

案由分类需求(刑民知行 4 分类)

此需求为刑民二分类补充。

1.数据描述

训练集需求数据字段包括案由、标题、正文等。数据量为刑事、民事（除知产外）、知产、行政各 10 万篇。案例最好是一审判决书、少量二审和终审，比例为 8:1:1。案由仅标注刑事、民事、知产、行政四类，不需要详细案由编号。

2.数据处理

正文字段去掉从法院判决段落之后的部分，举例：如“判决如下，依据《刑法》第一百条...”。如有纯文本字段则使用纯文本字段导出，即不含法宝之窗、法宝联想等内容。

3.子案由分布

补充可选择实现需求，刑事判决书约 480 个案由，如可将各子案由篇数提取一致则较为理想。如某个案由不足 200 篇，则随机复制补齐至平均水平。民事、知产、行政同样要求案由判决书平均。

4.数据存储

可存储至 mysql 数据库中。

5.测试集

另建一个测试集，要求与训练集相同，民事、刑事、知产、行政案例数为各 1 万篇，字段仍为案由、标题、正文。

6.需求日期

月底前。