九、建立垃圾邮件筛选系统

主要分为以下几个步骤：

1. 准备数据集：根据老师给的数据集网站，下载了其中的第一个集合，总共包含了3672个正常邮件和1500个垃圾邮件。为了数据的均匀性，从正常邮件中随机抽选了1500个。随后将1500个正常邮件和1500个垃圾邮件都分成1200：300，前者作为训练集，后者作为测试集。现在我们得到了一个2400容量的训练集和一个600容量的测试集，其中的正常邮件和垃圾邮件都是对半分布的。

2. 创建词典：邮件从第三行开始才是正文，针对每封邮件从第三行开始统计正文中出现的各个单词频率，创建一个词典。再根据已建的词典删去那些非文字类的符号，比如标点符号或者特殊字符。留取频率前3000的单词。

3. 特征提取：创建词典除去删除非文字类内容，还有另外的作用，就是作为特征——这里使用的特征就是单词频率。对于训练集，统计每一篇邮件的词频，就得到了一个1\*3000的向量，然后训练集一共包含2400封邮件，就是2400\*3000的矩阵。

4. 分类器：这里分别使用python当中scikit-learn库当中的朴素贝叶斯分类器和支持向量机方法。对于标签的获取可以从文件名当中匹配，正常邮件文件名包含’\*ham\*’，垃圾邮件则是’\*spam\*’。现在对于训练集有了特征矩阵，有了对应的标签，直接进行训练即可。

5. 性能测试：我们的测试集包含300封正常邮件，300封垃圾邮件，分别使用朴素贝叶斯和支持向量机方法得到的正确率如下：

|  |  |  |
| --- | --- | --- |
| Multinomial NB | Ham | Spam |
| Ham | 284 | 16 |
| Spam | 114 | 186 |

|  |  |  |
| --- | --- | --- |
| SVM (linear) | Ham | Spam |
| Ham | 258 | 42 |
| Spam | 45 | 255 |

可以看出，朴素贝叶斯的方法更擅长与找到正常邮件，但是对于垃圾邮件的识别错误率非常高，与之对比的则是支持向量机的方法二者都比较均衡。因此筛选垃圾邮件的时候更推荐使用SVM。

6. 评价总结：整个系统的思想都很简单，有很多细节可以提高。比如（1）训练集的数量：训练集的数量一方面直接影响分类器的训练，另一方面可以确保词典的全面性，词典不全面会导致测试集当中的一些单词无法进行统计，影响精度；（2）训练集的预处理：根据训练集得到的词典处于前面的基本都是‘the’，‘to’这种停用词，如果可以增加训练集的数量，可以同时进行停用词的排除，这样可以保证信息的纯度。