**数据仓库与数据挖掘实验验收**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 序号 | 姓名+学号 | 验收项目 | | | | | | 评分 |
| 训练集大小 | 测试集大小 | 字典大小 | 特殊处理 | 选择的分类器 | 性能 |
|  | 2021211184 郭晨旭 | 使用清华数据集, 一共80000篇, 分为10类, 每类8000篇 | 使用清华数据集,一共50000篇, 分为10类, 每类5000篇 | 6507 | 1. 选取词频大于30的词进行后续操作 2. TF-IDF 3. 卡方检验 4. 增加停用词和停用符号 5. 贝叶斯使用拉普拉斯处理零概率问题 6. 混淆矩阵 7. SVM网格调参 | 自己实现朴素贝叶斯  使用sklearn.svm实现svm | 贝叶斯:   1. 训练时间: 4.13 s 2. 测试时间: 65.22 s 3. 整体准确率: 90.65% 4. 整体召回率: 90.24% 5. 整体F1: 90.45%   SVM:   1. 训练时间: 35.49 s 2. 测试时间: 0.03 s 3. 整体准确率: 93.81% 4. 整体召回率: 93.70% 5. 整体F1: 93.75% |  |