Data Mining Homework2 总结报告

1、数据预处理

将文档分析和小写化，去掉数字、单字母项，使用NLTK库进行词性还原并去除非单词项，删除停用词。

2、统计词频

对18828个文档分别统计词频，结果存放在字典类型中并以josn文件格式保存于“word\_dict.json”文件中。



3、朴素贝叶斯分类

基本步骤如下：

STEP1：随机将20类文档每类分出20%作为分类测试集，其余为训练集。

STEP2：利用训练集计算20种分类单词概率分布。

STEP3：遍历测试集，计算测试文档在20种分类下单词概率累加结果。

STEP4：排序比较结算结果，确定文档分类。

重复以上步骤取得多次试验结果。

4、实验结果

通过相除取最大和相减取最小两种方式进行测试。结果如下：

相除取最大：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **测试序号** | **分类成功次数** | **测试文档总数** | **成功率** | **备注** |
| 0 | 2923 | 3759 | 77.76% |  |
| 1 | 2929 | 3759 | 77.91% |  |
| 2 | 2933 | 3759 | 78.02% |  |
| 3 | 2904 | 3759 | 77.25% |  |
| 4 | 2896 | 3759 | 77.04% |  |
| 5 | 2907 | 3759 | 77.33% |  |
| 6 | 2893 | 3759 | 76.96% |  |
| 7 | 2941 | 3759 | 78.23% | 最高 |
| 8 | 2928 | 3759 | 77.89% |  |
| 9 | 2871 | 3759 | 76.37% | 最低 |
| 10 | 2903 | 3759 | 77.22% |  |
| 11 | 2927 | 3759 | 77.86% |  |
| 12 | 2940 | 3759 | 78.21% |  |
| 13 | 2927 | 3759 | 77.86% |  |
| 14 | 2876 | 3759 | 76.50% |  |
| 15 | 2914 | 3759 | 77.52% |  |
| 16 | 2907 | 3759 | 77.33% |  |
| 17 | 2924 | 3759 | 77.78% |  |
| 18 | 2907 | 3759 | 77.33% |  |
| 19 | 2899 | 3759 | 77.12% |  |

相减取最小

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **测试序号** | **分类成功次数** | **测试文档总数** | **成功率** | **备注** |
| 0 | 2916 | 3759 | 77.57% |  |
| 1 | 2926 | 3759 | 77.83% |  |
| 2 | 2897 | 3759 | 77.06% |  |
| 3 | 2934 | 3759 | 78.05% |  |
| 4 | 2945 | 3759 | 78.34% |  |
| 5 | 2886 | 3759 | 76.77% |  |
| 6 | 2907 | 3759 | 77.33% |  |
| 7 | 2959 | 3759 | 78.71% | 最高 |
| 8 | 2916 | 3759 | 77.57% |  |
| 9 | 2897 | 3759 | 77.06% |  |
| 10 | 2900 | 3759 | 77.14% |  |
| 11 | 2885 | 3759 | 76.74% |  |
| 12 | 2887 | 3759 | 76.80% |  |
| 13 | 2869 | 3759 | 76.32% |  |
| 14 | 2926 | 3759 | 77.83% |  |
| 15 | 2925 | 3759 | 77.81% |  |
| 16 | 2886 | 3759 | 76.77% |  |
| 17 | 2887 | 3759 | 76.80% |  |
| 18 | 2872 | 3759 | 76.40% | 最低 |
| 19 | 2879 | 3759 | 76.58% |  |

吉晓辉

2018年11月25日