The Second Data Mining Experimental Report

学 号：201834862 姓 名：窦金峰

一、 实验任务

1、借助第一次数据处理代码，进行预处理

2、实现朴素贝叶斯分类器，测试其在 20 Newsgroups 数据集上的效果。

- The **20 Newsgroups dataset** is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups.
- **20news-18828.tar.gz** - 20 Newsgroups; duplicates removed, only "From" and "Subject" headers (18828 documents)

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

http://qwone.com/~jason/20Newsgroups/

二、 实验过程

1、dp: 对于数据集进行预处理，分词：主要分为采用正则表达式及 split 进行分词、并将所有非字母符号作为分割符号；单词处理：去掉长度小于 3、去掉停用词、复 数变单数、去掉文档中 frequency 小于等于 15 的单词，并生成 word_doc_tf, word_idf, doc_word_tf

2.bys：利用 sklearn.model_selection 模块实现了随机划分 80%训练集和 20%测试集， 利用下式构建贝叶斯。多项式模型重复的词语视为其出现多次，在统计与判断时，都关注重复次数。

□ 对于多项式模型，P(“正规发票”|S) 的一种平滑算法是：

$$P(\text{“发票”}|S) = \frac{\text{每封垃圾邮件中出现“发票”的次数的总和}+1}{\text{每封垃圾邮件中所有词出现次数（计算重复次数）的总和}+\text{被统计的词表的词语数量}}$$

三、 实验结果

实验结果稳定在 86.8%左右。

```
C:\Users\Donna\Anaconda3\envs\newen\
  import imp
begin
10 0.9
20 0.75
30 0.7333333333333333
40 0.8
50 0.82
60 0.85
70 0.8714285714285714
80 0.85
90 0.8555555555555555
100 0.85
110 0.8636363636363636
120 0.8666666666666667
130 0.8769230769230769
140 0.8785714285714286
150 0.88
160 0.88125
170 0.8823529411764706
180 0.8833333333333333
190 0.8842105263157894
200 0.88
210 0.8809523809523809
220 0.8863636363636364
```

```
3600 0.8691666666666666
3610 0.8695290858725762
3620 0.8693370165745856
3630 0.8691460055096418
3640 0.868956043956044
3650 0.8684931506849315
3660 0.8685792349726776
3670 0.8686648501362397
3680 0.86875
3690 0.8682926829268293
3700 0.8686486486486487
3710 0.868733153638814
3720 0.8690860215053764
3730 0.8694369973190349
3740 0.8697860962566845
3750 0.8696
3760 0.8696808510638298
```

三、总结

1、在处理平滑方面，比较伯努利、多项式两种方法，最后选择多项式进行处理达到了良好的效果。

2、这次代码的构建在看过尹老师给的实践视频后更加快速的理解需求。