**基于概率论的分类——朴素贝叶斯**

1. 基本知识
2. 贝叶斯决策理论核心思想：选择具有最高概率的决策

如果，那么类别为1

如果，那么类别为2

2．之前我们通过频数概率来进行分类，它只从数据本身获得结论，并不考虑逻辑推理及先验知识。这里，我们引入先验知识和逻辑推理来处理不确定命题。

3. 条件概率

事件A在事件B发生的条件下发生的概率，即：

可以交换条件概率的条件与结果：

4. 真正需要比较的是，数据点来自的概率是多少？数据点来自的概率又是多少？

但是我们已知的只有，可以交换概率中的条件和结果：

则贝叶斯分类准则为：

如果，那么类别为1

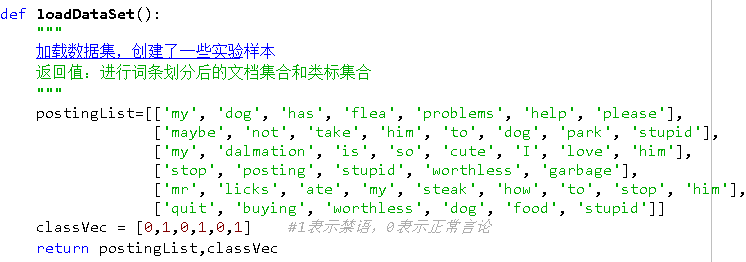
如果，那么类别为2

5. 朴素贝叶斯采用了“属性条件独立性假设”，对于已知类别，假设所有属性相互独立，假定每个特征独立地对分类结果产生影响，即每个特征同等重要。

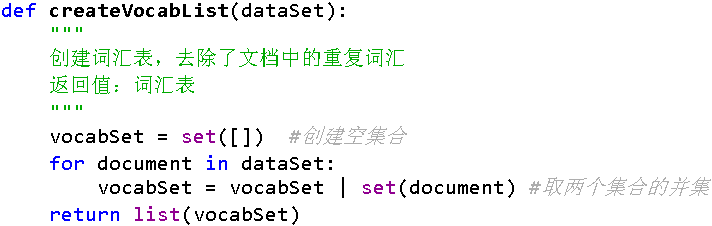
二、 使用Python进行文本分类

从文本中获取特征，首先拆分文本，将每个文本片段表示为一个词条向量，其中值为1表示词条出现在文档中，0表示词条未出现。

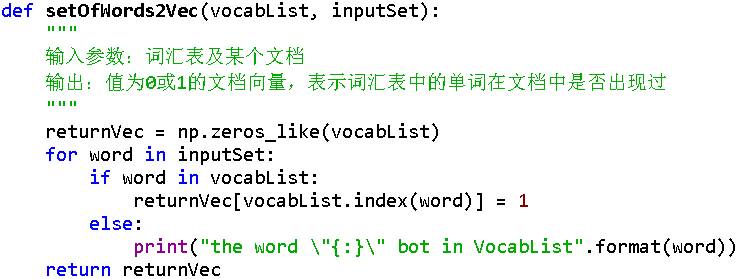
1. 首先将文本转换为数组向量



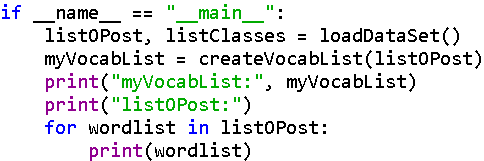
1. 创建一个包含在所有文档中出现的不重复词的列表，使用set集合即可解决这个问题。

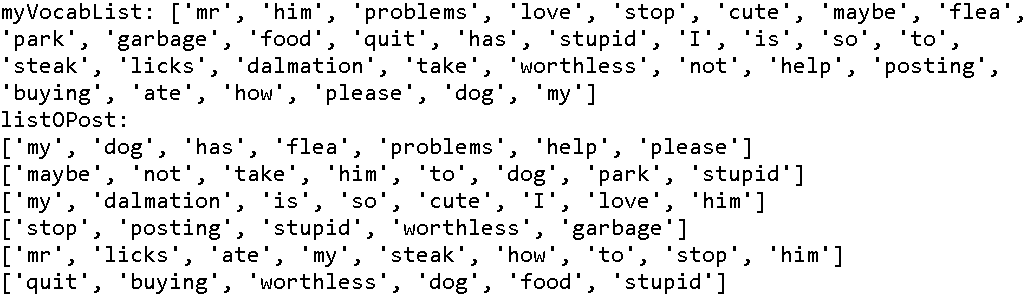


1. 词汇表中的单词在输入文档中是否出现过

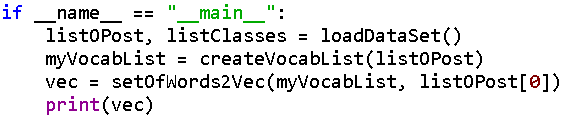


1. 查看词汇表：





1. 查看setOfWords2Vec()的运行效果：





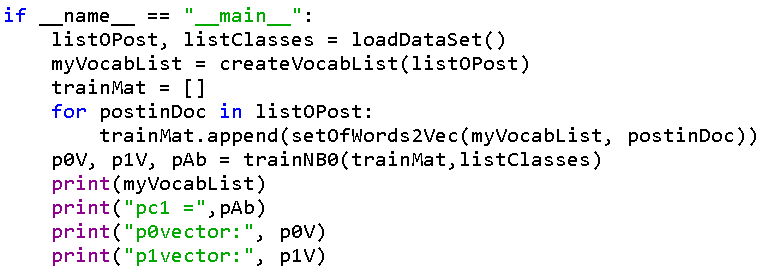
三、训练算法：从词向量计算概率

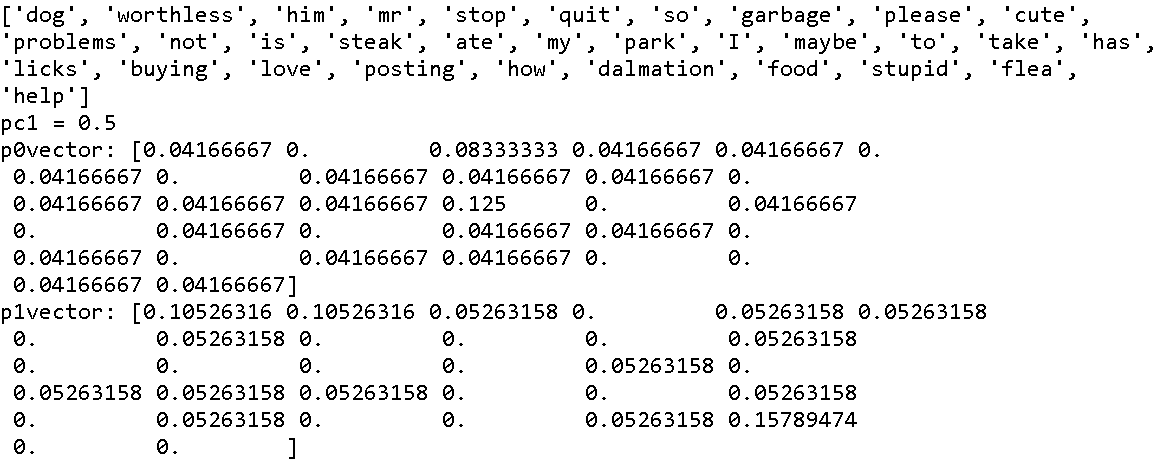


P1Vect是在已知是侮辱性句子的前提下，各个词汇出现的概率向量

P0Vect是在已知是非侮辱性句子的前提下，各个词汇出现的概率向量

PAbusive是一个句子是侮辱性句子的概率，即



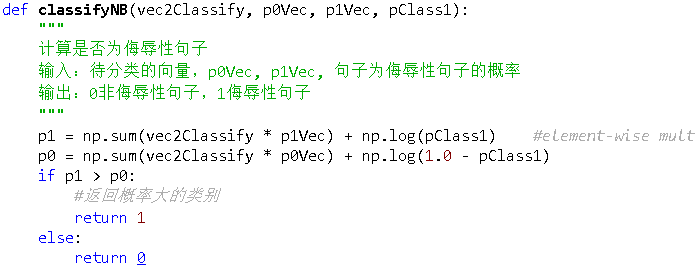


查看结果，可以发现一个句子是侮辱性句子的可能性为0.5，我们一共有6句话，其中3句为侮辱性句子，由此可以得知计算正确。Dog在非侮辱性句子中只出现了1次，对应的概率为0.04166667。在p1Vector中，可以看到倒数第三个数的值是该向量中的最大值，由此可以推断stupid是最能表征侮辱性句子的单词。

四、测试算法：根据现实情况修改分类器

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，即计算。如果其中有一个概率为0，则所有乘积的概率均为0。为了消除这种影响，可以将所有词的出现数初始化为1，并将分母初始化为2。

同时，有可能出现下溢出问题，由于太多很小的数相乘造成。可以尝试使用自然对数来解决这个问题。在代数中有ln(ab)=ln(a)+ln(b)。通过求对数可以避免下溢出或者浮点数舍入导致的错误。同时，采用自然对数进行处理不会有任何损失。





封装所有操作，以便于查看实现结果。

