山东大学 计算机科学与技术 学院

信息检索 课程实验报告

学号：201600150109	姓名：沈棋韬	班级： 16 人工智能班
实验题目：NBC 朴素贝叶斯分类器		
实验内容： • 实现朴素贝叶斯分类器，测试其在 20 Newsgroups 数据集上的效果。		
实验过程中遇到和解决的问题： （记录实验过程中遇到的问题，以及解决过程和实验结果。可以适当配以关键代码辅助说明，但不要大段贴代码。） 问题一：一开始使用空间向量模型，包含这个词则写它的频数，不包含这个词则写 0，导致每个文本都必须包含接近 2 万个维度，因为很稀疏，所以占用空间大并且效率低。后来用字典，只记录包含的词和对应的出现次数，减少占用内存并且提高效率。 问题二：因为要测试一篇文本与每一个分类的相似度，需要建立每个分类的词袋模型。我使用了字典，分别记录每个分类包含 feature 中的词，和对应 的频数。 问题三：在进行朴素贝叶斯分类的时候，按分类来进行，计算每一个分类时，先按顺序取类，找出对应的文件夹名，再对这个文件对于每一个分类计算可能性，把可能性最大的首选。一开始没有取对数，因为 python 数据精度原因导致每个分类的可能性都为 0。在取了 log 之后可以正常分类。		

```
In [15]: test_result
Out[15]:
[[175, 1, 0, 0, 0, 0, 0, 0, 2, 0, 0, 1, 0, 0, 2, 4, 0, 0, 0, 6],
 [1, 177, 8, 8, 2, 7, 3, 0, 0, 1, 0, 4, 2, 3, 1, 1, 0, 0, 1, 0],
 [0, 9, 206, 28, 6, 11, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0],
 [0, 5, 7, 185, 17, 2, 8, 1, 0, 0, 0, 1, 2, 0, 0, 0, 0, 0, 0, 0],
 [1, 2, 3, 8, 215, 0, 5, 0, 0, 0, 0, 0, 5, 2, 1, 0, 0, 0, 0, 0],
 [0, 18, 2, 4, 3, 201, 0, 1, 2, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0],
 [0, 1, 1, 11, 4, 1, 196, 7, 1, 1, 0, 1, 6, 0, 4, 0, 0, 0, 1, 0],
 [0, 1, 1, 1, 1, 0, 5, 231, 8, 4, 0, 0, 6, 0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 1, 1, 0, 1, 4, 251, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 [0, 1, 0, 0, 0, 0, 3, 1, 1, 222, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0],
 [0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 235, 0, 0, 0, 0, 0, 0, 0, 2, 0],
 [0, 1, 3, 0, 3, 0, 0, 1, 0, 0, 0, 255, 1, 0, 0, 0, 4, 0, 4, 1],
 [0, 8, 0, 12, 12, 2, 6, 2, 0, 0, 1, 2, 204, 0, 0, 0, 0, 0, 0, 0],
 [1, 3, 0, 0, 0, 0, 1, 0, 2, 1, 0, 0, 2, 224, 1, 0, 2, 1, 1, 2],
 [0, 6, 0, 0, 0, 2, 0, 1, 0, 0, 0, 0, 0, 0, 236, 0, 0, 0, 0, 1],
 [10, 0, 1, 0, 3, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 233, 1, 0, 1, 5],
 [0, 0, 0, 0, 0, 0, 0, 2, 2, 0, 0, 2, 0, 0, 0, 0, 229, 0, 8, 0],
 [2, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 243, 4, 1],
 [1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 2, 0, 0, 2, 1, 12, 2, 166, 2],
 [22, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 7, 6, 0, 3, 114]]
```

图中每一行代表每一个样本分类，每一列表示被分成了哪一类。因此对角线上的数字越大表示正确率越高，别的数字越小表示错误率越低。

	precise	recall	score
alt.atheism	82.2	91.6	86.6
comp.graphics	75.6	80.8	78.1
comp.os.ms-windows.misc	88.4	77.4	82.6
comp.sys.ibm.pc.hardware	71.7	81.1	76.1
comp.sys.mac.hardware	79.9	88.8	84.1
comp.windows.x	88.9	85.9	87.4
misc.forsale	84.8	83.4	84.1
rec.autos	90.2	89.5	89.9
rec.motorcycles	93.0	97.3	95.1
rec.sport.baseball	96.5	96.5	96.5
rec.sport.hockey	99.2	97.5	98.3
sci.crypt	94.8	93.4	94.1
sci.electronics	88.3	81.9	85.0
sci.med	97.0	92.9	94.9
sci.space	95.2	95.9	95.5
soc.religion.christian	94.3	91.4	92.8
talk.politics.guns	89.8	94.2	92.0
talk.politics.mideast	98.8	96.4	97.6
talk.politics.misc	86.9	86.5	86.7
talk.religion.misc	86.4	74.0	79.7

图中显示了每一个分类的精确度，召回率和调和平均数作为最后的分数。

$$F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$

结论分析与体会：
一开始实验的正确率很低，因为低频词去除的不够多，导致不能有效分类，最后把在词袋中次数小于 10 次的词去除以后，词袋的大小缩小了好几倍，说明有很多词都是没有辨识度但是在影响分类效果的。