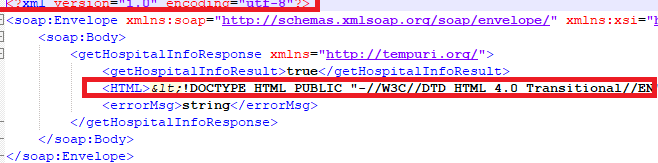
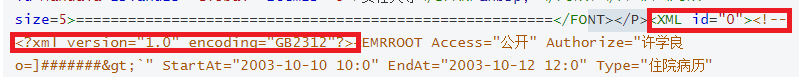
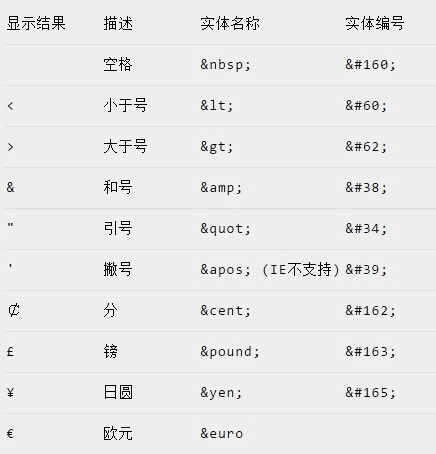
1.病历结构  
 病历基本结构是XML嵌套<html>标签，html标签中又嵌套xml





2.转义字符---以完成

P&gt;&lt;FONT color=steelblue&gt;&lt;STRONG&gt;&amp;nbsp;&amp;nbsp;&amp;nbsp;&lt;FONT color=#000&gt;主诉:&lt;/FONT&gt;&lt;/STRONG&gt;&lt;/FONT&gt;&lt;FONT color=steelblue&gt;&amp;nbsp;&lt;/FONT&gt;&lt;FONT color=#000000&gt;发现血压升高伴头晕30年，加重3天&lt;/FONT&gt;&lt;/P



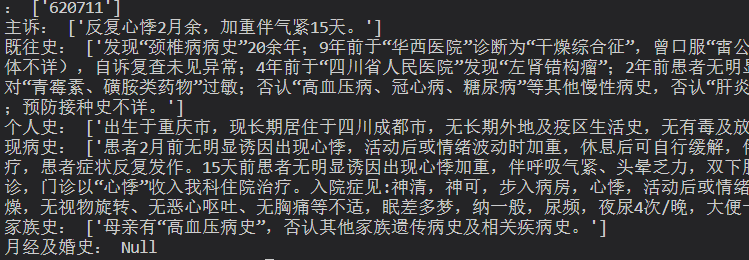
3.字体格式

病历中存在字体格式不同而导致结构不同



**<P><FONTcolor=steelblue><STRONG>&nbsp;&nbsp;&nbsp;</STRONG><FONT color=#000000><STRONG>主诉:</STRONG>反复咳嗽、咯痰、心累、气紧6年，加重3天&nbsp; </FONT></FONT></P>**

**<P><FONT color=steelblue><STRONG>&nbsp;&nbsp;&nbsp;<FONT color=#000>主诉:<FONT color=#ff0000>反复咳嗽、</FONT><FONT color=#4682b4><FONT color=#ff0000>心累、气紧6年，</FONT>双下肢水肿4年，加重12天</FONT></FONT></STRONG></FONT><FONT color=steelblue>&nbsp;&nbsp;</FONT></P>**



#### 4.中英文替换以及大量空格

病历中存在大量的空格符，为了显示格式。兵器存在中英文标点问题。

#### 5.病历所要提取的内容

包括结构化的字段和非结构化的一诉五史、病程记录等大段文本。

结构化字段：

其中包含患者及患者家属的个人信息，包括：姓名，住址，联系人姓名，联系人电话，公民身份证，电话。

其次，结构化字段中包含与患者和病房相关的字段信息：病案号、HIS内部标识符、入院中医诊断、入院主要诊断、首次病程记录、入院中医主病、入院中医主证描述、入院西医诊断、病区、入院日期、、入院科室、入院病区、出院日期、住院天数、死亡时间、死亡原因等。

Html中文本部分：初步诊断、入院日期、出院日期



