实验二　朴素贝叶斯分类器

在实验一中，对文档集 20 Newsgroups dataset 建立了向量空间模型，本实验在实验一的基础上用朴素贝叶斯分类器实现了对文档的自动分类。

1.数据划分。为避免实验一中文件读取上的麻烦，直接从 sklearn.datasets 上获取 20newsgroups 文档集，并用 sklearn.model_selection 中的 train_test_split 函数实现对训练集和测试集的划分，这里训练集设为 80%，测试集设为 20%。

```
#训练数据与测试数据的划分
from sklearn.datasets import fetch_20newsgroups
from sklearn.model_selection import train_test_split

newsgroups = fetch_20newsgroups()
X_train,X_test,y_train,y_test=
train_test_split(newsgroups.data,newsgroups.target,test_size=0.2,random_state=1)
```

其中 random_state 表示随机状态，这里设为 1，表示划分训练集和测试集的第一种随机状态。

2.Tfidf 向量化处理。分别对所划分的训练集和测试集的文档建立 Tfidf 权重的向量空间模型，同实验一。

```
#Tfidf 向量化处理
from sklearn.feature_extraction.text import TfidfVectorizer

vectorizer = TfidfVectorizer(stop_words='english', lowercase=True)
#以 X_train 建立字典并向量化
train_vector = vectorizer.fit_transform(X_train)
test_vector = vectorizer.transform(X_test)
```

3. 训练并预测。用 sklearn.naive_bayes 中的 MultinomialNB 函数实现多项式模型的分类算法，对测试数据中的 0 频次项，采用拉普拉斯平滑，平滑因子 alpha 设置为 0.01，预算数据类别先验设置为 false，训练完成后直接对上一步得到的测试数据 Tfidf 向量化后的向量 test_vector 进行预测。预测结果为以测试集文档数目为长度的列表，每个元素为对每个文档所分的类别对应的索引。

```
#多项式模型分类

from sklearn.naive_bayes import MultinomialNB
#训练 拉普拉斯平滑参数 alpha 设为 0.01
mnb_clf = MultinomialNB(alpha=0.01, fit_prior=False)
mnb_clf.fit(train_vector, y_train)

# 预测
pred = mnb_clf.predict(test_vector)
print(pred)
```

4. 评分。采用 Micro-F1 进行打分，其中 precision 和 recall 分别为总体的精确率和召回率。

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

用 metrics.classification_report(y_test, pred)命令返回详细的打分结果如下：

document	precision	recall	f1-score	support
0	0.88	0.90	0.89	97
1	0.82	0.85	0.83	114
2	0.90	0.87	0.88	112
3	0.82	0.84	0.83	127
4	0.87	0.93	0.90	112
5	0.94	0.93	0.93	115
6	0.88	0.79	0.83	124
7	0.91	0.89	0.90	108
8	0.96	0.96	0.96	99
9	0.95	0.97	0.96	113
10	0.96	0.96	0.96	108
11	0.95	0.99	0.97	120
12	0.89	0.87	0.88	119
13	0.97	0.98	0.98	119
14	0.94	0.96	0.95	118
15	0.90	0.95	0.92	128
16	0.92	0.98	0.95	111
17	0.97	0.97	0.97	123
18	0.94	0.89	0.92	110
19	0.86	0.73	0.79	86
avg / total	0.91	0.91	0.91	2263

最终 F1 得分 0.91，说明分类效果较好。

通过这次实验，我对朴素贝叶斯分类器实现步骤及预测结果的评价方法有了更深入的了解。