**中医病案文本挖掘**

**实验二 中医病案数据清洗**

**1、实验目的**

掌握病案数据的清洗方法，探讨数据清洗的规则。

**2、实验要求**

（1）分组，不多于3人；对病案数据一致性、缺失值、无效值、异常值、分解等处理，形成统一规范的数据。能用python编码的尽量用代码处理。

（2）报告撰写要求：小组分工情况；每个任务应按步骤、方法、处理过程及结果展示的结构来写，处理的结果应有相应文件存储，处理过程包含程序代码。上交文件包括报告和新的数据文件，数据文件格式不限。

**3、实验任务**

**（1）数据筛选：**根据主诉与现病史、自诉和症状两个属性，保留与由失眠相关症状的病人信息。

**（2）缺失值处理：**如主诉列，有些病人有，有些没有，根据现病史、自诉和症状，确定主诉；在现病史、自诉和症状列中，有些复诊病人的记录是病史如前又补充了一些内容，需把复诊病人的该列补充完整。

**（3）分解：**对处方中含有各字的中药分解，如煅龙牡是煅牡蛎和煅龙骨两味药；如焦楂曲是炒山*楂*和炒神曲两味药，等等。

**（4）一致性：**有复诊病人的“现病史、自诉和症状列”描述的一致性情况。各列的值是否描述了相应内容，如脉象、舌象是否混杂。

**（5）根据原始病案的数据情况，将数据整理为能够为后续特征提取和数据分析等处理带来方便的数据文件。**

作业提交时间：2020.4.07前。