工作计划-编码

|  |  |  |  |
| --- | --- | --- | --- |
| 8.24-8.26 | 积累预备知识，了解课题背景和算法原理 | 完成 |  |
| 8.27 | 第一次汇报 | 完成 |  |
| 8.28-8.31 | 完成架构和初步分模块的编码工作 | 完成 |  |
| 9.1-9.2 | 初步测试并准备汇报 | 完成 |  |
| 9.3-9.5 | 导出模型，继续测试完善 | 完成 | 导出模型出现问题，工作延后 |
| 9.5-9.7 | 继续完善UI界面 | 完成 |  |
| 9.8-9.9 | 完善系统，准备最终答辩 | 完成 |  |

9.3

模型训练存在训练时间长的问题，修改训练集后压缩到了每次完整训练需1小时。同时，训练过程中会出现混淆矩阵不匹配的情况，经检查发现代码在随机选取训练集时有一定概率会漏掉一类，导致最后匹配，修改了训练代码后得到解决。

9.4

在完成模型的训练和导出读取功能，准备做单样本的预测时，我们发现存在两个问题：

其一是将样本做文本预处理后作为模型输入时，出现了特征数量不一致的报错。这是因为测试样本与训练集的文本特征确实不一定匹配，对此我们提出了两种解决方案：一种是在构造词频统计器时先读取原训练集进行fit\_transform，再对测试样本进行transform，最后导出词频矩阵；另一种是提前提取一个训练集所有关键词的文件，并作为一项补在测试样本中，这样直接对测试集fit\_transform时也能输出完整特征的词频矩阵。经测试两种方法是等价的，为了预测效率更高我们选用了第二种方法。

其二是即使补齐了特征，输入到模型中预测值也非常不准确。经检查发现是因为原参考代码在训练时对数据根据最小最大值做了标准化处理，而测试集没有训练集的最小最大信息，自然也无法做相同的缩放处理。查询资料发现我们使用的随机森林分类器是非线性的，本来就不需要做标准化，修改代码重新训练模型后该问题解决。