# 日志分类模型训练与测试

## 环境介绍

**1.python：2.7.14**

**2.scikit-learn：0.19.1**

**3.nltk:3.2.4**

## 准备日志数据

**1.异常日志准备**

**1)搜集java和spring框架全部的异常（包含error和exception）**

**类型;**

**2)根据异常类型去检索对应的带堆栈信息的内容，将其作为异常日志数据集。**

**2.正常日志准备**

**1)搜集linkdata系统的ERROR级别的日志（日志尽量覆盖到所有的情况）;**

**2)通过tf-idf去统计词频并根据词频清洗日志，词频可以根据训练模型的反馈不断的去进行调整。**

**最后，正常日志与异常日志的合集作为最终的训练集与测试集。**

1. **正常日志处理的具体流程：**
2. **使用getMoreDataFromES方法查询出es中的正常数据集；**
3. **计算词频CountVectorizer；**
4. **通过wordcount方法获得词频统计列表，数据结构为（包含该词的文档数，词，所有出现改词的文档列表[(文档index,文档id)]）；**
5. **使用delword方法获取要删除的高频词所对应的文档id列表（指定词频）；**
6. **使用delid删除高频词所对应的文档id列表（指定保留词频）；**
7. **重复以上步骤不断更新训练的数据。**

**具体代码位置：**

**app/jupyter\_test/error\_predict/esdata\_normal\_preprocessing.ipynb**

## 文本向量化

**文本数据属于非结构化的数据，一般要转换成结构化的数据，方能进行实施机器学习算法实现文本分类。**

**常见的做法是将文本转换成『文档-词项矩阵』。矩阵中的元素，可以使用词频，或者[TF-IDF值](http://link.zhihu.com/?target=https://en.wikipedia.org/wiki/Tf%E2%80%93idf" \t "https://zhuanlan.zhihu.com/p/_blank)等。**

### 1.数据预处理

**1）使用nltk的word\_tokenize()进行文本切分；**

**2）去除停用词，停用词可以使用nltk和sklearn的词库，也可以根据文本内容自定义词库；**

### 2.文本特征提取

**使用sklearn.feature\_extraction.text的TfidfVectorizer将文本转换成词频矩阵。**

**分析转化后的结果，可以根据实际情况去进一步调整参数，**

**1）扔掉出现过于频繁。而又对匹配相关文档没有帮助的词语；**

**2）扔掉出现频率非常低。仅仅有非常小可能出现在文档中的词语；**

## 分类器训练及[交叉验证预测](http://www.cnblogs.com/taceywong/p/5930498.html" \o "发布于2016-10-04 15:19)

1. **训练** **KNeighborsClassifier和****DecisionTreeClassifier等分类器；**
2. **通过 sklearn.model\_selection中的cross\_val\_score进行交叉验证，计算各个训练模型的准确率和标准偏差。**

## 检测linkdata异常日志

降低正常日志中高频词的文档数，并检测正常日志与异常日志的成功与失败情况，如下表所示，

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 训练 | | | | 测试 | | | | | |
| 正常数据词频/个 | 正常训练数据/条 | 异常训练数据/条 | 准确率（标准偏差） | 正常测试数据总数  /条 | 正常检测成功/条 | 正常检测失败  /条 | 异常测试数据总数/条 | 异常检测成功  /条 | 异常检测失败/条 |
| 20 | 1618 | 804 | KNN: 0.680165 (0.429971) CART: 0.811579 (0.254661) | 10780 | 10721 | 33 | 378 | 58 | 320 |
| 10 | 1043 | KNN: 0.578431 (0.456046) CART: 0.760214 (0.261020) | 10693 | 61 | 61 | 317 |
| 5 | 692 | KNN: 0.479423 (0.456621) CART: 0.700541 (0.258434) | 9671 | 1083 | 61 | 317 |

注:KNN: KNeighborsClassifier；

CART:DecisionTreeClassifier。