**人工智能原理文本分类实验验收**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 序号 | 学号+姓名 | 验收项目 | | | | | | 评分 |
| 训练集大小 | 测试集大小 | 字典处理 | 特殊处理 | 选择的分类器 | 性能 |
|  | 学号： 2021211184  姓名： 郭晨旭 | 共 10 类，  每类 8000 篇  □ 自编爬虫  □ 自行爬取  ☑ 已有语库  □ 混合 | 共 10 类，  每类 5000 篇  □ 自编爬虫  □ 自行爬取  ☑ 已有语库  □ 混合  ☑训练集/测试集之比  8:5 | ☑ 降维  维度 6507  方法  1.TF-IDF  2.卡方检验  ☑ 停用词是否合理问题： 分词后自己实现去除停用词,对老师给的停用词表进行了补充,并且增加了停用符号 | ☑ 改进加权  方法： TF-IDF和卡方相乘  ☑ 其他改进  方法：   1. 筛选文章长度 2. 选取词频大于30的词进行处理 | ☑ 自编贝叶斯  ○ 零概率处理  使用拉普拉斯处理零概率  ☑ SVM  ○ libsvm  ○ 其他 sklearn.svm  ○ 核 线性核/高斯核  ○ 调参 网格调参  ○ 交叉验证  □ 其他分类器 | 自编贝叶斯:  ◎最高的正确率=99.52%  ◎最高的召回率=98.30%  ◎最低的正确率=80.48%  ◎最低的召回率=82.46%  ◎平均正确率=91.62%  ◎平均召回率=91.21%  ◎训练时间=4.04s  ◎测试时间=65.73s  ◎死机 无  ◎问题 无  SVM:  ◎最高的正确率=99.00%  ◎最高的召回率=99.46%  ◎最低的正确率=92.28%  ◎最低的召回率=85.90%  ◎平均正确率=96.15%  ◎平均召回率=96.11%  ◎训练时间=176.05s  ◎测试时间=0.028s  ◎死机 无  ◎问题 无 |  |

评语：