

Homework2:NBC

姓名：刘敏

学号：201814820

- **实验要求：**

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：**

朴素贝叶斯算法实现：每个测试样本属于某个类的概率 = 某个类中出现样本中单词的概率的乘积（类条件概率）* 某个类出现的概率（先验概率）

即 $p(\text{cate}|\text{doc})=p(\text{word}|\text{cate})*p(\text{cate})$

具体计算类条件概率和先验概率时，朴素贝叶斯分类有两种模型：

① 多元分布模型：以单词为粒度，不仅仅计算特征词出现/不出现，还要计算出现的次数

类条件概率 $p(\text{word}|\text{cate})=(\text{类 cate 中单词 word 出现在所有文档中的次数之和}+1)/(\text{类 cate 中单词总数}+\text{训练样本中不重复的特征词总数})$

先验概率 $p(\text{cate})=\text{类 cate 中单词总数}/\text{训练样本中特征词总数}$

② 伯努利模型：以文件为粒度

类条件概率 $p(\text{word}|\text{cate})=(\text{类 cate 中出现 word 的文件总数}+1)/(\text{类 cate 中文件总数}+2)$

先验概率 $p(\text{cate})=\text{类 cate 中文件总数}/\text{整个训练样本中文件总数}$

为了实现更好的分类结果，我们将采取细粒度也就是**以单词为粒度**的多元分布模型构造朴素贝叶斯分类器，其实现过程如下所述：

1. 由 `getCateWordsFre(data_path)`函数统计训练样本中每个类中单词总数以及每个单词出现的次数,分别存储在字典 `cateWordsNum` 和 `cateWordsFre` 中。

其中, cateWordsNum 中 <key, value> 对表示 <类别, 单词总数>;

cateWordsFre 中 <key, value> 对表示的是 <类别_单词, 该单词出现的次数>

2. computeCateProb(trainCate, testFilesWords, cateWordsNum, trainTotalNum, cateWordsFre)方法则实现了类条件概率以及先验概率的计算, 从而得到 $p(\text{cate}|\text{doc})$ 。需要注意的是计算概率时用到了平滑技术以及取对数的操作
3. 然后使用 NBProcess(train_path, test_path, classifyResultByNB)函数对测试样本进行分类, 也就是将该文档在某类别下具有最高概率值的类别作为其最终分类结果, 并将结果写入文件 classifyResultByNB.txt 中, 用于统计准确率
4. 最后就是计算 NBC 模型的准确率了, 由函数 computeAcc(rightCate, resultCate)实现, 这里用到测试样本中文档本身的类别, 通过将其与预测类别进行比较, 得到分类正确的数目, 以计算准确率。

实验结果:

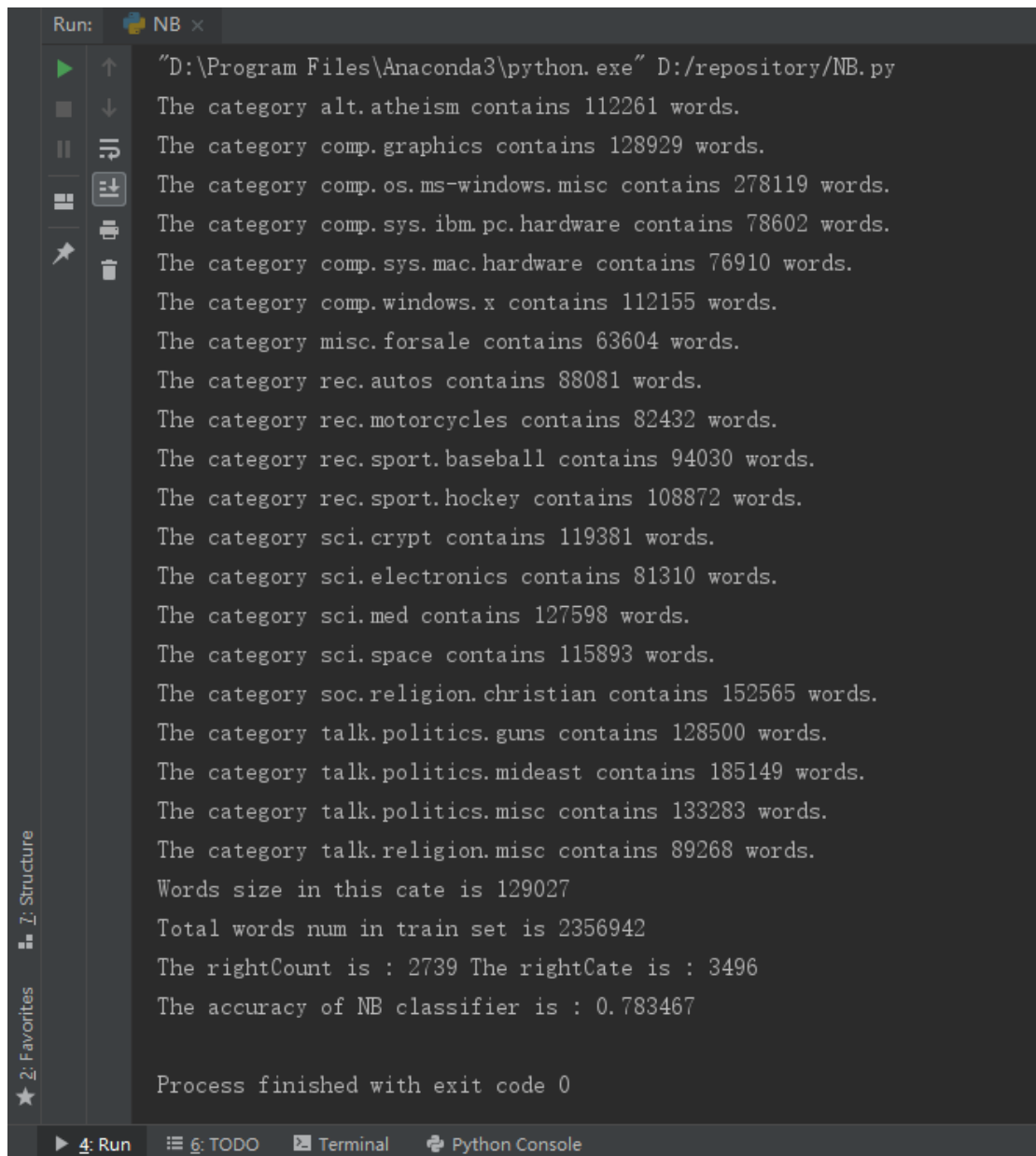
部分测试样本自身类别 from classifyRightCate0.txt

```
Preprocess.py × createDict.py × NB.py × classifyRightCate0.txt ×
4      51121 alt. atheism
5      51122 alt. atheism
6      51123 alt. atheism
7      51124 alt. atheism
8      51125 alt. atheism
9      51126 alt. atheism
10     51127 alt. atheism
11     51128 alt. atheism
12     51130 alt. atheism
13     51131 alt. atheism
14     51132 alt. atheism
15     51133 alt. atheism
16     51134 alt. atheism
17     51135 alt. atheism
18     51136 alt. atheism
19     51139 alt. atheism
20     51140 alt. atheism
21     51141 alt. atheism
22     51142 alt. atheism
23     51143 alt. atheism
24     51144 alt. atheism
25     51145 alt. atheism
26     51146 alt. atheism
27     51147 alt. atheism
28     51148 alt. atheism
29     51149 alt. atheism
30     51150 alt. atheism
31     51151 alt. atheism
32     51152 alt. atheism
33     51153 alt. atheism
34     51154 alt. atheism
35     51155 alt. atheism
36     51156 alt. atheism
37     51157 alt. atheism
38     51158 alt. atheism
```

部分测试样本预测类别 from classifyResultByNB.txt

Preprocess.py ×	createDict.py ×	NB.py ×	classifyRightCate0.txt ×	classifyResultByNB.txt ×
1	51060	soc.religion.christian		
2	51119	soc.religion.christian		
3	51120	soc.religion.christian		
4	51121	talk.politics.misc		
5	51122	alt.atheism		
6	51123	alt.atheism		
7	51124	alt.atheism		
8	51125	alt.atheism		
9	51126	alt.atheism		
10	51127	alt.atheism		
11	51128	alt.atheism		
12	51130	alt.atheism		
13	51131	alt.atheism		
14	51132	alt.atheism		
15	51133	alt.atheism		
16	51134	alt.atheism		
17	51135	alt.atheism		
18	51136	alt.atheism		
19	51139	alt.atheism		
20	51140	talk.politics.guns		
21	51141	alt.atheism		
22	51142	alt.atheism		
23	51143	alt.atheism		
24	51144	talk.politics.guns		
25	51145	talk.politics.guns		
26	51146	alt.atheism		
27	51147	alt.atheism		
28	51148	talk.politics.mideast		
29	51149	alt.atheism		
30	51150	alt.atheism		
31	51151	talk.politics.misc		
32	51152	alt.atheism		
33	51153	alt.atheism		
34	51154	alt.atheism		
35	51155	alt.atheism		

实验运行结果展示：（最终分类准确率为 0.78）



```
Run: NB x
"D:\Program Files\Anaconda3\python.exe" D:/repository/NB.py
The category alt.atheism contains 112261 words.
The category comp.graphics contains 128929 words.
The category comp.os.ms-windows.misc contains 278119 words.
The category comp.sys.ibm.pc.hardware contains 78602 words.
The category comp.sys.mac.hardware contains 76910 words.
The category comp.windows.x contains 112155 words.
The category misc.forsale contains 63604 words.
The category rec.autos contains 88081 words.
The category rec.motorcycles contains 82432 words.
The category rec.sport.baseball contains 94030 words.
The category rec.sport.hockey contains 108872 words.
The category sci.crypt contains 119381 words.
The category sci.electronics contains 81310 words.
The category sci.med contains 127598 words.
The category sci.space contains 115893 words.
The category soc.religion.christian contains 152565 words.
The category talk.politics.guns contains 128500 words.
The category talk.politics.mideast contains 185149 words.
The category talk.politics.misc contains 133283 words.
The category talk.religion.misc contains 89268 words.
Words size in this cate is 129027
Total words num in train set is 2356942
The rightCount is : 2739 The rightCate is : 3496
The accuracy of NB classifier is : 0.783467

Process finished with exit code 0
```

总结：朴素贝叶斯算法比较简单，主要是在多元分布模型和伯努利模型的选取上，我选择了多元分布模型，就是以单词为粒度计算分类概率，相较于以文件为粒度的伯努利模型，本模型具有更细层次的粒度，所以我认为其分类效果应该是优于伯努利模型的，只是没有进行实验验证，之后有时间会试一下。