实验报告

201844911 杨子玉

1.实验数据来源：20news-18828.tar.gz - 20 Newsgroups

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

2.相关方法：

1)TF-IDF 是一种用于信息检索与数据挖掘的常用加权技术。TF 意思是词频(Term Frequency)，IDF 意思是逆文本频率指数(Inverse Document Frequency)。

2)VSM：把对文本内容的处理简化为向量空间中的向量运算

3)KNN: 邻近算法， K 最近邻，就是 k 个最近的邻居的意思，说的是每个样本都可以用它最接近的 k 个邻居来代表.

3.预处理文本数据集:

1)将实验数据分成两部分：80%的 data_train 和 20%的 data_test

2)对文本进行分词、大小写进行统一以及词干提取分析，去除停用词等处理

3)对词频大于 9 小于 10000 创建字典 dictionary.csv

4.得到每个文本的 VSM 表示：

遍历文本数据，计算 TF-IDF 值，得到每个文本(包括训练数据和测试数据)的 VSM 向量表示

5. 实现 KNN 分类器，测试其在 20 测试数据上的准确率

对训练数据形成 KNN 分类器，选出其中距离最近的 k=40 个样本，返回类别标签，其中出现次数最多的标签为预测结果。根据预测结果与其本身的类别进行比较，得到准确率。

6.实验结果如下图所示

形成的准确率大都在 0.75 以上

```
In [1]: runfile('C:/Users/Administrator/Desktop/xu/vsm+knn.py',
wdir='C:/Users/Administrator/Desktop/xu')
Divided into two parts
train_set_end
test_set_end
```

```
1 Accuracy:     0.7201383346634743
2 Accuracy:     0.7201383346634743
3 Accuracy:     0.7334397446129289
4 Accuracy:     0.7499334929502527
5 Accuracy:     0.7517956903431764
6 Accuracy:     0.759244479914871
7 Accuracy:     0.7613727055067837
8 Accuracy:     0.7648310720936419
9 Accuracy:     0.7749401436552275
10 Accuracy:    0.7770683692471402
11 Accuracy:    0.7741420590582602
12 Accuracy:    0.7773343974461293
13 Accuracy:    0.7781324820430966
14 Accuracy:    0.7799946794360202
15 Accuracy:    0.7826549614259112
16 Accuracy:    0.7834530460228785
17 Accuracy:    0.7805267358339985
18 Accuracy:    0.7826549614259112
19 Accuracy:    0.782388933226922
20 Accuracy:    0.782388933226922
21 Accuracy:    0.7829209896249002
22 Accuracy:    0.782388933226922
23 Accuracy:    0.7813248204309656
24 Accuracy:    0.7834530460228785
25 Accuracy:    0.7829209896249002
```

```
25 Accuracy:    0.7829209896249002
26 Accuracy:    0.7815908486299548
27 Accuracy:    0.7831870178238893
28 Accuracy:    0.7821229050279329
29 Accuracy:    0.7839851024208566
30 Accuracy:    0.7866453844107475
31 Accuracy:    0.7855812716147912
32 Accuracy:    0.7842511306198457
33 Accuracy:    0.7818568768289439
34 Accuracy:    0.7826549614259112
35 Accuracy:    0.7847831870178239
36 Accuracy:    0.7842511306198457
37 Accuracy:    0.7858472998137802
38 Accuracy:    0.7855812716147912
39 Accuracy:    0.7845171588188348
40 Accuracy:    0.785049215216813
41 Accuracy:    0.782388933226922
42 Accuracy:    0.7810587922319766
43 Accuracy:    0.7831870178238893
44 Accuracy:    0.7842511306198457
45 Accuracy:    0.7845171588188348
46 Accuracy:    0.7845171588188348
47 Accuracy:    0.7837190742218675
48 Accuracy:    0.7842511306198457
49 Accuracy:    0.7837190742218675
```