**人工智能文本分类实验验收**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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 |
| 14 | ➀学号： 2018210547  姓名： 胡天翼  分工： 语料库构建、数据预处理  ➁学号： 2018210074  姓名： 熊宇  分工： 朴素贝叶斯实现及SVM | 共10类，  每类5000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 1：1 | √ 降维  维度 2000  方法 TF-IDF词袋模型[[1]](#endnote-1)  √ 停用词是否合理问题： 利用jieba分词选择名词，自动去除停用词 | √ 改进加权方法： 利用TF-IDF值取对数运算[[2]](#endnote-2)  √ 其他改进  方法：  改进一：筛选文档长度[[3]](#endnote-3)  改进二：去除人名[[4]](#endnote-4)  该进三：词频选职责[[5]](#endnote-5)  改进四：使用词袋模型[[6]](#endnote-6)  √ 是否有混淆矩阵 | √ 自编贝叶斯  √ 零概率处理 | ◎最高的正确率=97.31%  ◎最高的召回率=94.14%  ◎最低的正确率=78.31%  ◎最低的召回率=74.56%  ◎平均正确率=89.03%  ◎平均召回率=88.51%  ◎训练时间=7秒  ◎测试时间= 1919秒  ◎死机：无  ◎问题：无 |  |
| √ SVM（选作）  √ libsvm  ○ 其他  ○ 核  ○ 调参  ○ 交叉验证  □ 其他分类器 | ◎最高的正确率=98%  ◎最高的召回率=96%  ◎最低的正确率=86%  ◎最低的召回率=86%  ◎平均正确率=92.7%  ◎平均召回率=92.7%  ◎训练时间=4秒  ◎测试时间=0秒  ◎死机：无  ◎问题：无 |  |

评语：

1. TF-IDF模型：TFIDF的主要思想是：如果某个词或短语在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。计算数据集中词频（TF）和逆文档频率（IDF），两者相乘得到TF-IDF。将文档去噪分词、提取名词，并处理为TF-IDF模型后，利用TfidfVectorizer方法得到指定维度的字典，并生成训练集、测试集词袋模型。引入逆文档频率规避了基础贝叶斯选择最大出现次数时可能导致的的弊端。 [↑](#endnote-ref-1)
2. 在此基础贝叶斯方法上我们利用TF-IDF进行改进，使用TF-IDF将测试集中单词出现的次数转为一个0~1的TF-IDF权值，它在原贝叶斯方法的基础上额外考虑的单词的出现频率（重要程度）。与此同时取对数减少乘法运算次数，并在对数内加1以处理零概率。因此，计算后验概率式子变为

   max{ (ln(n1 \* P(x1|yi) + 1) + ln(n2 \* P(x2|yi) + 1) + …… + ln(nn \* P(xn|yi) + 1)) \* P(yi) }（ni为当前测试集第i个词的TF-IDF值） [↑](#endnote-ref-2)
3. 改进一：对文档进行长度筛选，从爬取到的200万篇文档中，筛选长度为2000~3000字的文章进行处理，从而使得数据集文档词数接近。 [↑](#endnote-ref-3)
4. 改进二：在提取名词、去除停用词的基础上，利用Jieba得到的人名属性，去掉人名，减少无用名词的影响。 [↑](#endnote-ref-4)
5. 改进三：利用TF-IDF模型建立词典时，去除在少于0.1%的文档中出现的词，在高于15%的文档中出现的词。本实验共选择10类文本，每类占比10%，如果一个词在15%以上的文档中都出现，说明它在不止一个类别中多次，那么该词语不具有好的代表性，应该去掉。对于少于0.1%的词，同样不具有代表性。 [↑](#endnote-ref-5)
6. 改进四：使用词袋模型，方便持久化到本地 [↑](#endnote-ref-6)