目录

[一、肿瘤术语及药物知识的关联数据基础与医学meta分析知识库（田敏） 1](#_Toc507943758)

[1.选题题目 1](#_Toc507943759)

[2.数据来源 1](#_Toc507943760)

[3.开发语言、开发工具 1](#_Toc507943761)

[4.知识库结构 2](#_Toc507943762)

[5.知识库的功能要求 4](#_Toc507943763)

[6.数据库关联 5](#_Toc507943764)

[二、用于合并的临床路径\_临床路径、药物不良相互作用知识库（江超） 6](#_Toc507943765)

[1.选题题目 6](#_Toc507943766)

[2.数据来源 6](#_Toc507943767)

[3.开发语言、开发工具 6](#_Toc507943768)

[5.知识库的功能要求 8](#_Toc507943769)

[6.数据库关联 9](#_Toc507943770)

[三、基于医学知识库的临床医学专业课程“知识主题-课程”学习网络（谢祎玉） 10](#_Toc507943771)

[1.选题题目 10](#_Toc507943772)

[2.数据来源 10](#_Toc507943773)

[3.开发语言、开发工具 10](#_Toc507943774)

[4.知识库结构 11](#_Toc507943775)

[5.知识库的功能要求 11](#_Toc507943776)

[6.数据库关联 12](#_Toc507943777)

[四、颅脑癌症影像知识库（朱安琪） 13](#_Toc507943778)

[1.选题题目 13](#_Toc507943779)

[2.数据来源 13](#_Toc507943780)

[3.开发语言、开发工具 14](#_Toc507943781)

[4.知识库结构 14](#_Toc507943782)

[5.知识库的功能要求 17](#_Toc507943783)

[6.数据库关联 20](#_Toc507943784)

[五、融合地理位置与领域知识的医疗健康信息推荐知识库（朱安琪） 21](#_Toc507943785)

[1.选题题目 21](#_Toc507943786)

[2.数据来源 21](#_Toc507943787)

[3.开发语言、开发工具 21](#_Toc507943788)

[5.知识库的功能要求 22](#_Toc507943789)

[6.数据库关联 24](#_Toc507943790)

[六、基于网络医患交流数据的医患知识表示系统（李易时） 25](#_Toc507943791)

[1.选题题目 25](#_Toc507943792)

[2.数据来源 25](#_Toc507943793)

[3.开发语言、开发工具 25](#_Toc507943794)

[4.知识库结构 25](#_Toc507943795)

[5.知识库的功能要求 26](#_Toc507943796)

[6.数据库关联 27](#_Toc507943797)

[七、心理健康网络咨询对话知识库（岳雪琪） 28](#_Toc507943798)

[1.选题题目 28](#_Toc507943799)

[2.数据来源 28](#_Toc507943800)

[3.开发语言、开发工具 28](#_Toc507943801)

[4.知识库结构及功能要求 28](#_Toc507943802)

[5.数据库关联 29](#_Toc507943803)

# 一、肿瘤术语及药物知识的关联数据基础与医学meta分析知识库（田敏）

**1.选题题目**

基于PubMed医学数据库：利用文本挖掘的方法进行有关肿瘤的原始研究的一般资料、原始研究临床特征、总效应值的数据、原始研究的方法学质量等数据的抽取和计算；基于Meta分析工具（针对数字型数据）和自然语言处理等技术（针对字符串型数据）：对以上肿瘤疾病的数据进行医学Meta分析研究。

**2.数据来源**

PubMed等医学数据库提供的文献；可限定检索词定量下载文本；若为“肿瘤（Neoplasm）”主题相关的文献，数据量大致为3-4G；原始数据格式为txt文本。

**3.开发语言、开发工具**

（1）MetaMap、SemRep等软件

用于对英文文本的命名实体识别，可提取出其中的疾病。也可自行设计命名实体识别的方法进行提取。注：这两款软件都是按照UMLS的标准进行术语的提取。

【UMLS一体化医学语言系统】

https://www.nlm.nih.gov/research/umls/

用户名：tianminwhu

密码：tm5303019

（2）Python语言

nltk、stanfordNLP等自然语言处理工具包可对文本进行自然语言处理，进行分析数据的信息提取。

Word2Vec或Glove可从文本中训练词向量，作为语义的表示之一。

TensorFlow等深度学习框架可以提供包括卷积神经网络CNN、循环神经网络RNN等在内的深度学习支持，可用于命名实体识别、词义消歧、文本分类、聚类等。

（3）Meta统计软件或工具

包括但不限于：

R中的rmeta、meta程序包，它们是R语言专门进行meta分析的一个程序包。

（参考：用R进行Meta分析

<https://jingyan.baidu.com/article/7f41ecec00e633593c095c7a.html>）

STATA、REVMAN等软件可更专业、深入地对数据进行meta分析。

包括：异质性检验、计算统计结果的可信区间、森林图展示结果等（注意不同的分析软件方法、结果不唯一）。

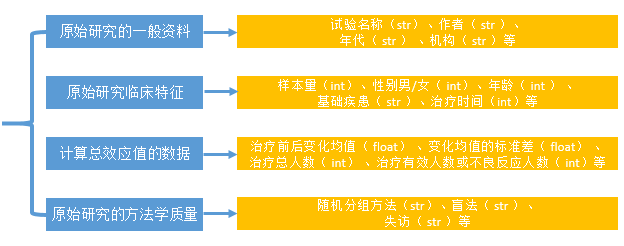
**4.知识库结构**

**4.1 逻辑结构**

①首先通过文本分类、聚类选取适用于Meta分析的文献

参考传统机器学习或者深度学习如CNN的方法对文本进行分类、聚类，选取合理的指标用于评价文献用于Meta分析的质量，再从中选取得分高的文献用于后续的信息抽取和分析。

②从选择的文本中抽取出重要的可利用信息，包括：



其中，str表示字符串，int表示整数，float表示浮点数。

例子：

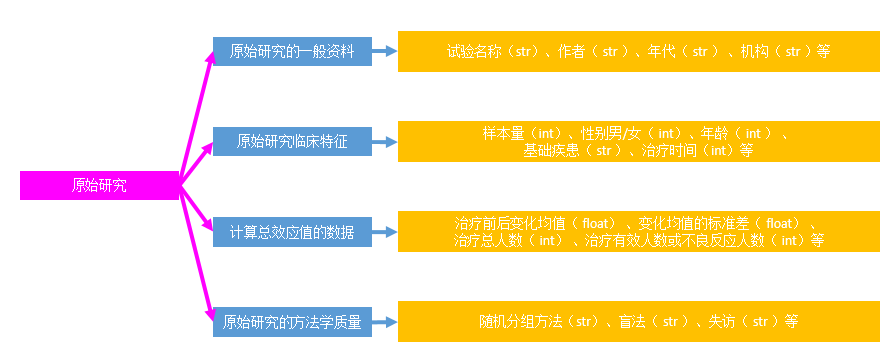


A.概念及其属性

包括但不限于：实验名称（字符串）、作者（字符串）、所属机构（字符串）、隐藏（字符串）、盲法（字符串）、随访（字符串）、年龄（考虑是否离散化）、性别（二值）、严重度（不同程度）、病程（字符串）、治疗方法（字符串）、剂量（考虑是否离散化）、疗程（字符串）、对照措施（字符串）、死亡（二值属性）、残废（不同程度）、事件数（数值）等。

其中，实验名称、作者、年代、所属机构都属于“原始研究的一般资料”，样本量、性别男/女、年龄、基础疾患、治疗时间等都属于“原始研究临床特征”，治疗前后变化均值、变化均值的标准差、治疗总人数、治疗有效人数或不良反应人数等都属于“计算总效应值的数据”，随机分组方法、盲法、失访等都属于“原始研究的方法学质量”。

B.关系



对于一项原始研究（类），它包含“原始研究的一般资料（类）” “原始研究临床特征（类）” “计算总效应值的数据（类）” “原始研究的方法学质量（类）”。其中，

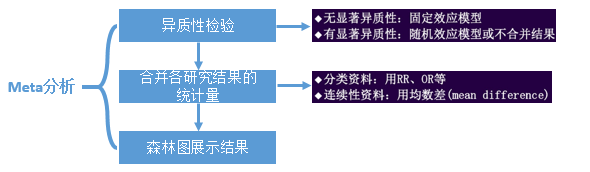
（1）“原始研究的一般资料”具有属性：实验名称、作者、年代、所属机构。

（2）“原始研究临床特征”具有属性：样本量、性别男/女、年龄、基础疾患、治疗时间等。

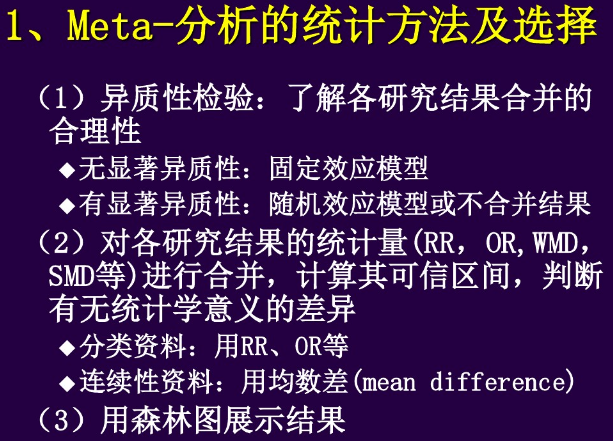
（3）“计算总效应值的数据”具有属性：治疗前后变化均值、变化均值的标准差、治疗总人数、治疗有效人数或不良反应人数等。

（4）“原始研究的方法学质量”具有属性：随机分组方法、盲法、失访等。

③对研究结果进行Meta分析，可参考（注意不同的分析软件方法、结果不唯一）：



此方法的具体说明：



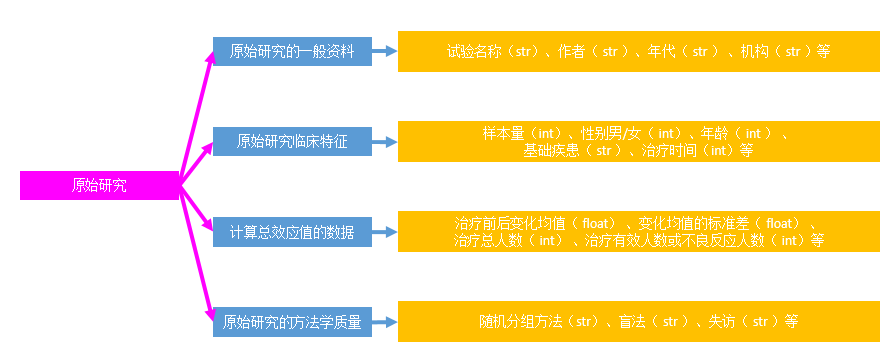
**4.2物理结构**

所有数据按知识库的逻辑结构以知识库本体Owl形式存储。

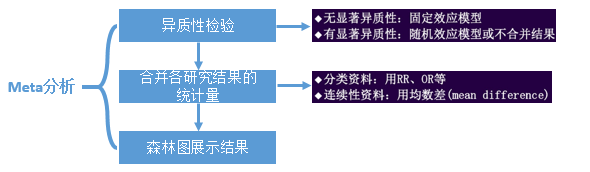
**5.知识库的功能要求**

**（1）知识库内部需求**

①首先从医学数据库格式为txt文本的原始数据中通过机器学习或深度学习进行文本的选择，然后通过自然语言处理等技术进行文本挖掘以抽取出重要的可利用信息，可结合传统的条件随机场、隐马尔科夫与目前较为流行和高效的深度学习（CNN、RNN等）来进行命名实体识别、词义消歧、文本分类、聚类等操作，实现数据的合理、准确提取。对于一个研究样本，参考：



②提取数据之后对数据进行Meta分析，在字符串格式的数据上可利用自然语言处理上面的词义消歧、语义消歧等技术，对字符串进行合理、有效合并或区分，在int、float格式的数据上需使用专用的软件或工具进行专业处理，得到科学结果，参考：



**（2）知识库外部需求**

知识处理的结果表现形式为知识库本体Owl文件，应用形式为利用owl文件围绕各自项目需求，基于JAVA或Python开发web项目，并提供知识库集成的API（各自owl的调用方法），为以后知识库集成做准备。

**6.数据库关联**

通过肿瘤疾病名称可以和江超的临床指南知识库以及朱安琪的颅脑影像知识库关联，通过机构信息可以和朱安琪的地理信息知识库关联。

# 二、用于合并的临床路径\_临床路径、药物不良相互作用知识库（江超）

## 1.选题题目

为了解决基于药物不良相互作用的多共病（comorbidities）临床路径合并问题，首先设计多疾病临床路径不良反应查询系统：基于临床路径知识库，药物不良反应知识库，实现两个及以上疾病的临床路径特定任务节点之间的不良反应查询（输入各临床路径的节点，输出不良反应）。

## 2.数据来源

1. **数据来源**

临床路径数据主要来源有两个：

1. 国家卫计委发布的1010个临床路径，特点：结构明晰，内容粗放，入径条件严格，入径难度大。
2. NICE (https://www.nice.org.uk/guidance)临床路径，特点：结构内容不够清晰，需要整理，入径条件比较宽松。

药物不良相互作用的数据来源为：drugs.com

使用两个临床路径来源互为对照，整理出肿瘤（癌症）、心脑血管疾病（高血压，脑卒中及冠心病等）、糖尿病、慢性呼吸系统疾病四大慢性病的临床路径，并根据涉及的药物从drugs.com整理drug-drug、drug-disease等不良反应。

1. **获取方式**

网站获取

1. **大概数据量**

典型的肿瘤疾病、高血压、脑卒中、冠心病、糖尿病、慢性呼吸系统病等6-10种疾病的临床路径及相关药物的不良反应。

1. **原始数据格式**

文本，网页。

## 3.开发语言、开发工具

1. Python语言

临床路径及药物网页html的爬取，数据的格式化处理。

1. Protégé工具

本体构建工具，用于手动构建知识库框架，可以使用Java API（Jena）、python API（owl库）进行大规模自动扩充本体。

1. Web 项目开发框架

Django（python），spring MVC（java）用于开发能够通过网页访问的查询系统。

#### 4.知识库结构

**4.1逻辑结构**

（1）**概念及属性**

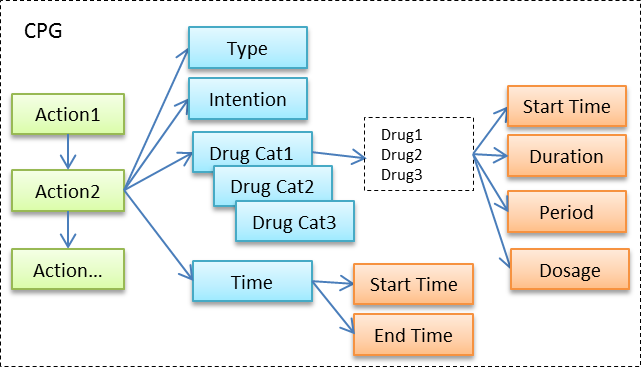


图1：临床路径模型

临床路径总的可分为几个action（医疗行为），action之间有时间上的先后次序，相邻action之间采用before/after关系连接。

action具有属性type（行为类型，取值为test，medication……），intention（行为目标，表达形式为<属性，增加/减少/稳定>二元组，属性代表身体的某一属性，action的目的在于“增加/减少/稳定”该属性），Drug category（药物类型），time（action的时间属性，可表述为里程碑式的开始/结束条件，或者是具体的时间日期等形式）； Drug Category下面有drug子类，drug具有的属性为：start Time（开始时间，相对时间，相对于action开始时间），Duration（持续服用时间）, period（服用周期），Dosage（剂量）。

**（2）药物不良相互作用**

表1：Drug-drug

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Drug1 | Drug2 | Type | Significance | Effect time | References |
|  |  | adverse effects /interfere | major/  moderate/  minor/  null | text |  |

表2：Disease-disease

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Drug | Disease | Type | Significance | Effect time | References |
|  |  | adverse effects | major/  moderate/  minor | text |  |

**4.2 物理结构**

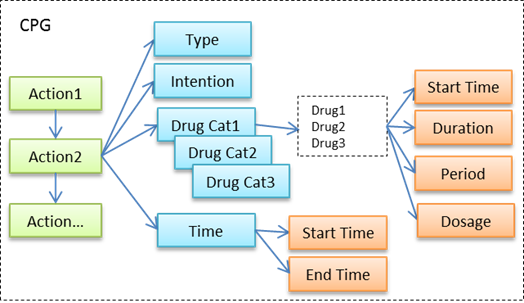
临床路径为Owl本体文件形式。

药物不良反应存储为关系数据库。

## 5.知识库的功能要求

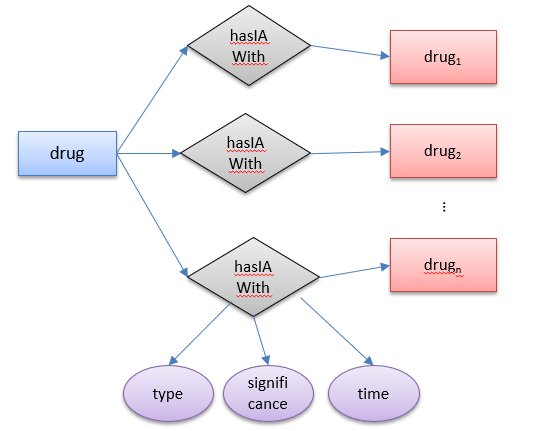
**（1）知识库内部要求：**

1. 使用两个临床路径来源互为对照，整理出肿瘤（癌症）、心脑血管疾病（高血压，脑卒中及冠心病等）、糖尿病、慢性呼吸系统疾病四大慢性病的6-10个临床路径，按照指定逻辑结构，将临床路径划分为几个按时间线性排序的action，action之下为Drug Category，Drug Category之下为drug。（action与Drug category之间若还有并行及其他等无法处理问题，可增加一个task类调节）



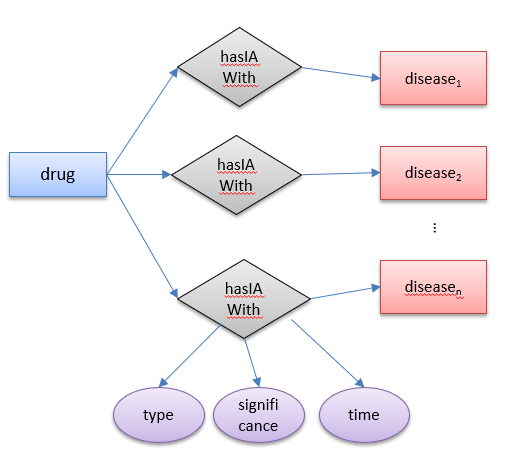
1. 并根据涉及的药物从drugs.com整理drug-drug、drug-disease等不良反应。

Drug-drug



Drug-drug之间具有hasIAwith关系，该关系具有type（相互作用类型，adverse/interfere），significance（严重程度，major/moderate/minor/null），time（不良反应作用有效时间区间）。

**Drug-disease**



属性同上。

**（2）知识库外部要求：**

有时候冲突不仅仅是药物之间的冲突，如两个临床路径中的医疗任务中的药物并未有药理冲突（如血栓患者被给以抗凝药物减少血栓，而这个患者由于手术等造成的流血就需要增加血液凝固），无法依靠药物冲突知识来解决。因此采用intention（目的）的方式来解决冲突。

intention的表达形式（属性，增加/减少/稳定），匹配方式则为首先识别“属性”，一致则再匹配情态动词（增加/减弱/稳定），情态动词不一致，则代表intention冲突，返回冲突结果。

实现的功能为：根据选定的多个临床路径特定的节点粒度（action/task/drug category），返回相应的不良反应结果，action/task粒度级，返回intention冲突结果；drug category及drug级，返回drug interaction。

## 6.数据库关联

涉及到疾病名称/药物名称的知识库均可产生一定关系的互联。

**三、基于医学知识库的临床医学专业课程“知识主题-课程”学习网络（谢祎玉）**

**1.选题题目**

基于医学知识库的临床医学专业课程“知识主题-课程”学习网络：选武汉大学临床医学五年制专业课程简介、教学大纲、电子教案、教材等作为输入，通过主题挖掘技术挖掘出各课程各章节的主题及其关联，从而得出一个“知识主题-课程”的关联网络，并进一步实现“知识主题-课程”体系网络知识库与可视化查询系统”。

研究目标：以相关教学和学习理论为指导，通过对临床医学五年制课程体系及知识主题的深度挖掘与汇聚，建立临川医学“知识主题-课程”体系网络，以揭示课程间、知识主题间以及课程与知识主题间的联系，并将其应用于课程实践，为学生及相关教学与管理人员提供深度、显性的知识服务，推进课程知识管理及课程体系建设。

**2.数据来源**

数据来源：武汉大学临床医学五年制专业课程简介、教学大纲、电子教案、教材等课程资料作为输入，选取临床医学五年制专业的14门主干课程进行研究，包括解剖学、组织胚胎学、生理学、生物化学与分子生物学、药理学、病理学、病理生理学、医学微生物学、医学免疫学、诊断学、内科学、外科学、妇产科学和儿科学，采用人民卫生出版社第八版教材的目录作为课程章节划分的依据。

分词词典：采用中国生物医学文献数据库中的主题导航的主题词作为分词依据，

http://www.sinomed.ac.cn/zh/subjectSearch.html

停用词表：整合“哈工大停用词词库”、“百度停用词表”等停用词表，去重后得到一份较为全面的停用词表

数据量：原始数据量大概3.7G

原始数据结构：ppt、word、txt等，用于主题挖掘的输入文件为txt

**3.开发语言、开发工具**

（1）python语言：对原始数据进行处理，从而得到LDA的输入文本，利用python的jieba包对原始数据进行分词处理

（2）LDA算法：对课程文本进行主题挖掘，从而得到课程章节的潜在主题

（3）HLDA算法：对LDA得到的结果进行进一步的层次主题挖掘

（4）语义网络：得到“知识主题-课程”学习网络

（5）owl语言：将“知识主题-课程”学习网络用owl文件进行知识表示

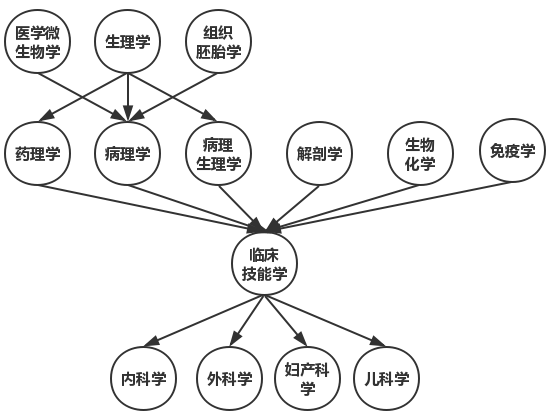
（6）JAVA语言或python语言：开发可视化的查询系统（web项目），并提供知识库集成的API（各自owl的调用方法），为以后所有的知识库集成做准备

（7）Protégé：用于构建知识库框架

**4.知识库结构**

**4.1 逻辑结构**

根据调研构建出临床医学专业主干课程的偏序有向图，揭示课程之间的逻辑与时序关联，为主题挖掘的结果提供验证和参考依据。



采用Protégé进行本体构建时，数据的层次关系为：学科 -> 课目 -> 章节 –> 主题

**4.2 物理结构**

用于数据预处理的分词词典和停用词表为txt格式，课程文本为txt文本格式

“知识主题-课程”关联网络采用owl文件进行存储

**5.知识库的功能要求**

**（1）知识库内部要求：**

a. 调研确定武汉大学临床医学专业五年制的课程体系（可看教改项目或教研论文明确哪些课是一个课程体系，校内外网络资源），收集某课程体系（应包含专业基础课及专业课）所有课程资料，确定进行研究的课程，调研并构建课程之间的偏序有向图，显性表明课程之间的逻辑与时序关联；以及调研Disease Ontology、Linked Life Data等医学知识库。

b. 将收集到的原始数据进行数据预处理操作，主要包括分词、去停用词等，得到可以用于主题挖掘的输入文件。

c. 分析资料，确定分析挖掘的方案，挖掘出“基于医学知识库的临床医学专业课程“知识主题-课程”学习网络。利用LDA方法对各课程的课程简介及课程大纲挖掘出各课程的关键主题词，利用HLDA方法对各课程的电子教案或电子教材进行层次主题挖掘，挖掘出课程中包含的所有知识主题词及这些知识主题词之间的关联，形成每一门课程的层次知识主题树，最终形成14门课程的“知识主题-课程”关联网络，采用语义网络和owl语言进行知识表示。

d. 设计开发临床医学类“知识主题-课程”体系网络（可视化）查询系统，基于主题挖掘得到的层次主题树进行相关知识的查询。

f. 进行检验与评估。

详细数据：将收集的资料按照章节划分，14门科目，共385章，每章一个txt文件（从0开始编号，到384）病理生理学：0-18（19）；病理学：19-37（19）；儿科学：38-55（18）；妇产科学：56-94（39）；解剖学：局部解剖学：95-103（9），系统解剖学：104-125（22）；临床技能学：126-133（8）；内科学：134-143（10）；生理学：144-154（11）；生物化学：155-181（27）；外科学：182-252（71）；药理学：253-300（48）；医学免疫学：301-323（23）；医学微生物学：324-358（35）；组织胚胎学：359-384（26）

**（2）知识库外部要求：**

数据存储：得到的“知识主题-课程”学习网络采用owl语言表示，并能够提供知识库集成的API调用接口，为以后知识库集成做准备。

结果评估：

a.以简介上的课程关系检验，

b.聘请该专业的研究生作为专家对知识点层面网络进行专家评估。

**6.数据库关联**

与李易时的基于网络医患交流数据的医患知识表示系统可以共用中文医学分词词典

# 四、颅脑癌症影像知识库（朱安琪）

## 1.选题题目

颅脑癌症影像知识库基于影像数据库、影像诊断知识数据库、影像学文献知识数据库实现快速查询和推荐颅脑癌症影像相关知识信息检索、推荐服务等功能。（本知识库面向医生而非患者）

## 2.数据来源

**1）数据来源：**

疾病相关的知识信息内容来源于国内外的相关期刊文献、经典书籍、诊疗指南、诊断标准等，以及WHO、CDC、MesSci等权威网站。

医院及专家相关知识信息内容来源于各医院官方网站、中国医院的排行榜（复旦大学或 中国医学科学院医学信息研究所等所出的排名）、相关权威网站等。

相关文献信息为构建数据库时所查阅的参考文献，及在知网和WOS上以疾病名为关键字检索得到的最近10篇文献。

推荐参考以下网页内容（只是推荐参考，不一定全面）：

<http://browser.ihtsdotools.org/> SNOMED CT国际临床参考术语集合

<https://link.springer.com/article/10.1007/s00401-016-1545-1> 2016世界卫生组织中央神经系统肿瘤分类:总结

<https://mp.weixin.qq.com/s/eHXrAIBdCQQZfmAHEwZT4A> 2016年WHO中枢神经系统肿瘤分类解读（中文）

<https://medlineplus.gov/braintumors.html>美国国家医学图书馆——脑瘤

<http://www.medsci.cn/>梅斯医学MedSci-临床医生发展平台

<http://www.medsci.cn/sci/icd-10.asp>梅斯ICD-10疾病编码查询系统

<https://www.cdc.gov/>疾病预防控制中心

<http://www.360doc.com/content/17/0323/11/40659465_639421744.shtml>颅内肿瘤的影像诊断学（上）

<http://www.baike.com/wiki/%E6%98%9F%E5%BD%A2%E7%BB%86%E8%83%9E%E7%98%A4>星形细胞瘤——互动百科

<http://www.360doc.com/content/17/0105/12/15509478_620234364.shtml> 【WHO分类】中枢神经系统肿瘤（2016版）

<http://rank.cn-healthcare.com/>中国医院排行榜

<http://top100.imicams.ac.cn/>中国医院科技影响力排行

<http://www.fudanmed.com/institute/news222.aspx>中国医院及专科声誉排行榜

**2）获取方式：**

网络获取

**3）大概的数据量**：

需获得疾病名称、基本信息、临床特征、生理状态、典型的病症、检查项目、相关文献等资料。需获得医院名称、医院地址、相关科室信息、相关设备信息、相关专家信息等资料。

所需的肿瘤名称参见2016年世界卫生组织中枢神经系统肿瘤分类中的所有中枢神经系统肿瘤。

**4）数据格式：**

存储为owl文件格式。

**※**所需的的具体数据项请参考知识库结构部分的内容。

## 3.开发语言、开发工具

## 4.知识库结构

**4.1逻辑结构**

A.概念及其属性

【关于疾病知识组织】

**包括但不限于：**疾病分类（字符串）；疾病名称（字符串）；疾病英文名称（字符串）；中文别名（字符串）；英文别名（字符串）；ICD编码（字符串）；ICD英文主题词（字符串）；MeSH编码（字符串）；MeSH英文主题词（字符串）；疾病概要（字符串）疾病定义（字符串）；疾病定义来源（字符串）；病因（字符串）；临床特诊表现（字符串）；影像学表现描述（字符串）；影像学表现图片（.jpg格式图片）；临床诊断（字符串）；鉴别诊断（字符串）；辅助检查项目（字符串）；相关文献资料（字符串）；等

【关于医院知识组织】

**包括但不限于：**医院名称（字符串）；所在省份（字符串）；所在市（字符串）；详细地址（字符串）；科室名称（字符串）；科室技术排名（字符串）；拥有影像设备名称（字符串）；专家名称（字符串）；专家职称（字符串）；专家擅长的领域（字符串）；从业经历（字符串）；发表的文献（字符串）；就诊时间（字符串）；等

总数据表

|  |  |  |  |
| --- | --- | --- | --- |
| 概念资料 | 疾病名称（中） | 医院资料 | 医院名称 |
| 疾病名称（英） | 医院地址 |
| 中文别名 | 科室名称 |
| 英文别名 | 科室电话 |
| ICD编码 | 科室技术排名 |
| ICD英文主题词 | 拥有影像设备名称 |
| MeSH编码 | 专家名称 |
| MeSH英文主题词 | 专家职称 |
| 疾病概要 | 专家擅长的领域 |
| 疾病分类 | 专家从业经历 |
| 就诊科室 | 就诊时间 |
| 基础资料 | 疾病定义 | 文献资料 | 文献名 |
| 疾病定义来源 | 作者 |
| 病因 | 发表时间 |
| 临床特诊表现 | 发表刊物名称 |
| 影像学表现描述 | 关键词 |
| 影像学表现图片 | 摘要 |
| 诊疗资料 | 临床诊断 | 所属机构 |
| 鉴别诊断 |  |  |
| 辅助检查项目 |  |  |

**※请以本体的形式存储在owl文件中。本体的构建可与我探讨。**

***注意相关医学术语（如疾病名等）要依据SNOMED CT标准录入***

关于数据的一些具体内容要求如下：

疾病分类来自WHO等权威网站，疾病名称编码遵循ICD10标准。参考内容如：

|  |  |
| --- | --- |
| 疾病名称（中） | 大脑星形细胞瘤 |
| 疾病名称（英） | Astrocytoma |
| 中文别名 | 星细胞瘤 |
| 英文别名 | Astroglioma;Glioma, Astrocytic |
| ICD编码 | M94000/3 |
| ICD英文主题词 | Astrocytoma |
| MeSH编码 | D001254 |
| MeSH英文主题词 | Astrocytoma |
| 疾病分类 | 星形细胞瘤 |
| 就诊科室 | 肿瘤科、神经外科 |
| 疾病概要 | 星形细胞肿瘤是指以星形胶质细胞所组成的肿瘤，是最常见的神经上皮性肿瘤。星形细胞瘤相当于Kernohan分级中除毛细胞型星形细胞瘤外的Ⅰ级与Ⅱ级，在ST Anne-Mayo分级中为1级与2级。星形细胞瘤为浸润性生长肿瘤，多数肿瘤切除后有复发可能，且复发后肿瘤可演变成间变性星形细胞瘤或多形性胶母细胞瘤。 |

疾病定义和概要来源包括书籍、期刊文献或者权威网站，可以由多个来源综合而成，需标注来源。

影像学表现描述一般分为CT/MRI平扫/增强等，都需要写明。临床表现中各条表现分开写，内容如：

|  |  |  |  |
| --- | --- | --- | --- |
| 星形细胞瘤 |  |  |  |
| 影像学表现描述 | CT | 平扫 | 肿瘤好发于额叶，其次为颞叶、顶叶、小脑、脑干，也可见于丘脑、胼胝体。病变主要累及髓质，但也可侵犯邻近的皮质。病灶为略高密度、低密度或低、等混杂密度，也可为囊性肿块。肿瘤形态不规则，边界多不清楚。一般没有出血和坏死表现，部分病例可见囊变，表现为肿瘤内低密度区。钙化的发生率为10%-20%，表现为点状或斑片状钙化，成人钙化少见，儿童钙化较多见。有学者认为儿童幕上原纤维型星形细胞瘤中1/3-1/2均有CT可见的钙化。一般没有或仅有轻度水肿和占位表现。 |
| 增强 | 多数肿瘤不强化或仅有轻度强化，囊性者囊壁可呈薄壁环状强化，并可见强化的壁结节。 |
| MRI | 平扫 | T1WI呈等或低信号，T2WI呈均匀或不均一高信号，边界较清楚。肿瘤周围一般无或仅有轻度水肿，占位表现轻微。一般无出血和坏死，少数可有囊变。 |
| 增强 | 多数肿瘤不强化。 |
| 临床表现 | 肿瘤的不断生长占据颅腔内空间，肿瘤阻塞脑脊液循环通路造成脑内积水和(或)脑水肿，脑脊液的回吸收障碍等均可造成颅内压增高。 | | |
| 正常颅腔容积比脑组织约大10%，当脑组织的体积增加8%～10%时尚可能无颅高压症状的出现，而当颅内占位病变占据150ml以上的容积时即可能产生相应的颅高压症状。 | | |
| 大脑半球的星形细胞瘤发病缓慢，病程较长，多数先出现由于肿瘤直接破坏所造成的定位体征和症状，随后又出现颅内压增高的症状。 | | |
| 小脑星形细胞瘤由于较早的影响脑脊液循环通路，多先出现颅内压增高症状，脑干的星形细胞瘤进展较快，病程较短，早期出现脑神经损害和锥体束征，而颅内压增高的症状常见于晚期。 | | |
| 颅内压增高的症状主要包括头疼、呕吐、视盘水肿，视力视野改变，癫痫、复视，头颅扩大(儿童期)和生命体征的改变等。 | | |

内容部分可借鉴《颅内肿瘤的影像诊断学》、百度百科、互动百科、各种文献资料等。

构建相关文献资料表时需含有文献名称、作者、年份、所属机构、关键词、摘要等信息。

**4.2 物理结构**

使用本体作为知识库构建的主要工具，其技术可行性为：

（1）知识的明示化，本体对描述领域知识的相关概念及概念之间的联系都给出了明确的定义。

（2）知识的形式化，本体基于描述逻辑和框架模型，具有计算机可理解的表现形式。

（3）知识的规范化，本体提供了引用某个领域知识的标准方式。

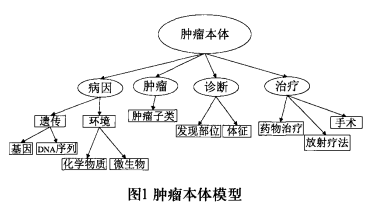
（4）知识的共享化，本体表达的领域知识反映了不同应用对该领域的共同认知

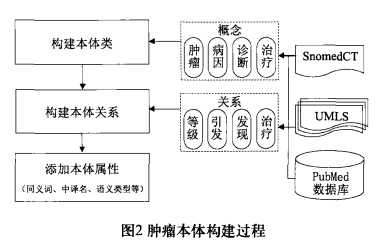
（5）知识的推理化，本体可以通过规则语言形成推理规则，通过Jena语义框架可以实现对知识的推理、查询等等。

## 5.知识库的功能要求

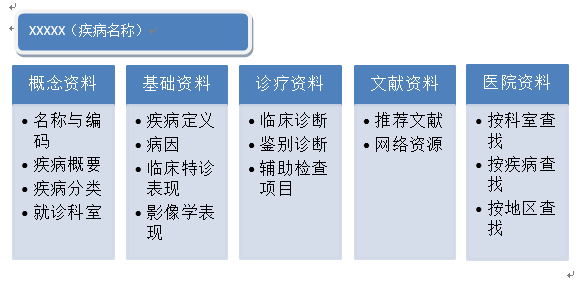
**（1）知识库内部要求：**

知识库主要是基于肿瘤本体的领域知识库。通过肿瘤名称、领域名称与医院本体相关联，医院本体的构建相对简单，因此在此给出肿瘤本体的构建参考模型。





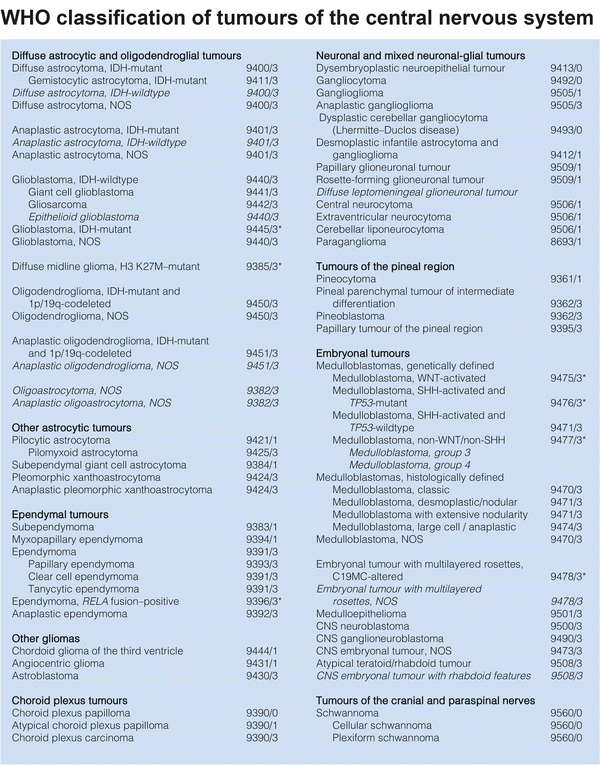
基本能实现一下结构的查询界面功能

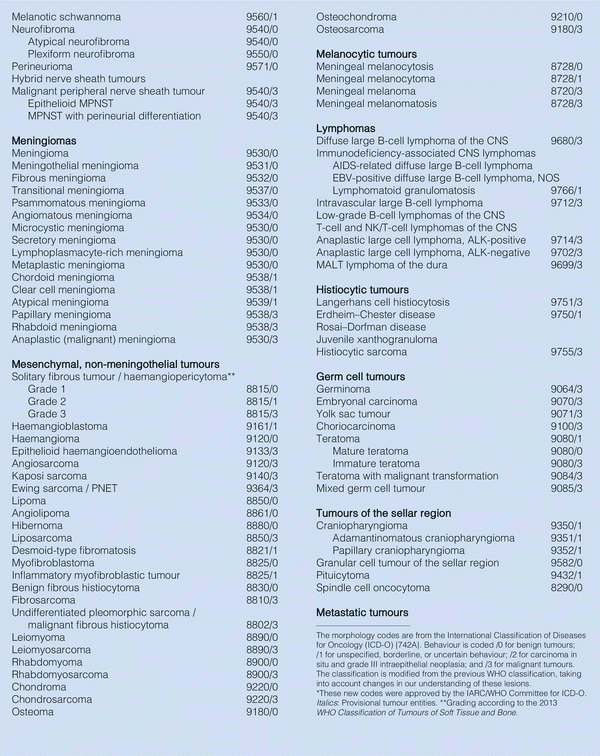


其中影像学表现包括影像学表现描述与图片，文献资料中的推荐文献即查资料时查到的参考文献，以及在WOS上检索得到与疾病名有关的最新20篇文献。

***※******注意相关医学术语要依据SNOMED CT标准录入。***

需要的疾病如WHO分类的神经中枢肿瘤项如下表，其中斜体标注的肿瘤项除外，其余项目都需收集详细资料。





**（2）知识库外部要求：**

知识处理的结果表现形式为知识库本体Owl文件，应用形式为利用owl文件围绕各自项目需求，基于JAVA或Python开发web项目，并提供知识库集成的API（各自owl的调用方法），为以后知识库集成做准备。

## 6.数据库关联

通过疾病名称可以与田敏、江超、李易时的知识库互联，通过病院名称可以朱安琪（地理）、李易时互联。

# 五、融合地理位置与领域知识的医疗健康信息推荐知识库（朱安琪）

## 1.选题题目

基于地名本体的地理位置与领域知识的医疗健康信息推荐系统。

## 2.数据来源

**1）数据来源**：

地理知识库的来源包括数据字典、地名词典、语义规则、领域专家知识、本体和元数据规范等。如USGS GNIS&GNS、ADL、GeoNames、Getty TNG、谷歌地图百度地图等地理知识信息源，还如可利用百度地图API获取具体的经纬度、邮编、地址、电话等相关信息。

**2）获取方式**：

网络获取

**3）大概的数据量**：

至少给出中国所有的肿瘤方面研究机构和病院等医疗健康机构的详细数据。

**4）数据格式**：

储存为.owl格式

## 3.开发语言、开发工具

**4.知识库结构**

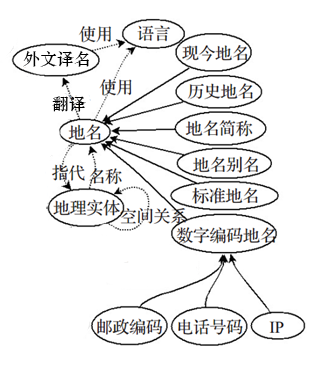
**（1）逻辑结构**

利用 Perez 和 Benjamin提出的五元组建模原语来表示地名本体

ToponymOntology =< TO\_C,TO\_R,TO\_F,TO\_A,TO\_I >

其中 TO\_C代表与地名相关的概念或者实体的集合；TO\_R代表地名实体或者概念间的关系的集合，表示概念之间相互作用；TO\_F代表一类特殊的关系地名函数的集合；TO\_A代表地名相关的永真断言的集合；TO\_I代表地名的实例的集合。

可参考以下地名本体构建模型（图左），展示了地名本体的概念及概念间关系，其中椭圆表示概念，实线箭头表示继承关系，虚线箭头表示关联关系。



**（2）物理结构**

使用本体作为知识库构建的主要工具，其技术可行性为：

（1）知识的明示化，本体对描述领域知识的相关概念及概念之间的联系都给出了明确的定义。

（2）知识的形式化，本体基于描述逻辑和框架模型，具有计算机可理解的表现形式。

（3）知识的规范化，本体提供了引用某个领域知识的标准方式。

（4）知识的共享化，本体表达的领域知识反映了不同应用对该领域的共同认知

（5）知识的推理化，本体可以通过规则语言形成推理规则，通过Jena语义框架可以实现对知识的推理、查询等等。

## 5.知识库的功能要求

（1）知识库内部要求：

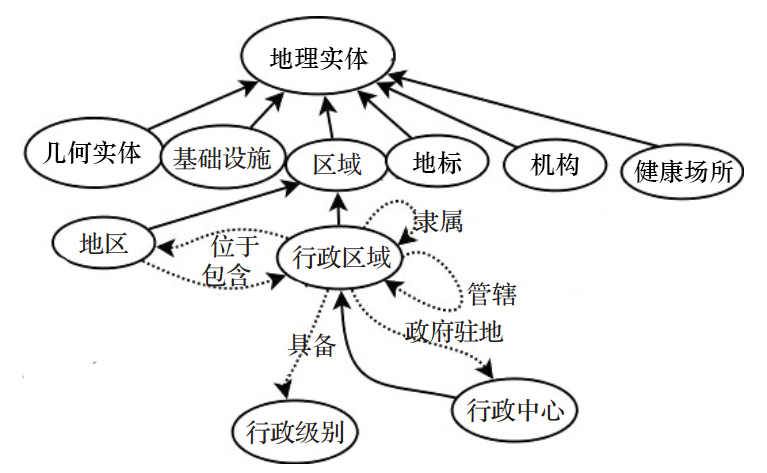
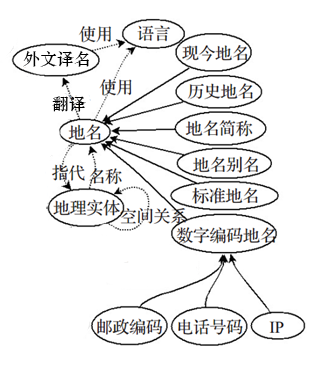
**主要功能：**

要求能够根据地理位置知识库中的医疗健康领域地名信息关联到医疗健康的一些领域知识。（具体领域知识详见另几个研究）。

级是构建基于地名本体的地理信息知识库，能够以地名实体为接口连接到相应的医疗健康领域知识。如搜索“珞珈山肿瘤”能够提取出珞珈山附近的有关肿瘤病院、疗养院、相关病院肿瘤科专家等信息，也能通过珞珈山推理出“武汉大学”得出研究机构为武大的相关肿瘤研究方面的领域知识（具体领域知识又其他几个库提供，主要是开发地名对应机构名称的接口）

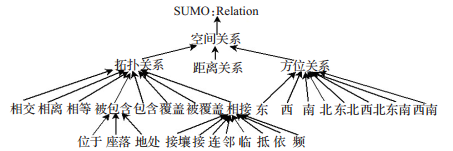
**模型构建要点：**

主要的本体模型构建如第四部分逻辑结构所示（左）。其中地理实体的构建可以参考图右的模型。

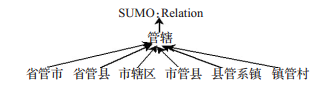


关于模型的一些具体要求如下：

* 地名：地名是人们对具有特定方位、地域范围的地理实体赋予的专有名称。其具有指位性和指类性。
* 标准地名：使用规范的语言文字书写，并经过官方认可的地名全称
* 地名简称：经过官方认可的约定俗成的地名的简单称谓
* 地名别名：某一地理实体的标准地名及其简称以外的其他现行名字
* 现今地名：目前正在使用的地名
* 历史地名：过去曾经使用过、目前已不再使用的地名
* 外文译名：对中文地名的外文翻译
* 地理实体：地理实体是指地球表层系统中与人类活动相关的物质实体，是地名所指代的具有一定空间位置或范围的实体。
* 几何实体：几何实体是应用几何思想对地理实体进行抽象成点线面体等基本几何要素组成的几何体。本数据库的几何实体利用 GML的几何模式来表述。
* 行政级别：行政级别是指行政区域在行政管理中所拥有的级别，在中国，它分为省级、地级、县级、乡镇级四级。
* 健康场所：健康场所是与医疗健康相关具体场所，须有医院、急救中心、敬老院、门诊部、药店、疗养院、药厂这些子分类，须有具体名称信息、地址信息、联系方式信息等。
* 机构：机构主要指医疗健康相关的学术机构，如学校学院研究所，须有具体名称信息、地址信息、联系方式信息等。
* 基础设施：主要指带基础交通设施，如长途汽车站、机场、加油站、停车场、火车站、码头、火车票民航售票处、公路、桥梁。
* 空间关系：空间关系是指地理实体之间存在的一些具有空间特性的关系，本文的地名本体中空间关系着重于拓扑关系、方位关系和距离关系。可参考下图



* 管辖关系：管辖关系是高行政级别的行政区域对低级别行政区域在政治、经济等方面的管理关系。在我国，管辖关系又分为省管市、省管县、市辖区、市管县、县管乡镇、乡镇管村五种，如图所示。管辖关系是一种传递关系



* 指代关系：地名具有一定的指代性，地名本体的指代关系是指人类理解的地名所表示的地理实体
* 名称关系：名称关系表示地理实体的名称是地名
* 隶属关系：隶属关系是地级行政级别的行政区域在地域、政治、经济等方面受高级行政区域的管理的关系的表示。隶属关系也是传递关系，同时隶属关系和管辖关系互为反关系

（2）知识库外部要求：

知识处理的结果表现形式为知识库本体Owl文件，应用形式为利用owl文件围绕各自项目需求，基于JAVA或Python开发web项目，并提供知识库集成的API（各自owl的调用方法），为以后知识库集成做准备。

## 6.数据库关联

通过具体机构名称/病院名称与田敏、李易时、岳雪琪的知识库相连。

# 六、基于网络医患交流数据的医患知识表示系统（李易时）

## 1.选题题目

基于网络医患交流数据的医患知识表示系统：采用来自互联网医疗咨询开放平台的医患交流数据，通过语义网络表示法，对医患交流过程中医生和患者的知识进行表示。系统构建了网络医患对话信息数据库，并实现自动对网络医患交流数据进行语义处理。

研究目标：该系统作为知识不对称研究的数据采集和预处理系统，设计用于知识不对称主题的静态研究和动态研究。研究目标是通过对网络医患交流的大数据进行挖掘，分析医患双方知识不对称的特征，探索网络医患交流知识不对称的规律。

研究内容：（1）知识不对称问题静态研究，将对话数据构建成医患双方的语义网络，通过语义网络特征的差异，比较分析双方的知识不对称，并探寻医患双方知识不对称问题的规律。（2）知识不对称问题动态研究，考虑时序性，相比于静态研究更关注话轮的顺序对知识不对称的影响，从知识构建的动态过程分析双方的知识不对称特征，并探寻医患双方知识不对称问题的规律。

## 2.数据来源

原始数据来源于互联网医疗咨询开放平台：春雨医生，按科室找医生，一级科室选择肿瘤及防治科，二级科室选择全部科室，地区选择全国，勾选“仅显示可咨询医生”。共40位医生，进入各个医生主页，展开“好评问题”并依次爬取其中的咨询对话数据。每位医生的好评问题展示最近的200条（20条\*10页），共计8000条原始数据。另：数据存在不定时更新

数据源网址：https://www.chunyuyisheng.com/pc/doctors/0-0-11/?is\_available=1&page=1

## 3.开发语言、开发工具

本知识库系统推荐python作为开发语言（可利用现成的网页爬虫代码）。

开发管理系统推荐GitHub。

文本编辑器推荐Vim和Sublime Text。

IDE推荐PyCharm。

数据存储用protégé存为OWL文件。

## 4.知识库结构

**（1）数据说明**

数据按加工程度可分两部分：一是采集自网络的医患交流和用户基本信息的数据（属于一手数据，对此部分数据处理只需进行简单的清洗和整理），需要通过互联网爬取，内容包括交流文本内容、患者特征信息和医生特征信息。二是对交流文本信息进行语义处理后的数据（属于二手数据，将对话语句进行规范化、结构化处理），需要通过自然语言处理，将复杂的、计算机不理解的文本内容按规则转化为计算机可读、可规范化存储的元组数据。

**（2）数据存储**

数据存储成owl文件的格式，用protégé软件编辑。（存储结构关系可参考 “肿瘤数据存储.owl”文件）。

爬取得到的数据包括交流文本内容、患者特征信息以及医生特征信息。交流文本内容包括每句发言的发言人，发言的文字信息，不存储包含语音内容的对话组合（此处进行数据清理）。患者特征信息主要有性别和年龄，医生的特征信息主要有所属医院科室、职称、咨询人数、好评率、学历、从业时间等信息。建议数据库存储格式如下：（id自定义）

Patient(Pid, Psex, Page)

Doctor(Did, Hospital, Title, Times, Evaluation, Education, Duration, Dage, Dsex)

Dialogue(Dnum, Pid, Did, Content)

**（3）数据处理**

参考自然语言理解的方法，对对话文本内容进一步处理，进行实体抽取和关系抽取（郭喜跃, 何婷婷. 信息抽取研究综述[J]. 计算机科学, 2015, 42(2):14-17.）。实体抽取主要是实体识别和谓词识别。实体抽取方法可按分词-词性标注-命名实体识别的步骤，对每一句对话进行处理和存储，存储形式为[实体、实体…[谓词]]，注意保留语序和发言人。（分词词表已有现成文件）。

在实体抽取基础上进行关系抽取，关系抽取方法可用基于SVM分类的方法，将谓词与语义关系进行匹配，存储形式为语义网络的形式。按照如下范式进行存储——

<语义网络>::=<基本网元>|Merge(<基本网元>,…)

<基本网元>::=<节点><语义联系><节点>

<节点>::=(<属性-值对>,…)

<属性-值对>::=(<属性名>:<属性值>)

<语义联系>::=<系统预定语义联系>|<用户自定义语义联系>

## 5.知识库的功能要求

本知识库分为两部分，包括基本数据和语义处理后的数据。

基本数据主要是患者于医生的特征数据，和对话原始文本。这部分的内部功能要求可以被其他库调用；可以根据患者id检索到其咨询的医生id、相应的对话文本；根据医生id检索到其回答的患者们的id以及相应的对话文本；要去知识库具有基础的统计和计算功能，能够实现统计医生完成的患者咨询数、计算不同特征医生数量等功能。

语义处理后的数据主要是对对话文本进行自然语言理解和语义关系提取。这部分首要要求是保留对话的话轮顺序；其次是要求把非结构化的口语对话转化为结构化存储的形式，略去不关键的文字；要求处理后的数据具有计算机可读易处理的结构；且和上文的原始数据知识库相互关联。

## 6.数据库关联

本知识库提供医生基本信息（包括姓名、性别、年龄、职称、学历、所属医院及科室等）、医生所属地区信息、医患交流信息（医生回答次数、好评率等）与朱安琪的地理知识库相关联。

在本体方面，与岳雪琪具有相同的属性如患者基本信息、医生基本信息等。

与谢祎玉使用同一分词词表。

# 七、心理健康网络咨询对话知识库（岳雪琪）

## 1.选题题目

心理健康网络咨询服务中的用户知识构建过程研究：抽取网络大数据，然后将咨询对话语义结构化处理，存储到知识库中，并提供查询及可视化功能。

## 2.数据来源

网络大数据（以网络健康问诊平台“春雨医生”中“心理科”中的咨询记录为例）。使用采集器抓取心理健康咨询对话内容，包含有多名医生及每个医生对应的多名患者之间的对话（至少6\*20段长对话，即6名医生及每名医生各自对应20名患者，每段对话不少于100条对话内容）。抓取的原始数据存储于TXT文档或Excel文档中。

## 3.开发语言、开发工具

## 4.知识库结构及功能要求

（1）采集：采集内容需包含有多名医生及每个医生对应的多名患者之间的对话（可使用八爪鱼采集器、火车头采集器等工具）。对医生及患者的身份及相关个人信息（性别、年龄等）进行提取并存储（根据平台上所提供的id确定为同一个人）并进行唯一的身份标识（如“d01”“p01”等）。

（2）语义处理：对发言内容按序进行语义分析，提取出关键词以及关键词之间的逻辑关系，用于构建表示患者知识的语义网络。对采集得到的对话内容，使用分词器进行分词（如庖丁解羊分词器），进行词性标注（如使用工具语料库在线），对患者的提问进行实体和谓词（提问对象，如“xx病的症状有哪些？”）或实体和实体（提问关系，如“xx药可以替代xx药么？”）的提取，并进行谓词消歧（使用谓词词典，如果提取谓词在谓词词典中，则直接使用该标准谓词；若不存在，则计算改谓词与谓词词典中每个词的语义相似度，取语义相似度最大的标准谓词）。实体和谓词分别作为语义网络的节点和联系，然后从医生的回答中提取内容对语义网络进行补充（补充另一节点或补充节点间关系）。

（3）存储：按照语义网络的基本网元，存储患者语义网络中的节点、节点间关系等信息。存储格式可参考工具语义网络语言SNet、语义场框架表、基元概念语义网络表等。需要标识哪些关系或节点是患者已知的，哪些是医生补充或是修改（如患者提出了自己的认识，即完整的节点-关系-节点，然后医生指出误区并纠正）的。可参考：

语义网络SN<P，D，E1，E2，……>（<患者id，医生id，基本网元1，基本网元2，…… >）

基本网元E<num，N1，R，N2>（<时序编号，节点1，关系，节点2>）

节点N<n\_name，n\_before>（<节点名称（咨询后），节点名称（咨询前）>）

关系R<r\_name，p\_before>（<关系名称（咨询后），关系名称（咨询前）>）

before为空代表该节点是患者咨询前未知的内容；name和before存储内容不同，代表医生对患者的认识进行了纠正。

将结构化的信息以owl文件存储，并提供原始文本及对应结构化存储内容的查看功能（语义网络可视化图像最优），可通过输入患者id或医生id进行查询。

## 5.数据库关联

医生以及患者基本信息：

医生（id、姓名、性别、年龄、职称、学历、所属医院及科室等）

患者（id、性别、年龄等）