1.导入数据

In [1]:

```
import pandas as pd
data=pd.read_table('../data/noteData.txt',sep='\t',header=None,nrows = 10000,names=["标签","短信内容"])
data.head()
```

Out[1]:

	标签	短信内容
0	0	商业秘密的秘密性那是维系其商业价值和垄断地位的前提条件之一
1	0	带给我们大常州一场壮观的视觉盛宴
2	0	有原因不明的泌尿系统结石等
3	0	23年从盐城拉回来的麻麻的嫁妆
4	0	感到自减肥、跳减肥健美操、

In [2]:

```
data.shape
```

Out[2]:

```
(10000, 2)
```

2.进行分词

In [3]:

```
import jieba
data['分词后数据']=data["短信内容"].apply(lambda x:' '.join(jieba.cut(x)))
data.head()
```

```
Building prefix dict from the default dictionary ...
Dumping model to file cache C:\Users\TIAN-J~1\AppData\Local\Temp\jieba.cache
Loading model cost 0.693 seconds.
Prefix dict has been built successfully.
```

Out[3]:

	标签	短信内容	分词后数据
0	0	商业秘密的秘密性那是维系其商业价值和垄断地位的前提条件之一	商业秘密 的 秘密性 那 是 维系 其 商业价值 和 垄断 地位 的 前提条件 之一
1	0	带给我们大常州一场壮观的视觉盛宴	带给 我们 大 常州 一场 壮观 的 视觉 盛宴
2	0	有原因不明的泌尿系统结石等	有 原因 不明 的 泌尿系统 结石 等
3	0	23年从盐城拉回来的麻麻的嫁妆	23 年 从 盐城 拉回来 的 麻麻 的 嫁妆
4	0	感到自减肥、跳减肥健美操、	感到 自 减肥 、 跳 减肥 健美操 、

3.导入停用词

In [4]:

```
f = open('../data/my_stop_words.txt','r')
my_stop_words_data = f.readlines()
f.close()
my_stop_words_list=[]
for each in my_stop_words_data:
    my_stop_words_list.append(each.strip('\n'))
```

4.提取特征和目标数据

In [6]:

```
X = data['分词后数据']
y = data['标签']
print(X)
print(y)
```

```
0                商业秘密 的 秘密性 那 是 维系 其 商业价值 和 垄断 地位 的 前提条件
之一
1                                  带给 我们 大 常州 一场 壮观 的 视觉 盛宴
2                                       有 原因 不明 的 泌尿系统 结石 等
3                                   23 年 从 盐城 拉回来 的 麻麻 的 嫁妆
4                                      感到 自 减肥 、 跳 减肥 健美操 、
                               ...
9995                                    明天 微软 Windows10 发布 啦
9996                             我 分享 了 百度 云里 的 文件 ： ？听力 xxxx
9997    痛是 真的 痛 就是 觉得 彻底 彻底 的 痛 没 了解 生活 的 真相 而痛 失去 了
重...
9998                                    微软 为了 重新 争夺 移动 市场份额
9999                             第一 时间 拿 起 手机 点开 微信 领取 红包 一气呵成
Name: 分词后数据, Length: 10000, dtype: object
0       0
1       0
2       0
3       0
4       0
       ..
9995    0
9996    0
9997    0
9998    0
9999    0
Name: 标签, Length: 10000, dtype: int64
```

5.模型训练and预测打分

In [7]:

```
from sklearn.model_selection import StratifiedKFold
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.pipeline import Pipeline

skf = StratifiedKFold(n_splits=10, random_state=1, shuffle=True)

for train_index, test_index in skf.split(X, y):

    X_train, X_test = X[train_index], X[test_index]
    y_train, y_test = y[train_index], y[test_index]

    pipeline = Pipeline([
                        ('vect', TfidfVectorizer(stop_words=my_stop_words_list)),
                        ('clf', MultinomialNB(alpha=1.0))])

    pipeline.fit(X_train,y_train)

    #进行预测
    predict = pipeline.predict(X_test)
    score = pipeline.score(X_test,y_test)
    print(score)
```

```
D:\Anaconda3\envs\pytorch\lib\site-packages\sklearn\feature_extraction\text.py:39
1: UserWarning: Your stop_words may be inconsistent with your preprocessing. Token
izing the stop words generated tokens ['若果'] not in stop_words.
  'stop_words.' % sorted(inconsistent))
```

```
0.946
0.952
```

```
D:\Anaconda3\envs\pytorch\lib\site-packages\sklearn\feature_extraction\text.py:39
1: UserWarning: Your stop_words may be inconsistent with your preprocessing. Token
izing the stop words generated tokens ['若果'] not in stop_words.
  'stop_words.' % sorted(inconsistent))
D:\Anaconda3\envs\pytorch\lib\site-packages\sklearn\feature_extraction\text.py:39
1: UserWarning: Your stop_words may be inconsistent with your preprocessing. Token
izing the stop words generated tokens ['若果'] not in stop_words.
  'stop_words.' % sorted(inconsistent))
```

```
0.955
0.945
```

```
D:\Anaconda3\envs\pytorch\lib\site-packages\sklearn\feature_extraction\text.py:39
1: UserWarning: Your stop_words may be inconsistent with your preprocessing. Token
izing the stop words generated tokens ['若果'] not in stop_words.
  'stop_words.' % sorted(inconsistent))
D:\Anaconda3\envs\pytorch\lib\site-packages\sklearn\feature_extraction\text.py:39
1: UserWarning: Your stop_words may be inconsistent with your preprocessing. Token
izing the stop words generated tokens ['若果'] not in stop_words.
  'stop_words.' % sorted(inconsistent))
```

```
0.949
0.956
```

```
D:\Anaconda3\envs\pytorch\lib\site-packages\sklearn\feature_extraction\text.py:39
1: UserWarning: Your stop_words may be inconsistent with your preprocessing. Token
izing the stop words generated tokens ['若果'] not in stop_words.
  'stop_words.' % sorted(inconsistent))
D:\Anaconda3\envs\pytorch\lib\site-packages\sklearn\feature_extraction\text.py:39
1: UserWarning: Your stop_words may be inconsistent with your preprocessing. Token
izing the stop words generated tokens ['若果'] not in stop_words.
  'stop_words.' % sorted(inconsistent))
```

```
0.954
0.942
```

```
D:\Anaconda3\envs\pytorch\lib\site-packages\sklearn\feature_extraction\text.py:39
1: UserWarning: Your stop_words may be inconsistent with your preprocessing. Token
izing the stop words generated tokens ['若果'] not in stop_words.
  'stop_words.' % sorted(inconsistent))
D:\Anaconda3\envs\pytorch\lib\site-packages\sklearn\feature_extraction\text.py:39
1: UserWarning: Your stop_words may be inconsistent with your preprocessing. Token
izing the stop words generated tokens ['若果'] not in stop_words.
  'stop_words.' % sorted(inconsistent))
```

```
0.944
0.948
```

```
D:\Anaconda3\envs\pytorch\lib\site-packages\sklearn\feature_extraction\text.py:39
1: UserWarning: Your stop_words may be inconsistent with your preprocessing. Token
izing the stop words generated tokens ['若果'] not in stop_words.
  'stop_words.' % sorted(inconsistent))
```

In [8]:

```
data["数据类型"] = pipeline.predict(X) #lambda x:x+1 if not 2==1 else 0
data['数据类型']=data["数据类型"].apply(lambda x:"垃圾短信" if x==1 else "正常短信")
data.head()
```

Out[8]:

	标签	短信内容	分词后数据	数据类型
0	0	商业秘密的秘密性那是维系其商业价值和垄断地位的前提条件之一	商业秘密 的 秘密性 那 是 维系 其 商业价值 和 垄断 地位 的 前提条件 之一	正常短信
1	0	带给我们大常州一场壮观的视觉盛宴	带给 我们 大 常州 一场 壮观 的 视觉 盛宴	正常短信
2	0	有原因不明的泌尿系统结石等	有 原因 不明 的 泌尿系统 结石 等	正常短信
3	0	23年从盐城拉回来的麻麻的嫁妆	23 年 从 盐城 拉回来 的 麻麻 的 嫁妆	正常短信
4	0	感到自减肥、跳减肥健美操、	感到 自 减肥 、 跳 减肥 健美操 、	正常短信

In [ ]:

```

```