# 数据清洗

数据清洗包括初步数据清洗，任务一数据清洗和任务二数据清洗。

## 全局数据清洗

全局数据清洗旨在将原始数据进行全局清洗，进而得到可适用于任务一、任务二、任务三和任务四的数据。其主要目的是消除数据中的错误、冗余和噪声，确保数据的完整性和标准数值化，从而有利于下一步的数据分析。

全局数据清洗的步骤包括数据剔除，数据完整化，数据数值化，数据标准化。这些步骤的详细意义和操作如下文所述。

### 数据剔除

数据剔除的目的是将原始数据中与所有任务都不相关的错误与冗余数据全部去除。分为两个方向：

1. 纵向特征剔除。

剔除与所有任务都不相关的纵向特征列。具体包含了表示数据来源的特征项（scite1, scite2,scite3,dbsource），与数值编码重复了的描述性语句的特征项（如targtype1\_txt 与targtype1意义重复了，且targtype1已经数值化了，所以targtype1\_txt可以删除），其他的一些不方便使用的描述性语句（如location, summary等）。具体见表1.

表1.

1. 横向样本剔除。

将基本上认为该事件不是恐怖主义行为的样本删除。即如果某事件的疑似恐怖主义（doubtterr）标记为‘0’，则表示基本上不怀疑该事件是恐怖主义行为，所以该样本属于错误样本，可以直接删除，这样更加有利于下一步的数据分析。

### 数据完整化

所有特征项中存在数据缺失的，统一标记为‘unknown’。

### 数据数值化

由于在后续数据分析中主要采用的是xgboost算法，而且本案例中的非数值数据是离散的，没有数值大小意义的，所以这里的数值化编码可以采用二进制编码格式。具体做法为：

对于非数值数据，如一个单词或一段描述性语句的字符串，先统计该特征项总共有多少个分布值f\_num, 然后一一对应地编号至Num = [0…f\_num], 再按照二进制编码格式将Num编码转化为Binary\_Num。

### 数据标准化

对于‘unknown’项，统一转化为数值‘-99’。

对于分布不均匀的数值特征，先统计该特征项总共有多少个分布值f\_num, 然后一一对应地编号至Num = [0…f\_num]。

# 问题一：依据危害性对恐怖袭击事件分级

人员伤亡和经济损失，时机、地域、针对的对象

地域： **region，provstate，specificity**

**911事件：5622，5623**