**大数据智能量刑**

**数据介绍**

案件案情事实数据分为训练集和测试集两部分，格式为csv格式。其中以训练集数据为例，各个字段含义如下：

|  |  |  |
| --- | --- | --- |
| 字段 | 含义 | 格式 |
| ids | 案件id信息，由案件正文内容计算hash得到 | str |
| fact | 案件对应的案情事实 | str |
| criminal | 需要进行罪名预测和法律条文预测的目标嫌疑人 | str |
| accusation | 指控信息，即嫌疑人涉及的罪名信息 | str |
| articles | 嫌疑人涉及到的法律条文编号 | str |

其中，具有多个类别的数据通过英文封号“;”进行区分。

测试集数据与之类似，需要参赛选手针对accusation和articles字段两部分进行预测。

除案情数据外，针对法律条文数据给出编号对应的法律条文，条文根据《中华人民共和国刑法》整理，格式为csv格式，分隔符为英文逗号：“,”，各个字段含义如下：

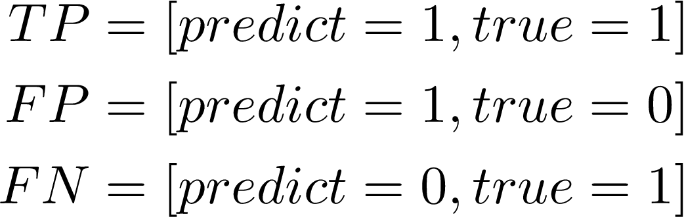
|  |  |  |
| --- | --- | --- |
| 字段 | 含义 | 格式 |
| article\_id | 法律条文对应编号 | str |
| article\_detail | 法律条文具体内容 | str |

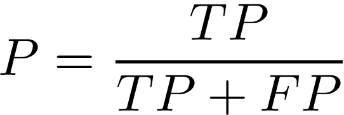
注意：如无特殊说明，数据集格式均为UTF-8无BOM格式。

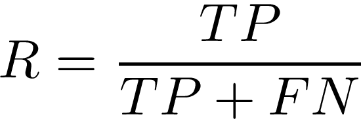
**评价指标**

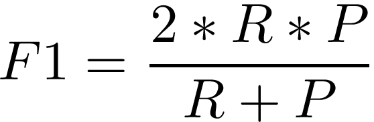
本次任务为嫌疑人罪名预测。

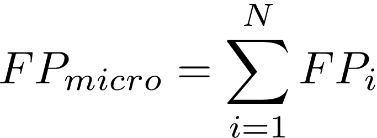
根据案情事实，针对嫌疑人罪名进行预测分类，本次任务选取案件类型较多的TOP30作为罪名空间。选手需要对罪名进行预测，每个案情事实可能对应多项罪名。评价指标使用宏平均的F1与微平均F1值加权的方式给出：

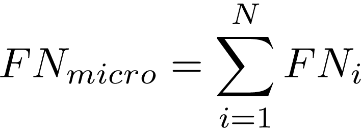


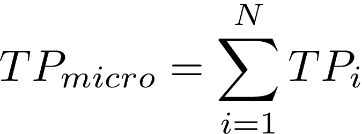


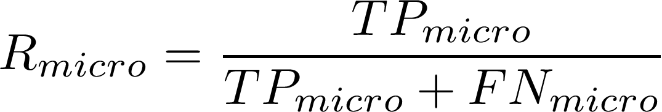
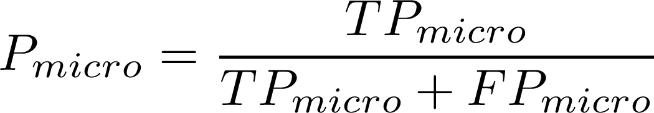


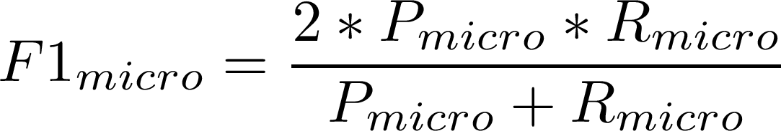


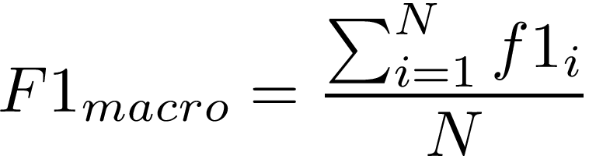














**提交格式**

提交格式参见test\_sample.csv文件，提交文件应为csv文件，首行为csv文件标题，使用英文逗号分隔，之后每行为一条案情事实预测的指控信息和法律条文结果。多条结果使用英文封号“;”进行分隔。

|  |  |
| --- | --- |
| ids | accusation |
| c84ffbe9e148db9fbf6c0e787ecc118d | 诈骗;贪污 |
| 662758700f0650d4523988f3e30d3067 | 交通肇事;诈骗 |
| 37b8606a5a6ddbc1acf49261db2ae42e | 故意毁坏财物 |
| ac62d0742a3b182123f3b11c6a3a77f7 | … |

**提交方式**

将结果文件以附件的方式发送至竞赛组委会邮箱:data@dlut.edu.cn。邮件标题格式要求：“队名”-“限选题1”-“结果提交”。例如：XXX队-限选题1-结果提交。若邮件标题未按要求书写，则提交结果无效。

限选题最终成绩，将取各参赛队伍提交的结果中评价指标最好的一组计算。初赛期间，各参赛队伍可反复提交结果，但每天只可提交一次结果。若一天内多次提交结果，则按时间最新的计算，其余将自动作废。

组委会将不定期在竞赛官网（bigdata.dlut.edu.cn）上公布已收到的结果评分和排名，请及时关注竞赛官网相关通知。