
4. from lxml import html
5.
6. def open_page(url):
7.     headers={'User-Agent':'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/104.0.5112.102 Safari/537.36 Edg/104.0.1293.63'}
8.     res=requests.get(url,headers=headers)
9.     res.encoding='gb2312'
10.    res=res.text.replace('\ufffd','')
11.    return res
12. #-----
13.
14.
15. def introduction(url):
16.     data=open_page(url)
17.     ht = html.fromstring(data)
18.     introduction= {}
19.
20.     introduction_key = ht.xpath('//div[@class="jib-articl fr f14 "]/div[{p}]/strong/text()'.format(p=1))[0]
21.     introduction_value = ht.xpath('//div[@class="jib-articl fr f14 "]/div[{p}]/p/text()'.format(p=1))[0]
22.     introduction[introduction_key] = introduction_value.strip()
23.
24.     for i in range(2, 5):
25.         if i == 2:
26.             basic_know_att_value = []
27.
28.             basic_know_key = ht.xpath('//div[@class="jib-articl fr f14 "]/div[{p}]/strong/text()'.format(p=i))[0]
29.             introduction[basic_know_key] = {}
30.
31.             basic_know_att_key = ht.xpath('//div[@class="jib-articl fr f14 "]/div[{p}]/p/span[1]/text()'.format(p=i))
32.
33.             basic_know_att_value_init = ht.xpath('//div[@class="jib-articl fr f14 "]/div[{p}]/p/span[2]/text()'.format(p=i))
```

```

34.         for value in basic_know_att_value_init:
35.             value = value.strip()
36.             basic_know_att_value.append(value)
37.         if '9144' in url:
38.             basic_know_att_value.insert(3, ' ')
39.             ill_all = ht.xpath('//div[@class="jib-articl fr f14 "]/div[{p}]/p
/ span[2]/a/text()'.format(p=i))
40.         for i in range(len(ill_all) - 1):
41.             basic_know_att_value.pop(4)
42.             basic_know_att_value.append(ill_all)
43.
44.
45.         for i in range(len(basic_know_att_key)):
46.             introduction[basic_know_key][basic_know_att_key[i]] = basic_k
now_att_value[i]
47.
48.         elif i == 3:
49.             com_sen_att_value = []
50.
51.             com_sen_key = ht.xpath('//div[@class="jib-articl fr f14 "]/div[{p
}]/strong/text()'.format(p=i))[0]
52.             introduction[com_sen_key] = {}
53.
54.             com_sen_att_key = ht.xpath('//div[@class="jib-articl fr f14 "]/di
v[{p}]/p/span[1]/text()'.format(p=i))
55.
56.             com_sen_att_value_init = ht.xpath('//div[@class="jib-articl fr f1
4 "]/div[{p}]/p/span[2]/text()'.format(p=i))
57.             for value in com_sen_att_value_init:
58.                 value = value.strip()
59.                 com_sen_att_value.append(value)
60.             medicine = ht.xpath('//div[@class="jib-articl fr f14 "]/div[{p}]/
p/span[2]/a/text()'.format(p=i))
61.             if '9992' in url or '9865' in url or '9987' in url:
62.                 com_sen_att_value.insert(1, ' ')
63.             for i in range(len(medicine) + 1):
64.                 com_sen_att_value.pop(4)
65.                 com_sen_att_value.insert(4, medicine)
66.
67.             for i in range(len(com_sen_att_key)):
68.                 introduction[com_sen_key][com_sen_att_key[i]] = com_sen_att_v
alue[i]
69.
70.         else:

```

```

71.         tip_key_t = ht.xpath('//div[@class="jib-articl fr f14 "]/div[{p}]/strong/text()'.format(p=i))
72.         if len(tip_key_t) != 0:
73.             tip_key = ht.xpath('//div[@class="jib-articl fr f14 "]/div[{p}]/strong/text()'.format(p=i))[0]
74.
75.             tip_value_init = ht.xpath('//div[@class="jib-articl fr f14 "]/div[{p}]/p/text()'.format(p=i))[0]
76.             tip_value = tip_value_init.strip()
77.
78.             introduction[tip_key] = tip_value
79.
80.     return introduction
81.
82. def symptom(url):
83.     data=open_page(url)
84.     ht = html.fromstring(data)
85.
86.     members = ht.xpath('//div[@class="jib-articl fr f14 jib-lh-articl"]/p')
87.
88.     sum = ""
89.     for member in members:
90.         member = member.xpath('string().').strip()
91.         member = str(member)
92.         sum += member
93.
94.
95.     return "常见症状",sum
96.
97. def inspect(url):
98.     data=open_page(url)
99.     ht = html.fromstring(data)
100.
101.     members = ht.xpath('//div[@class="jib-articl fr f14 jib-lh-articl"]/p')
102.
103.     sum = ""
104.     for member in members:
105.         member = member.xpath('string().').strip()
106.         member = str(member)
107.         sum += member
108.
109.
110.     return "检查",sum
111.

```

```
112. def diagnosis(url):
113.     data=open_page(url)
114.     ht = html.fromstring(data)
115.
116.     members = ht.xpath('//div[@class="jib-articl fr f14 jib-lh-articl"]/p')
117.
118.     sum = ""
119.     for member in members:
120.         member = member.xpath('string().').strip()
121.         member = str(member)
122.         sum += member
123.
124.     return "诊断鉴别",sum
125.
126. def treat(url):
127.     data=open_page(url)
128.     ht = html.fromstring(data)
129.
130.     members = ht.xpath('//div[@class="jib-lh-articl"]/p')
131.
132.     sum = ""
133.     for member in members:
134.         member = member.xpath('string().').strip()
135.         member = str(member)
136.         sum += member
137.
138.
139.     return "治疗",sum
140.
141. def cause(url):
142.
143.     data=open_page(url)
144.     ht = html.fromstring(data)
145.     members = ht.xpath('//div[@class=" jib-articl fr f14 jib-lh-articl"]')
146.
147.     sum = ""
148.     for member in members:
149.         member = member.xpath('string().')
150.         member = str(member.strip())
151.         sum += member
152.
153.     return "病因",member
154.
155.
```



```

156. def prevent(url):
157.     data=open_page(url)
158.     ht = html.fromstring(data)
159.     members = ht.xpath('//div[@class="jib-articl fr f14 jib-lh-articl"]/p')
160.
161.     sum = ""
162.     for member in members:
163.         member = member.xpath('string().').strip()
164.         member = str(member)
165.         sum += member
166.     return "预防",sum
167.
168. def neopathy(url):
169.     data=open_page(url)
170.     ht = html.fromstring(data)
171.     members = ht.xpath('//div[@class="jib-articl fr f14 jib-lh-articl"]/p')
172.
173.     sum = ""
174.     for member in members:
175.         member = member.xpath('string().').strip()
176.         member = str(member)
177.         sum += member
178.     return "并发症",sum
179.
180. #-----
181.
182. def extract(res,ser,modi):
183.     temp=re.findall(ser,res,re.S)
184.     return re.findall(modi,str(temp),re.S)
185.
186. def page_link(res,menu):
187.     ser_='<a href="(.*?)">'+menu+'</a>'
188.     return 'https://jib.xywy.com'+re.findall(ser_,res)[0]
189.
190. #-----
191.
192. def detail(link):
193.     chai='/il_sii_(.*?).htm'
194.     feature=(re.findall(chai,link))[0]
195.     print(feature)
196.     content=introduction('https://jib.xywy.com/il_sii/gaishu/'+feature+'.htm
    ')

```

```

197.     title1,value1=symptom('https://jib.xywy.com/il_sii/symptom/'+feature+'.htm')
198.     content[title1]=value1
199.     title2,value2=inspect('https://jib.xywy.com/il_sii/inspect/'+feature+'.htm')
200.     content[title2]=value2
201.     title3,value3=diagnosis('https://jib.xywy.com/il_sii/diagnosis/'+feature+'.htm')
202.     content[title3]=value3
203.     title4,value4=treat('https://jib.xywy.com/il_sii/treat/'+feature+'.htm')
204.     content[title4]=value4
205.     title5,value5=cause('https://jib.xywy.com/il_sii/cause/'+feature+'.htm')
206.     content[title5]=value5
207.     title6,value6=prevent('https://jib.xywy.com/il_sii/prevent/'+feature+'.htm')
208.     content[title6]=value6
209.     title7,value7=neopathy('https://jib.xywy.com/il_sii/neopathy/'+feature+'.htm')
210.     content[title7]=value7
211.     return content
212.
213. # def correct(res,item):
214.     # ser='<a title="(.)" href="/il_sii_3865.htm">穿透性角膜移...</a>'
215.     # #<a title="穿透性角膜移植术所致青光眼" href="/il_sii_3865.htm">穿透性角膜移...</a>
216.     # r=re.findall(ser,res)
217.     # if len(r)==0:
218.     #     return item
219.     # return r[0]
220.
221.
222. info={}
223. url='https://jib.xywy.com/html/toubu.html'
224. code=open_page(url)
225. ht = html.fromstring(code)
226. name = ht.xpath('//li[@class="pr"]/a/text()')
227. for i in name:
228.     info[i]={}
229.     ser_item='>'+i+'</a>\r\n                                     <ul class="jbk-sed-menu
        jb-body-nav bor pa none f12">(.*?)</ul>'
230.     modi_item='<li><a title="(.)" href="(.)">'
231.     items=extract(code,ser_item,modi_item)
232.     if len(items)==0:
233.         link=page_link(code,i)

```

```

234.         item_code=open_page(link)
235.         item_ht = html.fromstring(item_code)
236.         items= item_ht.xpath('//ul[@class="ks-ill-list clearfix mt10"]/li/a/
            @title')
237.         href=item_ht.xpath('//ul[@class="ks-ill-list clearfix mt10"]/li/a/@h
            ref')
238.         ix=0
239.         for s in items:
240.             print(ix)
241.             content=detail(href[ix])
242.             info[i][s]=content
243.             ix+=1
244.         else:
245.             for j in items:
246.                 info[i][j[0]]={}
247.                 link2='https://jib.xywy.com'+j[1]
248.                 item2_code=open_page(link2)
249.                 item2_ht = html.fromstring(item2_code)
250.                 items2=item2_ht.xpath('//ul[@class="ks-ill-list clearfix mt10"]/
                    li/a/@title')
251.                 href2=item2_ht.xpath('//ul[@class="ks-ill-list clearfix mt10"]/l
                    i/a/@href')
252.                 idx=0
253.                 for s in items2:
254.                     print(idx)
255.                     content=detail(href2[idx])
256.                     info[i][j[0]][s]=content
257.                     idx+=1
258. output_path = "疾病_edit4.json"
259. print(info)
260. with open(output_path, "w",encoding='utf-8') as f:
261.     json.dump(info, f, ensure_ascii=False)

```

2.test_关系.py

```

1. import json
2.
3.
4.
5. def my_split(data):
6.     return data.split(" ")
7.
8.
9. def name(data):
10.    return data.split(": ")[0]

```

```

11.
12. if __name__ == '__main__':
13.     output_path = "疾病_edit3.json"
14.     with open(output_path, "r", encoding='utf-8') as f:
15.         data=json.load(f)
16.     with open("test.txt", "w", encoding='utf-8') as f2:
17.         menu=['头部', '四肢', '生殖部位']
18.         mm=[]
19.         for i_1 in data:
20.             if i_1 in menu:
21.                 for i_11 in data[i_1]:
22.                     wri=i_11+'@@@部位-subClassOf-部位@@@'+i_1
23.                     f2.write(wri+'\n')
24.                     for i_12 in data[i_1][i_11]:
25.                         w=i_12+'@@@疾病/症状-有发生部位-部位@@@'+i_11
26.                         f2.write(w+'\n')
27.                         for item in data[i_1][i_11][i_12].keys():
28.                             if item=='治疗费用: ':
29.                                 w=i_12+'@@@治疗费用
@@@'+data[i_1][i_11][i_12][item]
30.                                 f2.write(w+'\n')
31.                                 continue
32.                                 n=name(item)
33.                                 if False==isinstance(data[i_1][i_11][i_12][item],
list):
34.                                     re=my_split(data[i_1][i_11][i_12][item])
35.                                     if item=='就诊科室: ':
36.                                         if len(re)==1:
37.                                             wde=i_12+'@@@疾病/症状-subClassOf-科室
@@@'+re[0]
38.                                             f2.write(wde+'\n')
39.                                         else:
40.                                             wde=i_12+'@@@疾病/症状-subClassOf-科室
@@@'+re[2]
41.                                             f2.write(wde+'\n')
42.                                             me=re[2]+'@@@科室-subClassOf-科室
@@@'+re[0]
43.                                             if me not in mm:
44.                                                 mm.append(me)
45.                                             continue
46.                                         for de in re:
47.                                             wde=i_12+'@@@'+n+'@@@'+de
48.                                             f2.write(wde+'\n')
49.                                         else:

```

```

50.                 for de in data[i_1][i_11][i_12][item]:
51.                     wde=i_12+'@@@'+n+'@@@'+de
52.                     f2.write(wde+'\n')
53.             else:
54.                 for i_2 in data[i_1]:
55.                     wri=i_2+'@@@疾病/症状-有发生部位-部位@@@'+i_1
56.                     f2.write(wri+'\n')
57.                 for item in data[i_1][i_2].keys():
58.                     if item=='治疗费用: ':
59.                         w=i_2+'@@@治疗费用@@@'+data[i_1][i_2][item]
60.                         f2.write(w+'\n')
61.                     continue
62.                     n=name(item)
63.                     if False==isinstance(data[i_1][i_2][item],list):
64.                         re=my_split(data[i_1][i_2][item])
65.                         if item=='就诊科室: ':
66.                             if len(re)==1:
67.                                 wde=i_2+'@@@疾病/症状-subClassOf-科室
@@@'+re[0]
68.                                 f2.write(wde+'\n')
69.
70.                             else:
71.                                 wde=i_2+'@@@疾病/症状-subClassOf-科室
@@@'+re[2]
72.                                 f2.write(wde+'\n')
73.                                 me=re[2]+'@@@科室-subClassOf-科室
@@@'+re[0]
74.                                 if me not in mm:
75.                                     mm.append(me)
76.                                 continue
77.                             for de in re:
78.                                 wde=i_2+'@@@'+n+'@@@'+de
79.                                 f2.write(wde+'\n')
80.                             else:
81.                                 for de in data[i_1][i_2][item]:
82.                                     wde=i_2+'@@@'+n+'@@@'+de
83.                                     f2.write(wde+'\n')
84.                 for it in mm:
85.                     f2.write(it+'\n')

```

3. get_node_relation.py

1. #读取txt 文件中的三元组
2. #将节点去重复, 写入 csv 文件
3. #将关系写入 csv 文件


```

4.
5. import pandas as pd
6.
7. import csv
8. # 读取三元组文件
9.
10.entity_list = []
11.tri_list = []
12.
13.entity_dict = {}
14.ent_num = 0
15.with open("关系.txt", "r", encoding="utf-8") as f:
16.    for line in f:
17.        line = line.strip().split('@@@')
18.        head = line[0].strip()
19.        rel = line[1].strip()
20.        tail = line[2].strip()
21.        if head not in entity_dict:
22.            entity_dict[head] = 'e' + str(ent_num)
23.            entity_list.append(['e' + str(ent_num), head, "疾病"])
24.            ent_num += 1
25.        if tail not in entity_dict:
26.            entity_dict[tail] = 'e' + str(ent_num)
27.            entity_list.append(['e' + str(ent_num), tail, "tail-type"])
28.            ent_num += 1
29.        #添加三元组, 使用实体id 和关系id
30.        if [entity_dict[head], entity_dict[tail], rel] not in tri_list:
31.            tri_list.append([entity_dict[head], entity_dict[tail], rel])
32.
33.#将去重复后的实体写入 csv 文件
34.
35.entity_header = ["entity:ID", "name", ":LABEL"]
36.with open ('entity.csv','w',encoding='utf-8',newline='') as fp:
37.    # 写
38.    writer =csv.writer(fp)
39.    # 设置第一行标题头
40.    writer.writerow(entity_header)
41.    # 将数据写入
42.    writer.writerows(entity_list)
43.
44.#将去重复后的三元组写入 csv
45.tri_header = [':START_ID', ':END_ID', ':TYPE']
46.with open ('triplets.csv','w',encoding='utf-8',newline='') as fp:
47.    # 写

```

```
48.     writer =csv.writer(fp)
49.     # 设置第一行标题头
50.     writer.writerow(tri_header)
51.     # 将数据写入
52.     writer.writerows(tri_list)
53. #h_r_t_name = [":START_ID", "role", ":END_ID"]
54. h_r_t = pd.read_csv("triplets.csv", decimal="\t", names=tri_header)
55. print(h_r_t.info())
```