## 第二章  系统需求分析

 电子病历信息管理系统，是一个跨学科、跨领域的项目，不仅需要了解医学方面的知识，还需要掌握软件开发技能。本系统根据2016年国家发布的《电子病历共享文档规范》，建立电子病历信息管理系统，对患者的病历进行提取、分析、存储和管理，利用数据库技术和文本处理技术，解析和存储电子病历；运用自然语言处理技术，对病历中的文本格式信息进行结构化处理，并通过json格式转换，进行存储和建模，供高效查询和展示；挖掘病历中的有用信息，进行二次使用，发掘其潜在价值，供医生和病人进行参考。

 系统的需求分析将系统按照功能进行划分，各个功能定义清晰，为后边的系统具体设计做好准备性工作。按照业务具体场景，将系统分为以下部分：电子病历管理功能、电子病历存储功能、电子病历展示系统。

### 2.1数据预处理

#### 2.1.1数据来源

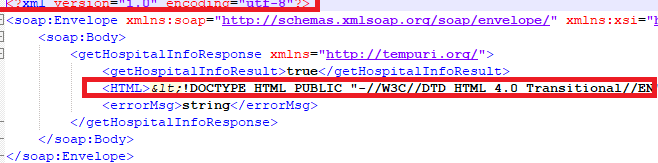
  电子病历中既有结构化数据（护理记录、病案首页等）又包含有大量的非结构化数据（入院记录、病程记录、出院记录等）。尤其是住院病历，其有用信息主要包含在非结构化的文本数据中，这些信息对于后续的辅助诊断、编码推荐（诊断编码和手术编码）、质控、支付保障等一系列研发起着至关重要的作用。所以如何有效提取、存储和管理这些非结构化数据，是处理电子病历数据的难题。

  本文数据主要来源于两个数据源。其一为来自于四川省中医院的9608份住院病历，格式为xml格式，每份病历主要包括入院记录、首次病程记录、病程记录、入院记录、出院证明书、病历评分表、病历评分表反面。其二来源于四川省石油医院435份电子病历，存储格式为SQL SERVER数据库，主要病历信息包含在两张数据表emr\_bl\_bl01和emr\_bl\_bl02中，数据表bl01为病历结构表，数据表bl02为病历内容表，其中病历内容存储格式为image数据类型，为16进制数据流。

#### 2.1.2数据解析

##### 一：xml型数据

根据数据保密性和中医院数据开放使用规范，我们拿到了中医院电子病历数据库的接口信息，通过接口调用，提取到了9066份2017-2019年三年病人的住院电子病历，其结构是xml格式，如图一所示，需要将其中的文本信息提取出来，存放到临时数据库中，以备后续处理使用。



图一：xml文件结构

中间<html>标签中包含有大量的文本信息，是我们最感兴趣的部分。针对数据特有得的格式我们进行了如下操作：

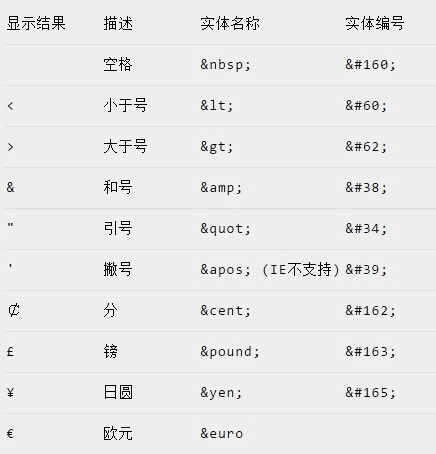
###### 数据脱敏和隐私保护

从接口提取到的数据中，包含有病人的个人信息，处于隐私保护目的，需要对病人信息设置权限，防止隐私泄露。

我们使用正则表达，将病历信息中得病人联系电话和病人得身份证号进行了字符替换，以做到隐私保护的目的。同时对于其他的与病人相关的基本信息，我们的临时数据仓库中，只保留病人的性别、发病节气、婚姻状况、民族、年龄等，此外我们用病案号来唯一标识病人

2.转义

提取到的xml源数据中，为了表示方便，将html标签转换为转义字符，其对照表达式如图二所示：



图二：转义字符

我们将转义字符转换为对应的html标签，供后续处理。

###### 3.去除xml外层格式，提取html标签内容

由于我们需要的信息全部包含在一对<html></html>标签中，外层的xml格式对所需提取的文本并无帮助，因此我们需要从xml格式中提取出html内容。

###### 4.解析病历

提取到的html格式文件，主要包含html标签、xml标签和文本信息。我们需要从其中提取出病人信息以及住院病历信息，病人信息主要包含在xml标签中，住院信息包含在html标签中，主要包括入院记录（包括主诉、现病史、既往史、个人史、月经及婚育史、家族史，中医望、闻、切诊，体格检查，辅助检查，初步诊断）、首次病程记录（包括病例特点，初步诊断，中医辨病辨证依据，西医诊断依据，鉴别诊断，诊疗计划）、病程记录（主要有查房记录，术后讨论等）、出院证明书、出院记录，包含有大量切复杂的文本信息，需要针对不同段落结构中的不同数据进行分级提取

###### 5.将提取的文本信息分部分存储到临时数据库中。

临时数据库的设计与构建，按照标准电子病历模板，将住院病历分为入院记录、首次病程记录部分、病程记录部分、出院记录、出院证明书，用来存储从文本中提取到的电子病历信息，下图表示入院记录示例：

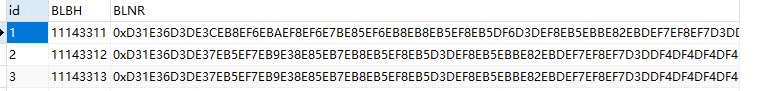


图三：临时数据库示例

这样，将复杂的电子病历提取到数据库中，一个字段代表一部分内容，上图中红色方框内表示入院记录中的主诉。

##### 二：数据库型数据

对于SQL SERVER数据库中的435份住院病历，其格式为数据库形式，病历内容和病人信息包含于数据库的表格中，病人的信息为结构化，可以直接冲表格中通过sql语句提取到临时数据库中，而病历内容为image数据类型，即十六进制数据流，如图所示：



图四：SQL SERVER数据库病历内容存储示例

要提取病历内容，首先需要将十六进制的数据流转换为原来的xml格式文本。转换后的结构如图所示：未完成

### 2.2 电子病历结构化

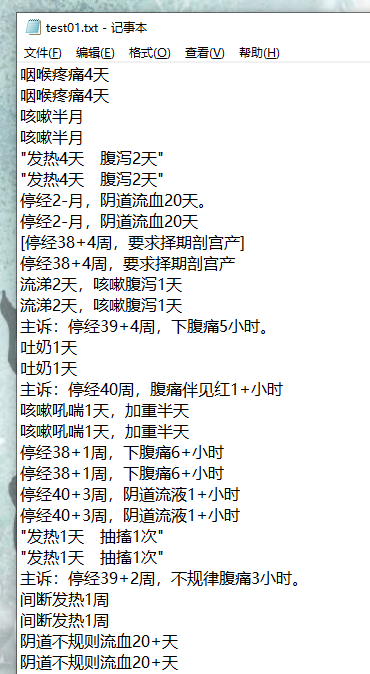
电子病历的结构化以及结构化之后提供的供计算机处理的信息是电子病历和传统的纸质病历的重要区别，只有经过结构化之后的电子病历才能够进行医学分析和临床决策支持，体现出电子病历的核心价值。只有经过结构化之后，才能进行语义层面的查询，数据才能更好的为科研所用。

临时数据库中，虽然将整份的电子病历，拆分成及部分，并且将每部分存储指定字段中，但仍是非结构化的，因为一个字段中存储了大量的长文本信息，不利于检索和后续使用。例如，我们要分析糖尿病的影响因素，在长文本形式的病历中，我们只能通过“糖尿病”这个关键词进行搜索，搜索结果中可能包含很多“无糖尿病”患者。而我们对病历结构化之后，可以对“糖尿病”指定一个标签：“1”代表患有糖尿病；“0”代表未患有糖尿病；“2”代表未作相关检查，不清楚是否患有糖尿病。这样，我们要查询糖尿病患者的病历，只需要指定该标签为1，大大提高了检索的效率和准确性。

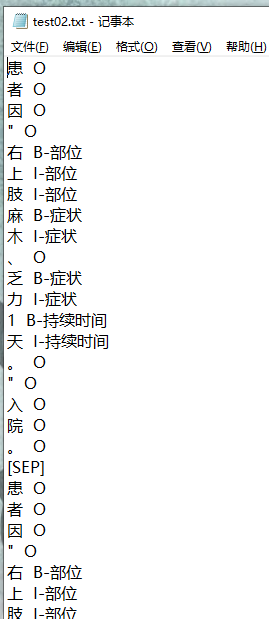
对于电子病历结构化，主要借助方法有NER、RE、正则等，其中最常用为NER和正则表达式。对于NER而言目前主流的模型有，Bi-LSTM+CRF，LatticleLSTM，BertNER（基于BERT的NER方法）。本文选取了基于BERT网络模型的NER方法，对电子病历提取出的长文本进行结构化。同时，对于任何机器学习模型或者深度学习模型，数据标注都毫无疑问是首先要解决的问题，我们分别对主诉、既往史、现病史、个人史、月经及婚育史、家族史、辅助检查、体格检查等进行标注，对于每个标注任务，我们分别设计其对应的标注规则，并训练BERT模型。

#### 2.2.1数据标注

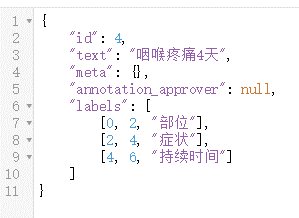
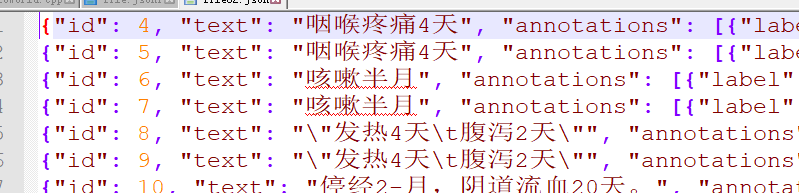
原始数据标注后处理：将原始的数据（未标注、未经模型预测）的数据导入标注工具中，标注后导出为JSON格式，后将JSON转为BIO格式，如下图所示：



原始未标注数据



模型训练数据的格式



标注工具导出数据格式

图四：标注数据说明

标注工作三人一组，两人负责标注，第三个人负责审阅，两人标注相同的部分不需修改，两人标注不相同部分，要按照标注规则仔细商讨修改。审阅完成后导入模型中进行训练。

#### 2.1.2 Bert模型训练

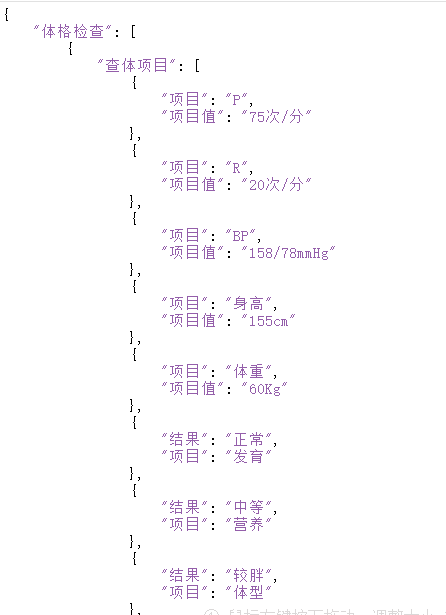
#### 2.1.3 格式转换

结构化的电子病历指从医学信息学的角度将以自然语言方式录入的医疗文书按照医学术语的要求进行结构化分析，并将这些语义结构最终以关系型结构的方式保存到数据库中。

由于模型的预测输出结果格式是“字+标签”的形式，因此需要对其进行组合拼接处理，使其还原成原始的完整词的形式。结构化工程的目的是使使得我们可以按照需求输入病历文本时，可以得到对应的结构化的NER实体词，并且这些实体词按照json格式进行返回。



图五：标注后数据

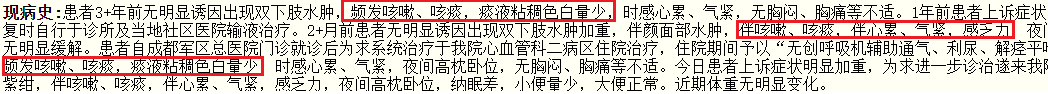


图六：格式转换后数据

## 第三章  数据存储

#### 3.1 标注后数据存储问题

如果同一个词出现在不同的句子中，经过BERT模型之后，如何能够确定这个词来自于那句话，是存储结构化的关键所在。如图七所示：

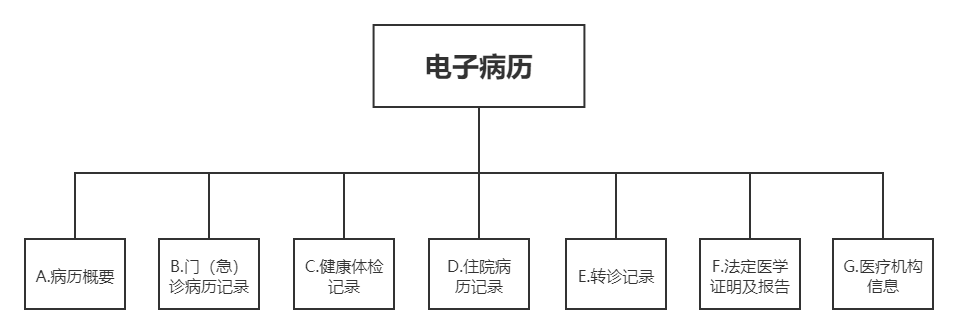


根据标注规则，上图中第一个红色方框中的内容：频发—状态，咳嗽—症状，咳痰—症状。第二个红色方框中：咳嗽—伴随症状，咳痰—伴随症状。第三个红色方框中：咳嗽—症状，咳痰--症状。我们发现同样的“咳嗽”、“咳痰”，在不同的语义环境中被备注为不同的标签，经过BERT模型后，这些字段被分割为独立的标签，如果随意还原，会造成语义歧义且不顺，那么如何确定此字段前后逻辑，继续有效的存储，是结构化电子病历的难点。。

#### 3.2 结构化存储

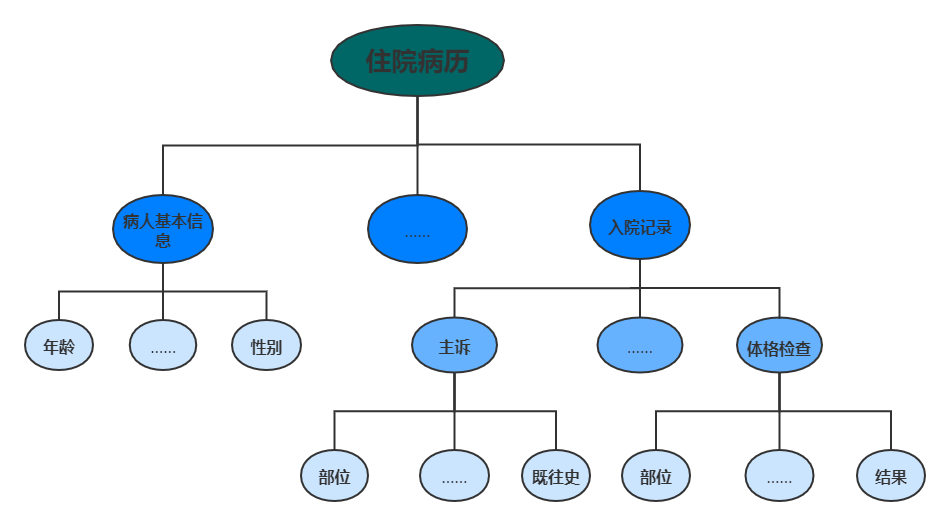
为了解决这一问题，我们将树的概念引入到电子病历中。树有且仅有一个根结点，有多个子结点，当子结点一下不再含有节点时成为叶子节点。一个结点可以包含一个或者多个叶子结点。将树的结构引入医学中，每个结点代表一个医学术语。结点之间的关系为父子关系，即从属关系，父结点包含子节点，子节点属于父结点，根结点只有子结点，叶子结点只有父结点。

根据电子病历的基本概念和系统架构，结合卫生部、国家中医药管理局关于《病历书写基本规范（试行）》、《电子病历共享文档规范2016》、《中医、中西医结合病历书写基本规范（试行）》的相关要求，将电子病历分为七个业务域的临床信息记录，如下图七所示：



图七：电子病历的基本内容

其中，D住院病历记录的概念树模型如图八所示：



图八：住院病历概念树

根结点表示住院电子病历，叶子结点为模型中的最小医学实体。第一层结点为“Ⅰ.病案首页”、“Ⅱ.病人基本信息”、“III.入院记录”、“Ⅳ.首次病程记录”、“V.病程记录”、“VI.出院证明书”、“Ⅶ .出院记录”，代表住院病历的段落结构。

段落结构Ⅱ“病人基本信息”包括：HIS号、病案号、性别、年龄、民族、婚姻状况、职业状况，其中HIS号用来唯一标识电子病历，病案号用来唯一标识病人。此段落对应的叶子结点用来描述病人信息，主要以结构化形式存在，直接从病历文本中提取。段落结构III、Ⅳ、V、VI、Ⅶ存在形式为长文本，也是我们系统重要处理的部分。

段落结构III.入院记录包括“”