# 数据预处理大纲

## 分列解释

- 标题 - 删除 - 信息已由和后面概括。
- 案号 - 删除 - 信息已由和后面概括。
- 案件类型 - 保留 - one hot
- 庭审程序 - 保留 - one hot
- 案由 - 保留 - one hot
- 文书类型 - 保留 - 0/1
- 法院 - 保留 - [省份，城市，中/高级/其他]
- 判决日期 - 保留
- 原告 - 保留
    - one hot，自然人告？检察院告？法人告？
- 被告 - 保留
    - one hot，法人？自然人？
- 第三人 - 保留 - 0/1 有无自然人
- 法官，审判长，审判员，书记员 - 删除
    - Assumption：与这些人为因素无关
- 头部1&2 - 删除 概括了之前所有信息了
- **当事人1&2** - 保留
    - 重点NLP处理犯罪人的前科、出身等。
- **庭审程序说明1&2** - 保留
    - 介绍审理原由
- **庭审过程1-6** - 保留
- **法院意见1&2** - 保留
- **判决结果** - 保留
- **庭后告知**1&2
    - 是否终审？可否上诉？
- 结尾1&2 - 删除 - 多余信息
- 附录1&2 - 保留- 提取所用法律条文

## 按处理需要分类

### 1. one-hot encoding

| 列名 | **one-hot encodings** | 关键假设 |
| --- | --- | --- |
| 案件类型 | [刑事，民事，...] | |
| 庭审程序 | [刑罚变更，一审，二审，复合，其他] | |
| 案由 | [罪名...] | 罪名分类是无争议的 |
| 文书类型 | [判定书，裁定书] | |
| 原告 | [检察院，法人，自然人] | |
| 被告 | [法人，自然人] | |
| 第三人 | [有，无] | |
| 法院省市 | [按照表格分] | |
| 法院分级 | [人民法院，中级，高级] | |

## 2. NLP处理

| 列名 | 处理大纲 | 关键假设 |
| --- | --- | --- |
| 当事人 | 对于犯罪者本身形成one-hot:<br>1. 年龄婚姻<br>2. 犯罪前科 | 犯罪者前科、年龄、婚姻状况会影响犯罪<br>籍贯信息需进一步商讨是否加入。 |
| 庭审程序说明 | | |
| 庭审过程 | | |
| 法院意见 | | 其重要性还需专家访谈 |
| 判决结果 | 1. 提取所用到的法律条文 | |
| 庭后告知 | | |

## 3. 关键信息提取

附录 - 提取法律条文