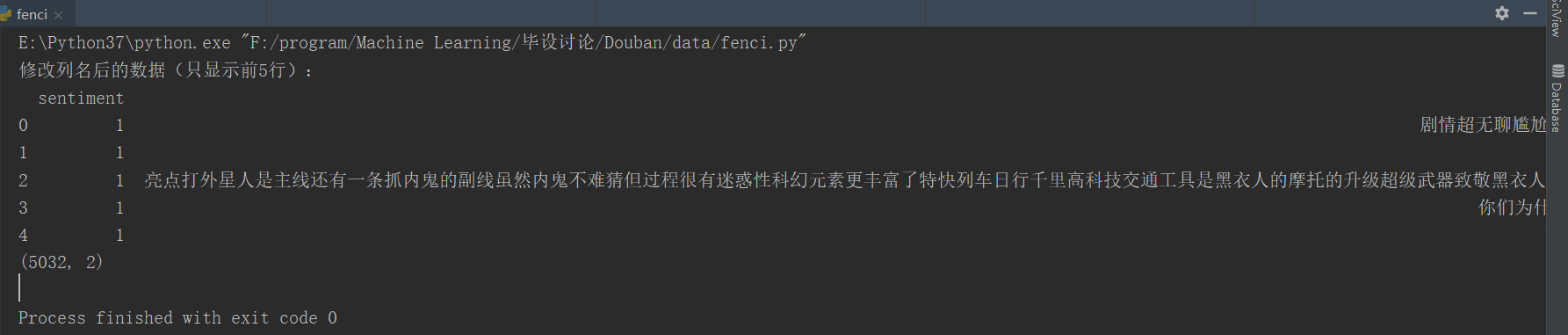
Project 1步骤

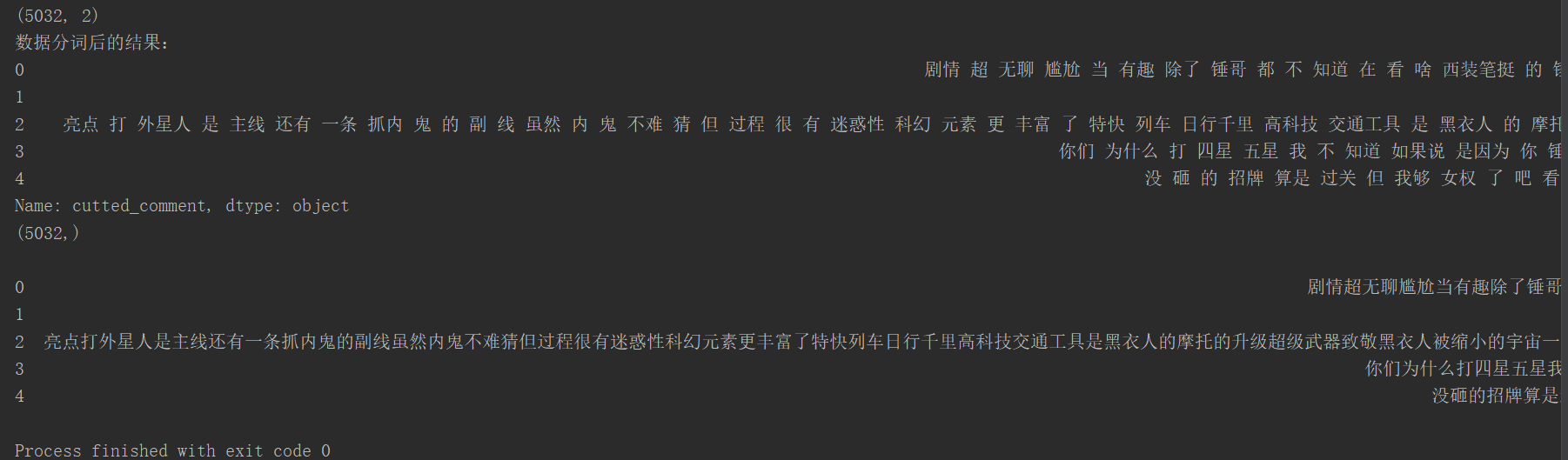
1. 读取评论文件，转换为DataFrame结构，输出如下：



 可以看到评论内容没有标点符号，比较干净。

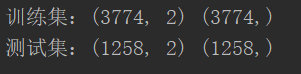
爬下来的数据有5032条评论，共两列，评价 + 文本

2. 用jieba分词, 上下可以对比 分词后的效果



3. 下一步将数据按3:1分为训练集和测试集。

得到 如下训练集和测试集



4. 但是文本数据中许多词属于无意义词，比如：一些、一个、如果、非常等等，这些称之为停用词，需要将这些词去除，下面引入停用词表 stopwords.txt

5.去掉停用词表和不去停用词矩阵的特征数量的变化：



特征数降了近800，有点效果，但是还欠缺。

6. 进一步改进

比如一个特征词如果在80%以上的评论中都出现了，那么这个词就没有很好的区分度（过于平凡），如果出现的过于稀少同样不合适（过于独特），所以我们把特征词出现的频率进行一些限制。

可以看出，直接由一万八千降到四千，有较大改善。



7. 使用贝叶斯预测分类

得到准确率



后续待改进：

1. 爬取得数据集还不够“干净”，比如有的评价里面都是不满意的词，但评星却很高。

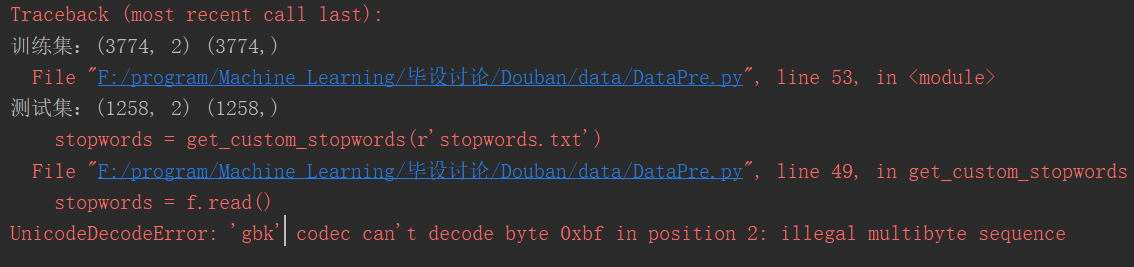
2、否定词的影响。比如“不是很好”分词成了“不是”，“很好”，出现了正向的词。

3、朴素贝叶斯的假设是特征词独立，而这在情感分析中一般不成立。

遇到的问题和技巧：

1. 下载下来的停用词表不是utf-8 格式的，为了简洁数据预处理的代码，另存为停用词表的过程中就可以转换

2. 打开停用词表时报错



参考<https://blog.csdn.net/qq_16761099/article/details/80559302>

可以知道，要处理的字符串本身不是gbk编码，但是却以gbk编码去解码 。比如，字符串本身是utf-8的，但是却用gbk去解码utf-8的字符串，所以结果出错。

只需在open 中加上一句以utf-8 形式打开即可



3. 利用贝叶斯预测分类时，运行报错

函数 train\_test\_split，ModuleNotFoundError: No module named 'sklearn.cross\_validation'

经过查询，发现

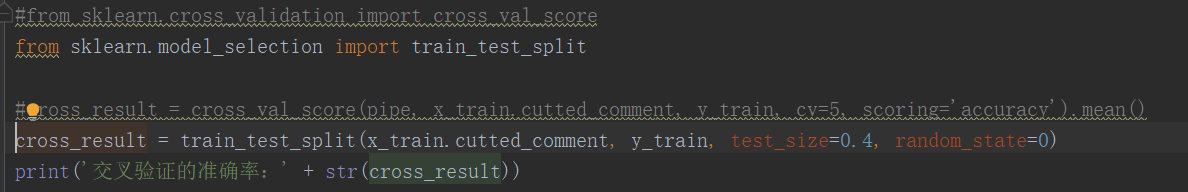
在sklearn 0.18及以上的版本中，出现了sklearn.cross\_validation无法导入的情况，原因是新版本中此包被废弃。

方法一：

改换成train\_test\_split，但是里面的参数不一样，经过查询博客

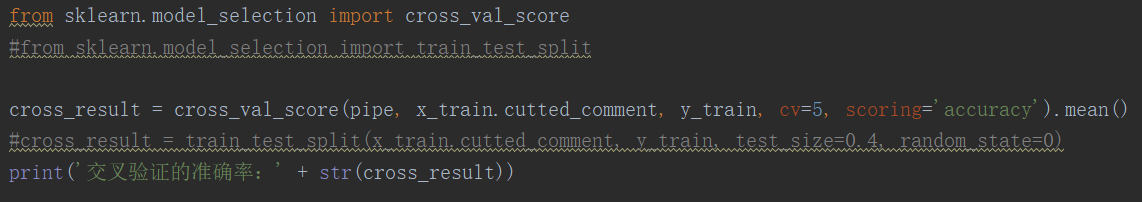
<https://www.jianshu.com/p/6beffa4b80be>

了解到train\_test\_split 的各个参数的用法后，替换成功。



方法二：博客<https://blog.csdn.net/qq_40475568/article/details/83277995>

只需将 cross\_validation 改为 model\_selection 即可，如下：



上述结果都是



参考文献：

《Keras入门与实践》