**实验Naive Bayes**

姓名: 李昀哲 学号： 20123101

**1 题目和数据：**

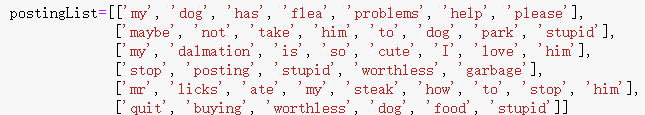
题目：Naive Bayes

* 基于朴素贝叶斯算法的言论过滤器
* 基于朴素贝叶斯算法实现垃圾邮件过滤

数据描述:

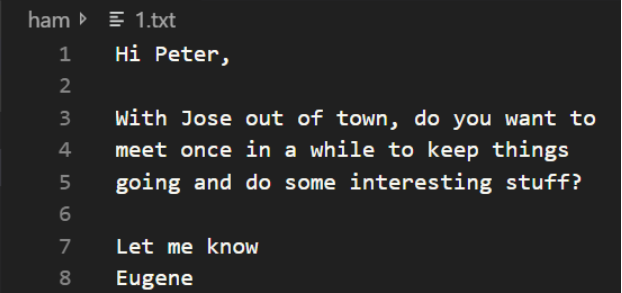
* **言论过滤器**

一组实验样本作为词条库，用以训练分类器，如图所示：



* **垃圾邮件过滤**

bayes\_email目录下有两个文件夹：ham和spam，其中spam文件下的txt文件为垃圾邮件。



**2 算法：**

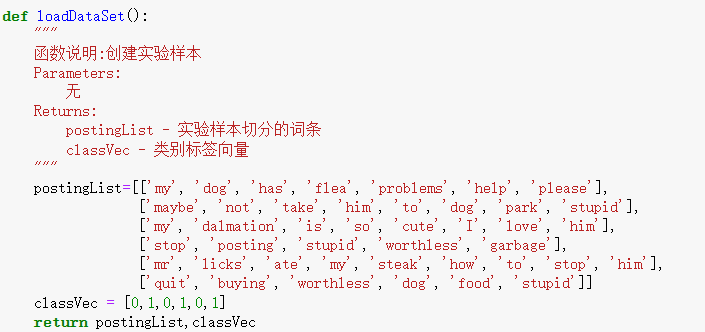
朴素贝叶斯法（Naive Bayes model）是基于[贝叶斯定理](https://baike.baidu.com/item/%E8%B4%9D%E5%8F%B6%E6%96%AF%E5%AE%9A%E7%90%86/1185949?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/%E6%9C%B4%E7%B4%A0%E8%B4%9D%E5%8F%B6%E6%96%AF/_blank)与特征条件独立假设的分类方法，和决策树模型相比，朴素贝叶斯分类器(Naive Bayes Classifier 或 NBC)发源于古典数学理论，有着坚实的数学基础，以及稳定的分类效率。同时，NBC模型所需估计的参数很少，对缺失数据不太敏感，算法也比较简单。理论上，NBC模型与其他分类方法相比具有最小的误差率。但是实际上并非总是如此，这是因为NBC模型假设属性之间相互独立，这个假设在实际应用中往往是不成立的，这给NBC模型的正确分类带来了一定影响。

其优点是：朴素贝叶斯算法假设了数据集属性之间是相互独立的，因此算法的逻辑性十分简单，并且算法较为稳定，当数据呈现不同的特点时，朴素贝叶斯的分类性能不会有太大的差异。换句话说就是朴素贝叶斯算法的健壮性比较好，对于不同类型的数据集不会呈现出太大的差异性。当数据集属性之间的关系相对比较独立时，朴素贝叶斯分类算法会有较好的效果。

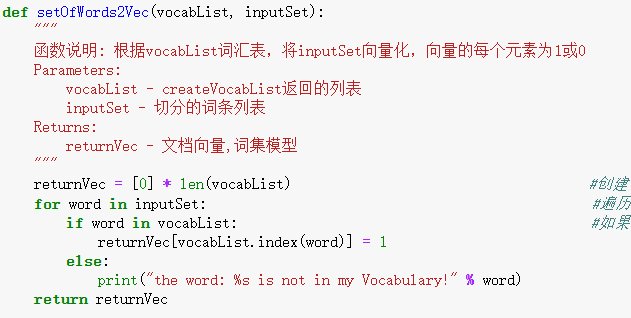
缺点是：属性独立性的条件同时也是[朴素贝叶斯分类器](https://baike.baidu.com/item/%E6%9C%B4%E7%B4%A0%E8%B4%9D%E5%8F%B6%E6%96%AF%E5%88%86%E7%B1%BB%E5%99%A8/22657525?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/%E6%9C%B4%E7%B4%A0%E8%B4%9D%E5%8F%B6%E6%96%AF/_blank)的不足之处。数据集属性的独立性在很多情况下是很难满足的，因为数据集的属性之间往往都存在着相互关联，如果在分类过程中出现这种问题，会导致分类的效果大大降低。

**3 代码及结果**

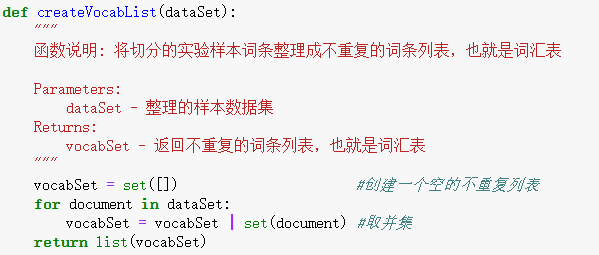
**数据集读入**



**生成词条向量**



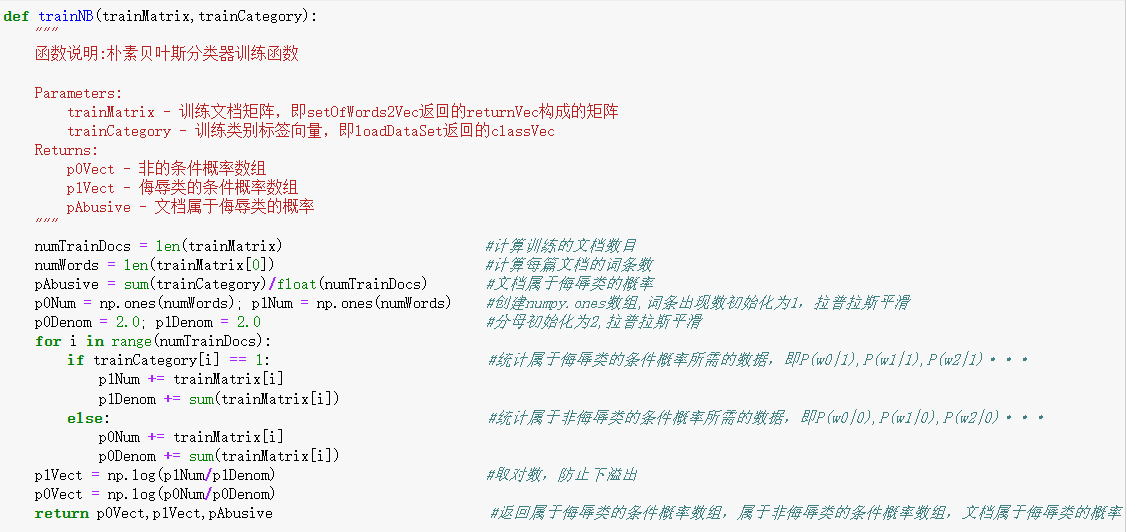
**创建词汇库**



**言论过滤**

利用贝叶斯分类器对文档进行分类时，要计算多个概率的乘积以获得文档属于某个类别的概率。如果其中有一个概率值为0，那么最后的乘积也为0。为降低这种影响，可以将所有词的出现数初始化为1， 并将分母初始化为2。这种做法就叫做拉普拉斯平滑(Laplace Smoothing)，是比较常用的平滑方法。

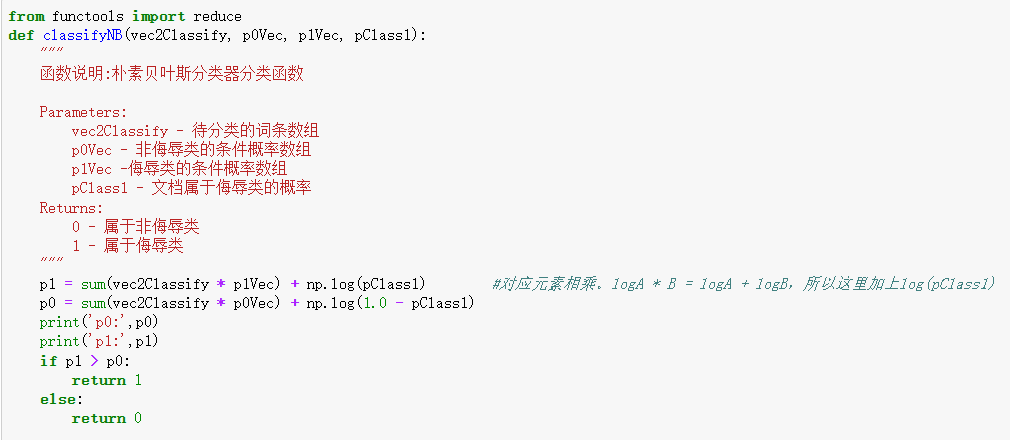
下溢问题。由于太多很小的数相乘会导致下溢问题。两个小数相乘，越乘越小，这样就造成了下溢出。在程序中，许多很小的数相乘，最后四舍五入计算结果可能就变成0。为了解决这个问题，对乘积结果取自然对数。通过求对数可以避免下溢出或者浮点数舍入导致的错误。



P(词条属于侮辱类|测试词条中每个词汇的分类)是要求的结果，P(词条中每个词汇的分类|词条属于侮辱类)可以计算出来，P(词条中每个词汇的分类)是提前计算好的，P(词条属于侮辱类)也是已知的0.5。最终计算结果，比较两个结果大小，值更大就表示测试词条属于该类。公式如下所示：



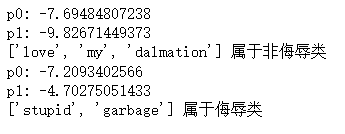
**朴素贝叶斯分类**



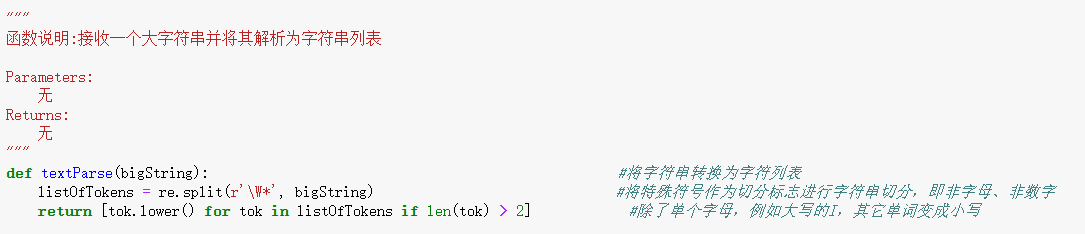
**测试**



**分类结果**



**文本解析**



**过滤测试**



**测试结果**

