

****

软 件 学 院

数据挖掘大作业报告

**题　　目 《红楼梦》后40回作者探究**

**小组成员 覃玲芸 30220152200779**

**武戈 24320152202832**

**2018 年 6 月 7日**

目录

[1 选题说明 3](#_Toc516598898)

[1.1 题目来源 3](#_Toc516598899)

[1.2 数据来源 3](#_Toc516598900)

[1.3 整体思路 3](#_Toc516598901)

[2 具体步骤 4](#_Toc516598902)

[2.1 数据预处理 4](#_Toc516598903)

[2.1.1 数据来源 4](#_Toc516598904)

[2.1.2 数据清洗 4](#_Toc516598905)

[2.1.3 数据分割 4](#_Toc516598906)

[2.2 对虚词进行的数据分析 5](#_Toc516598907)

[2.2.1 主成分分析 5](#_Toc516598908)

[2.2.2 分类器分类 6](#_Toc516598909)

[2.3 对高频词汇进行的数据分析 9](#_Toc516598910)

[2.3.1 主成分分析 10](#_Toc516598911)

[2.3.2 分类器分类 10](#_Toc516598912)

[2.3.3 “笑道”剔除后的研究 12](#_Toc516598913)

[3 总结 14](#_Toc516598914)

# **选题说明**

## 题目来源

知乎上的一篇贴子，用机器学习的角度分析红楼梦，躺在我的收藏夹里很久了，知道这次数据挖掘的大作业我决定是时候试一试了。

http://daily.zhihu.com/story/9631416

## 数据来源

众所周知，《红楼梦》其书版本众多，不同版本保留的完整度也不尽相同，我选择当下绝大多数读者能够阅读到的程高本，即曹雪芹所著庚辰本前80回和高鹗所著后四十回的合订本。

## 整体思路

我们打算从高频词汇和虚词两个角度出发来进行文本研究。

选择虚词是因为作家对虚词的使用习惯是较持久的，不会受到情节的影响，比如“吗”、“呀”、“呢”、“咧”这些词，同是虚词中的语气词，不同作家会选择不同的语气词对同一语境进行描写。

但是需要注意的是，高鹗是在曹雪芹创作的基础上进行续写的，他认真阅读过《红楼梦》的前80回文本，他熟悉了曹雪芹的用语习惯，也许会有意模仿。因此，除了虚词，我们选择了高频词汇作为另一个研究方向，正所谓“画皮画虎难画骨”，我们试图从高频词汇分析中发现高鹗模仿不出的、曹雪芹的独到之处。

在对高频词汇分析的过程中，我们注意到情节发展的影响，比如本人的主人公，“宝玉”这个词在全120回出现了3945次， “宝二爷”189次，但是随着情节的发展，如80回后贾府破败，“宝二爷”出现的次数大大减少，又如“黛玉”出现1095次，但是在黛玉魂归之后就很少再有“黛玉”一词出现。因此，我们从数据中提出了所有人物的名字。

# 具体步骤

## 数据预处理

### 数据来源

《红楼梦》程高本的120回文本。

### 数据清洗

剔除批注：这些批注系与作者关系较密切的人撰写，与小说情节相关，会对数据统计造成干扰，故需要剔除。

删除特殊符号：《红楼梦》中繁体字无法显示，被转化成特殊字符，对文本的分割分词造成困扰。

### 数据分割

1、把《红楼梦》全部文本分割成120个单独的txt文本文件。

2、把每个文本文档里面的全部标点符号替换成“#”

3、用算法分割每一个txt文件中的关键词，生成csv文件。

算法参照github： https://github.com/LouYu2015/analysis\_on\_the\_story\_of\_a\_stone

4、人为剔除分词csv中的人名，因为人名可能与情节发展相关，剔除人名可以除去情节发展的干扰。

虚词：古代文言文常用虚词17个现代常用虚词27个，我们分别统计了他们在前80回、后40回以及总120回的分布，发现两类虚词交错分布，因此我们有44个虚词用来统计。我们把每个虚词在每一章节出现的频率画图。如下：

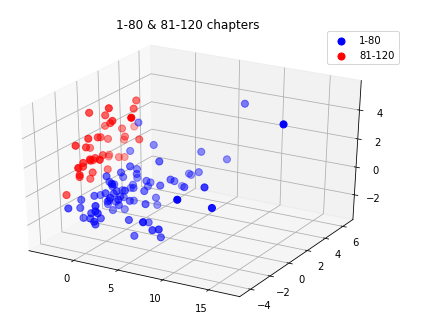
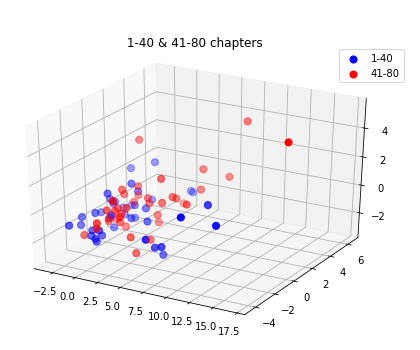
我们发现有几个比较特殊的词语。“皆”、“亦”两字在前80回中使用频繁，而到了81回以后使用次数直线下降，与前面的使用频率形成鲜明的对比。“吗”“咧”两字在前面使用很少，几乎没有出现，但是以81回为分界线，81-120回使用十分频繁，差别明显，这体现了前80回和后40回用词习惯上的明显的不同，前后出自两个人之后的可能性更大。

如上图，“亦”、“皆”在前80回中的出现频率明显高于后40回

如上图，“吗”、“咧”在后40回出现的频率明显高于前80回

## 对虚词进行的数据分析

### 主成分分析

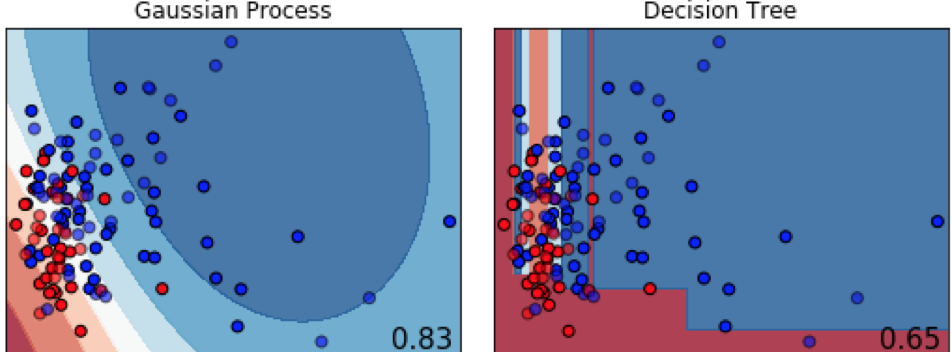
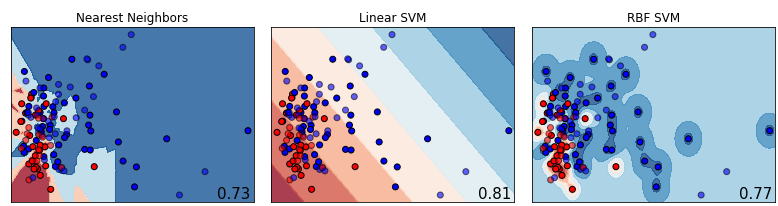
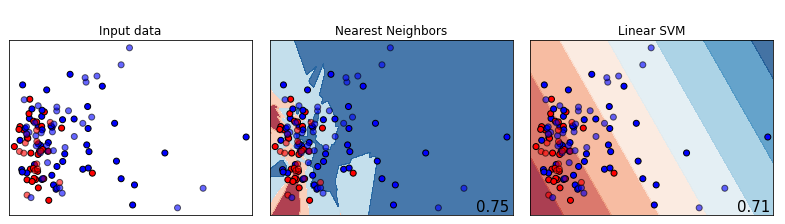
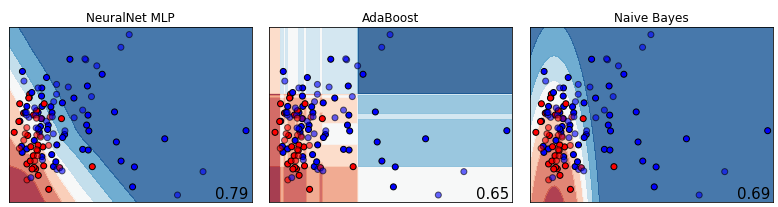
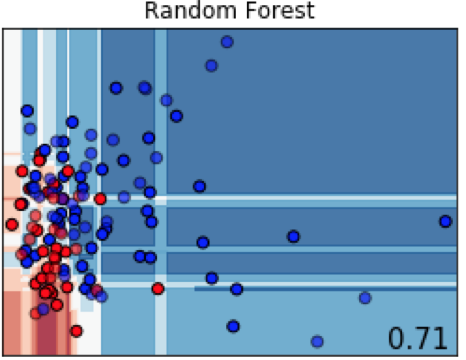


首先对数据进行降维处理，把44行（44个虚词）120列（120个章回）的csv文件导入，降到三维，用蓝色代表前80回，红色代表后40回，观察分布，可以看到两种颜色的点分布在不同的空间区域，总体来说蓝色的点分布在左下方，红色的点分布在蓝色的点上方的区域，两种点之间的界限较为明显，偏离总体数据集群的点较少。（如上右图）

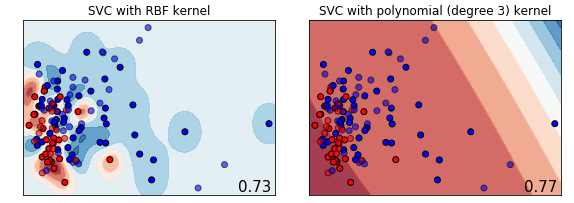
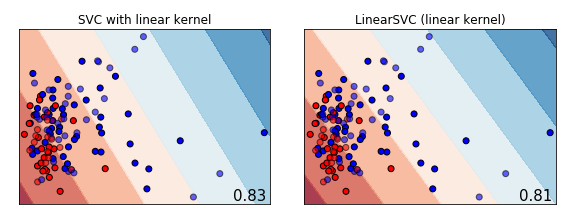
作为对比，我们对前80回进行了相同的处理，把前80回分割成1-40回和41-80回两部分，把44行（44个虚词）120列（120个章回）的csv文件导入，只读取前80行，降到三维，用蓝色代表1-40回，红色代表41-120回，观察分布，可以看到两种颜色的点分布较为密集，而且中间没有明显界限。（如上左图）

### 分类器分类

#### 1-80回和81回-120回

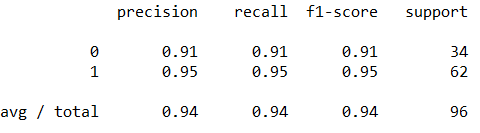
  

我们对1-80回和81-120回的虚词进行了归一化处理、降维，用分类器进行分类。我们从1-80和81-120回中随机选取60%的样本进行训练，对剩余40%的样本进行预测。结果如图。

可以看到，使用线性内核的SVM算法达到了较高的准确率，在多类分类(文本分析)效果上较好，因此采用kernal=’linear’的SVM进行分类预测。

比较四个SVM分类器的分类结果，使用线性内核的SVC结果更好，因此使用该分类器调整参数得到预测的结果如下：

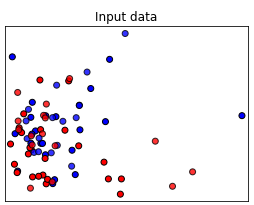
分类指标：precision类标签精确率 recall 召回率f1-score F1值 support样本数量



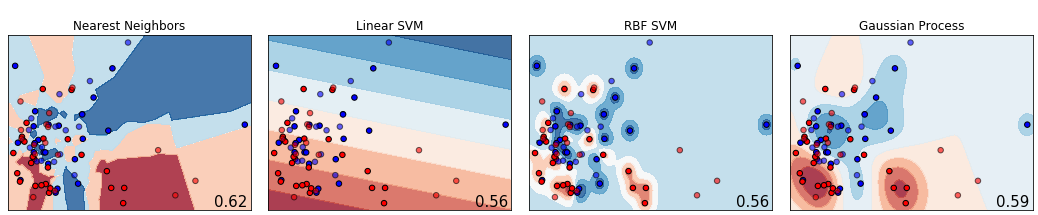
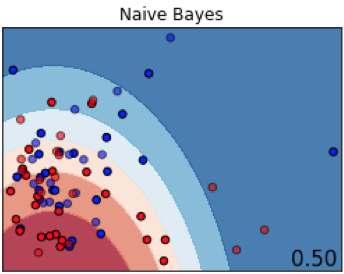
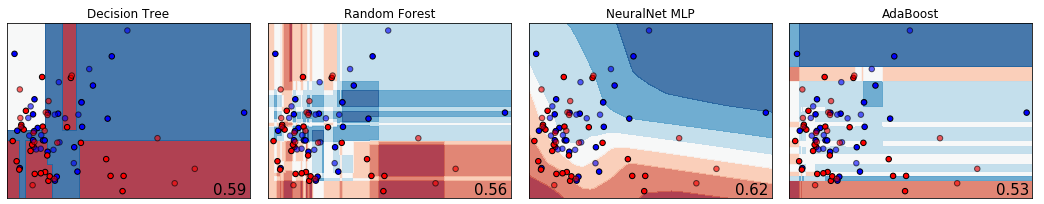
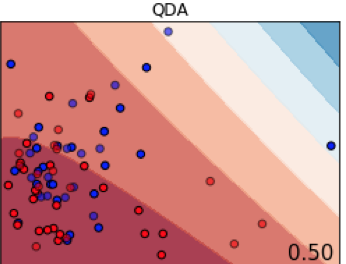
标记为0的是后40回，标记为1的是前80回。可以看到预测结果有较高的正确率。

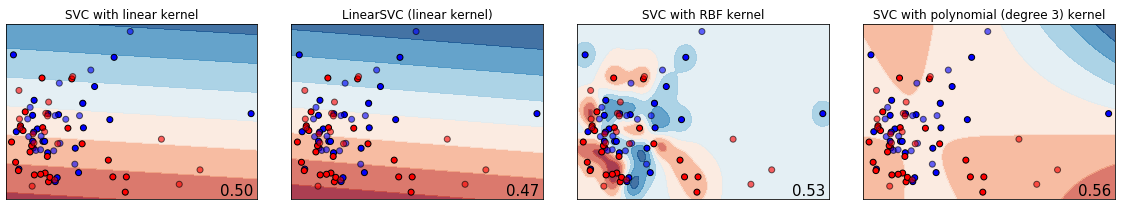
#### 1-40回和41回-80回

根据老师的建议，我们拿前八十回分成两个四十回来做个比较验证，前八十回分成一半，前40回和后40回分布如图，其中红色点代表前40回，蓝色的点代表后40回，查看两拨数据的相似程度。使用分类器的分类效果为：



他们几乎相互重合，联系密切，反映了是同一个作家的写作风格格所致。

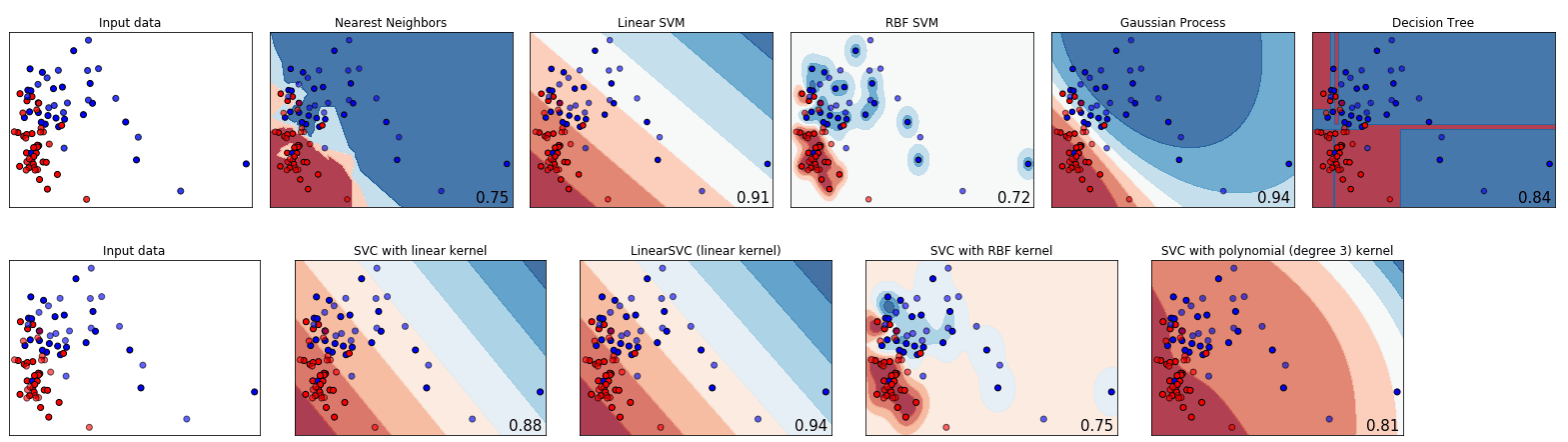
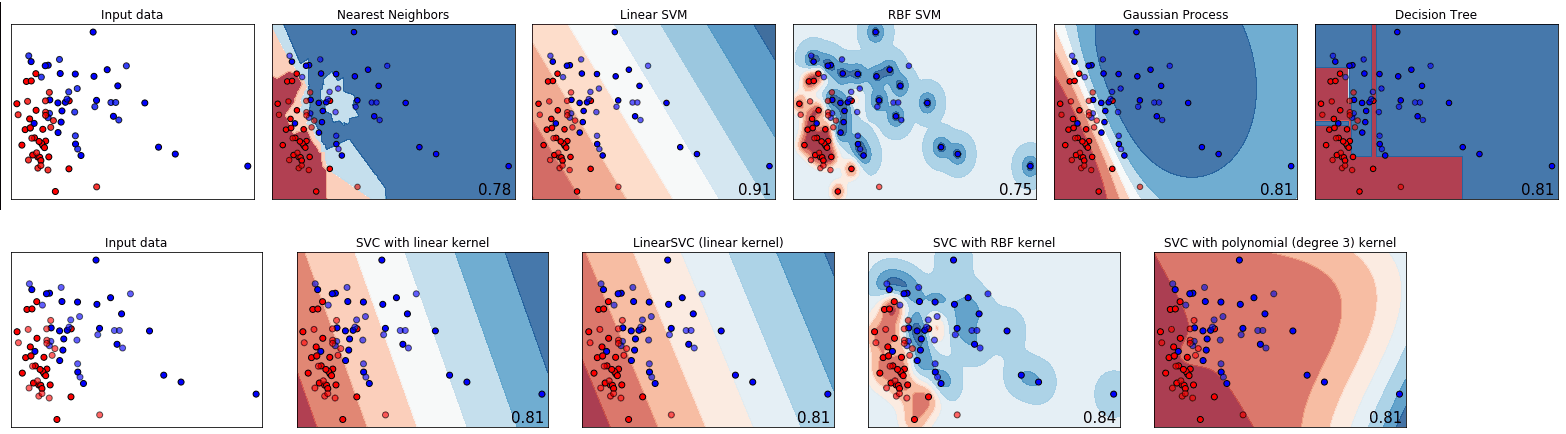
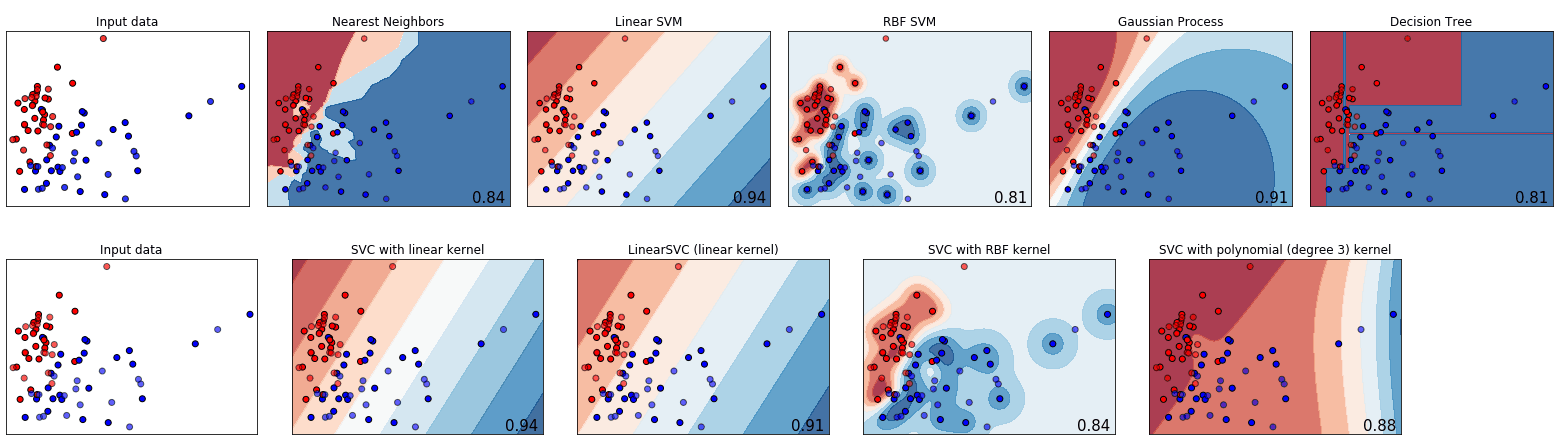
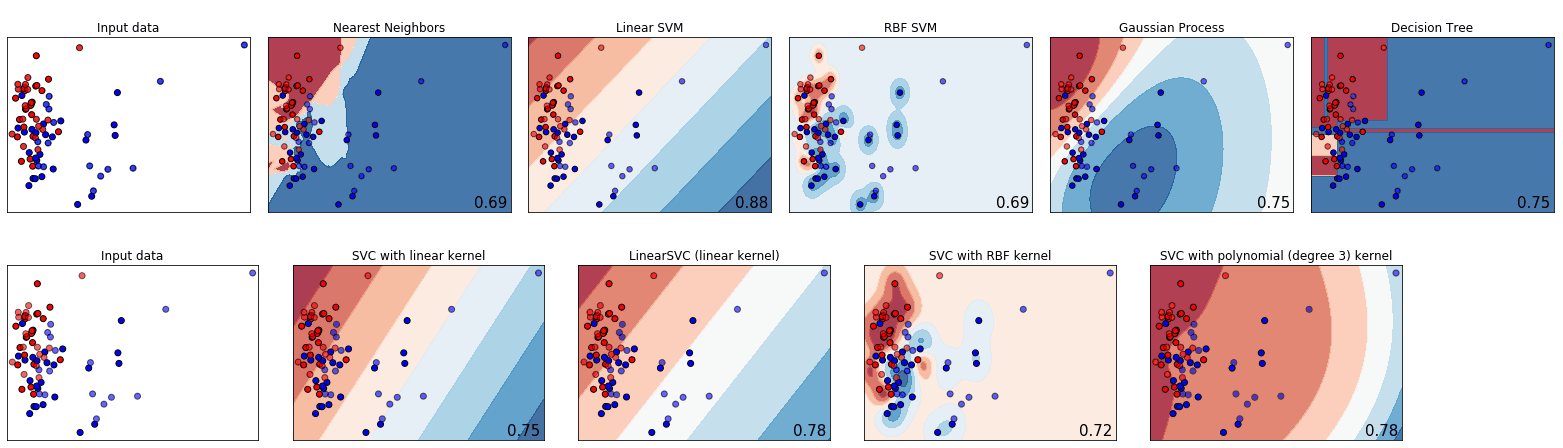


分类器的预测准确率很低，只有近百分之五十，说明前两个类的分类特征不明显，没有足够的证据支持前八十回分为两类。从实际情况可知，前80回确实只出自曹雪芹一人之手。

#### 前80回随机取40回与后40回

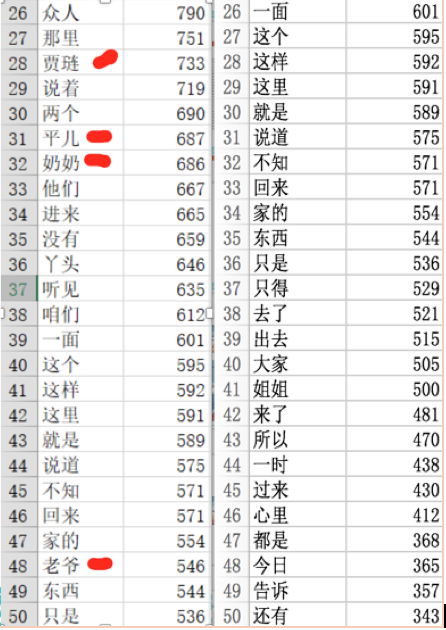
我们进行了四次随机抽取，从前八十回随机取40回作为第一类，80-120回作为第二类，考察两个类别的距离。

①②③④



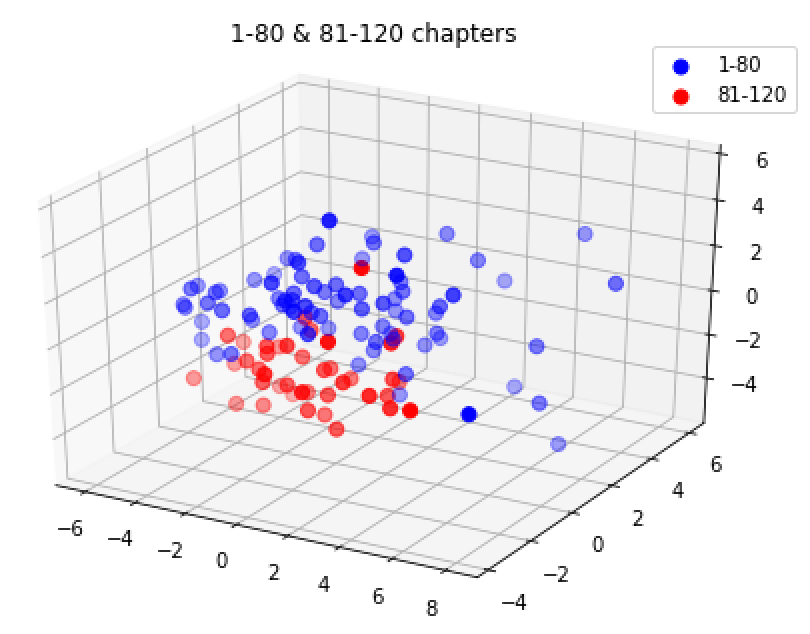
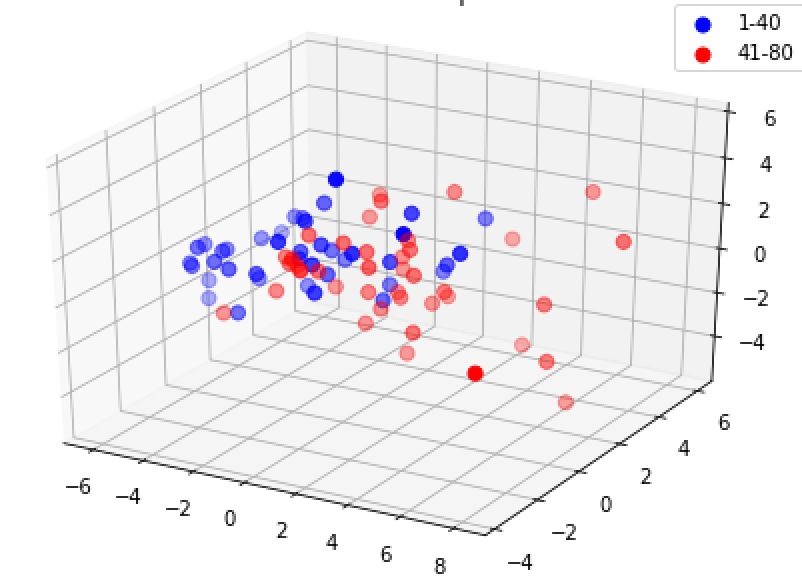
从预测准确性来看，最低的有71%，高的有91%，平均准确性在0.85左右，相对于前80回是两个人的分类准确性为0.5来说更高，也就是有更充分的可能性支持后四十回和前80回不出自同一人。

## 对高频词汇进行的数据分析



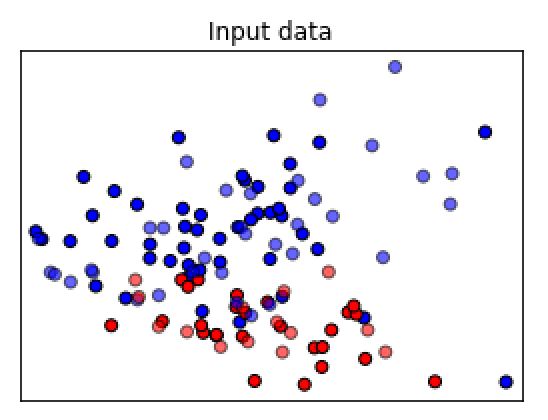
上图为使用分词得到的高频词汇的频率，其中标红的代表与人名相关的数据，在后续实验中剔除。

### 主成分分析

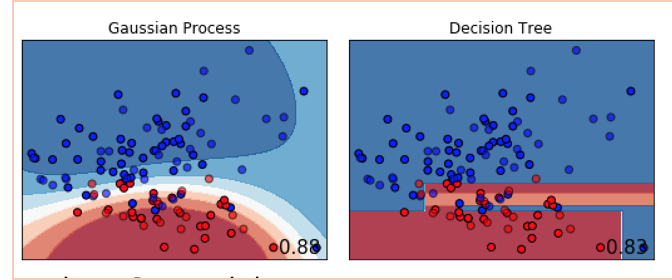
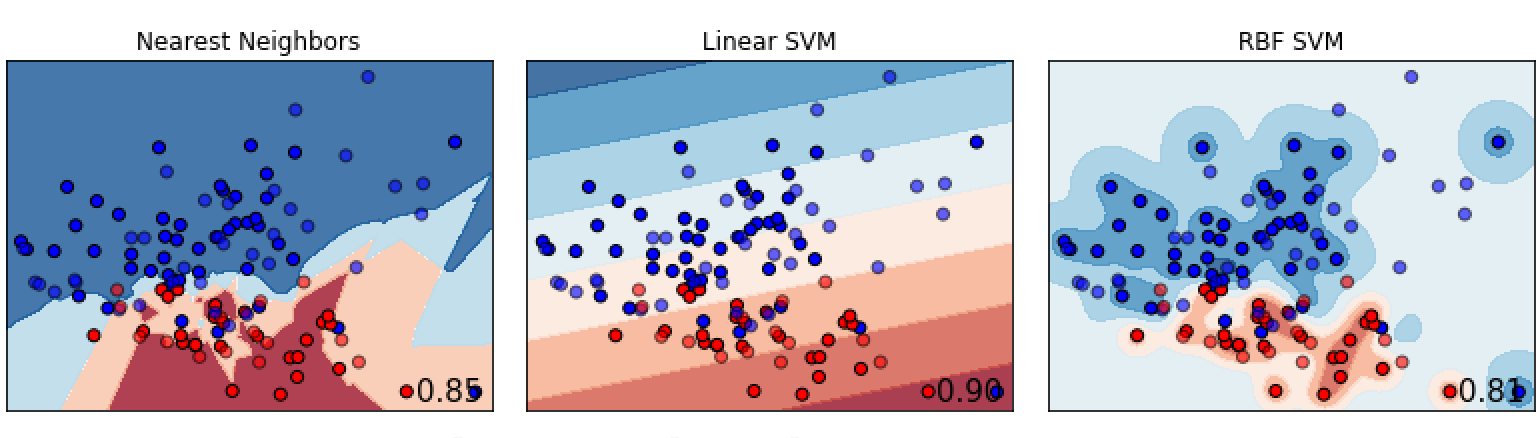
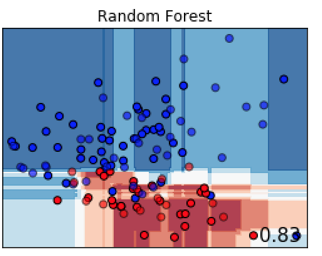
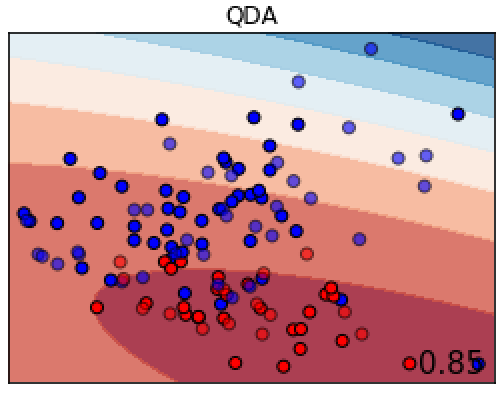
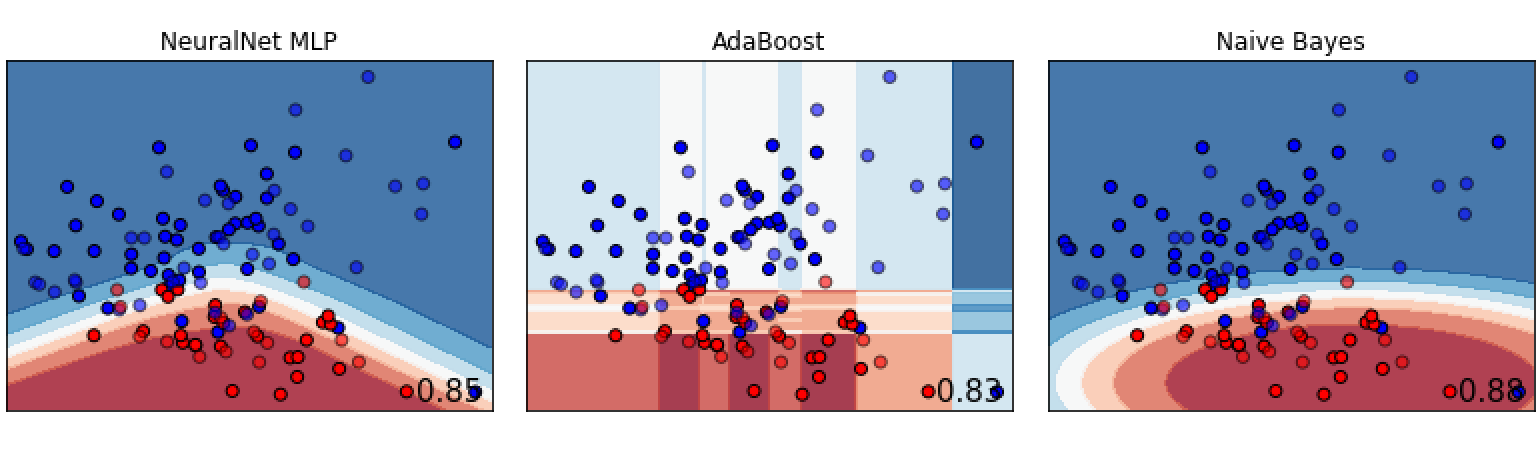


首先采用主成分分析法对数据进行归一化、降维处理，可以看到左图中为前80回分两类，1-40回何41-80回的结果，红点和蓝点分布较为集中。右图为1-80回随机抽取40回与81-120回的结果，红点和蓝点各自集簇分布，中间有较为明显的分层。

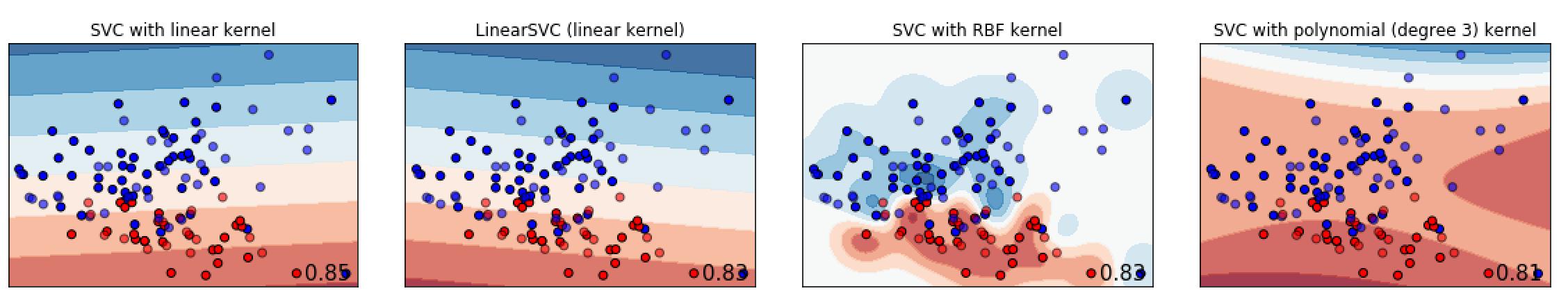
### 分类器分类



蓝色点为1-80回数据，红色点为81-120回数据，可以看出红色点和蓝色之间有较为明显的分界。

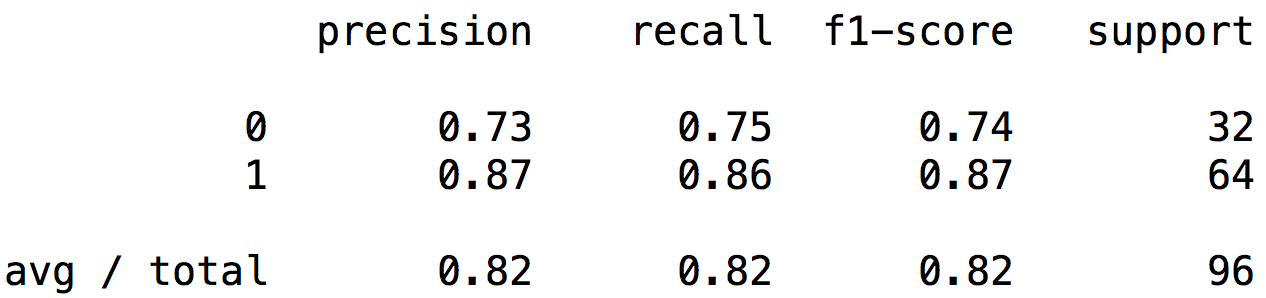
 

可以看到，使用线性内核的SVM算法达到了较高的准确率，在多类分类(文本分析)效果上较好，因此采用kernal=’linear’的SVM进行分类预测。



比较四个SVM分类器的分类结果，使用线性内核的SVC结果更好，因此使用该分类器调整参数得到预测的结果如下：

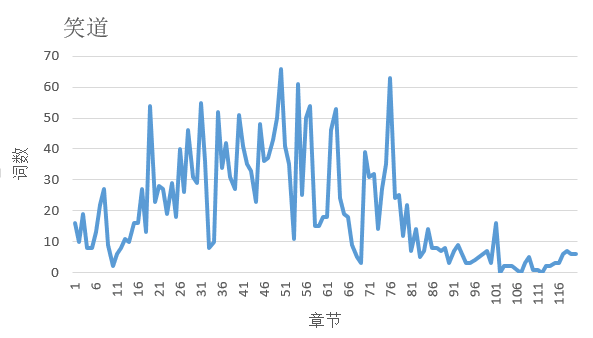
分类指标：precision类标签精确率 recall 召回率f1-score F1值 support样本数量



其中，标记为1的是1-80回，标记为0的是81-120回。 可以看到，通过这种分类方法可以得到较高的准确率。

### “笑道”剔除后的研究

#### 为什么剔除“笑道”

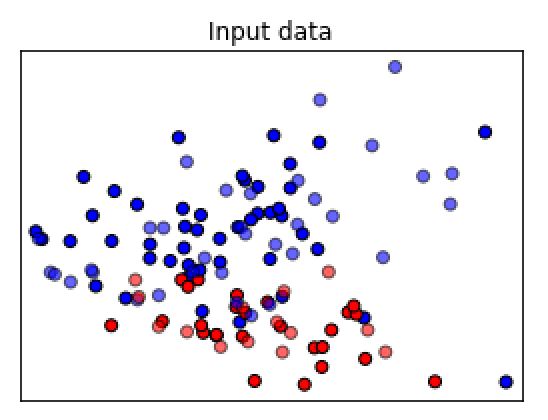
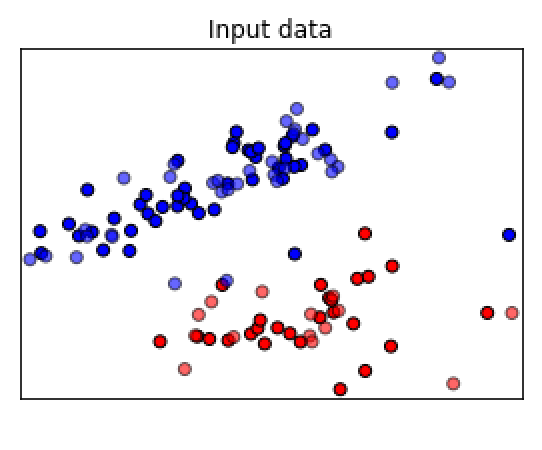


由上图笑道在全书出现的频率来看，总体分布近似先升再降，再加上“笑道”一词的频率最高，我们猜想改词可能与情节发展有关，如果真的与情节发展有关，则需要剔除。

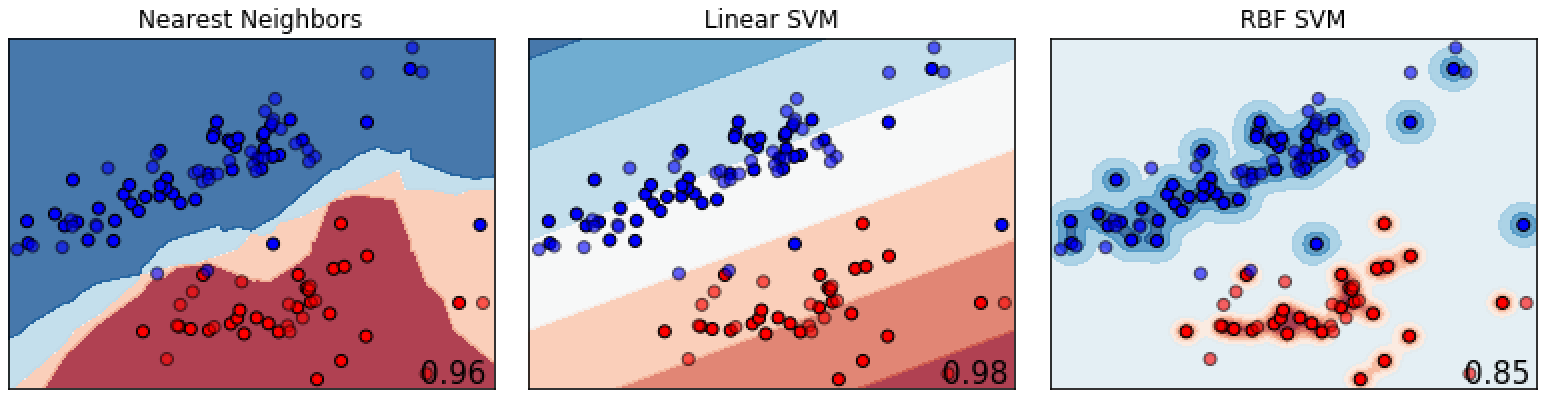
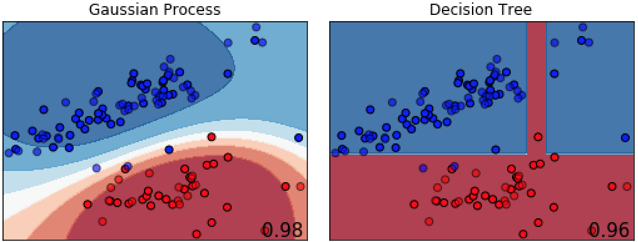


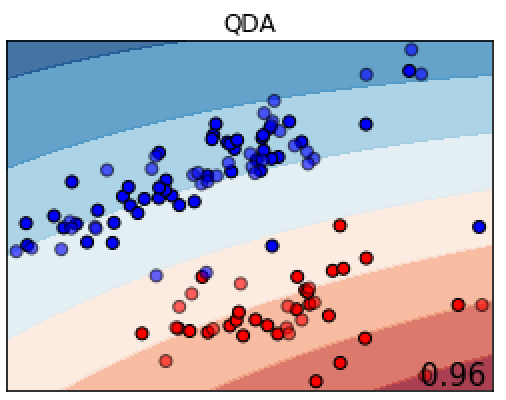
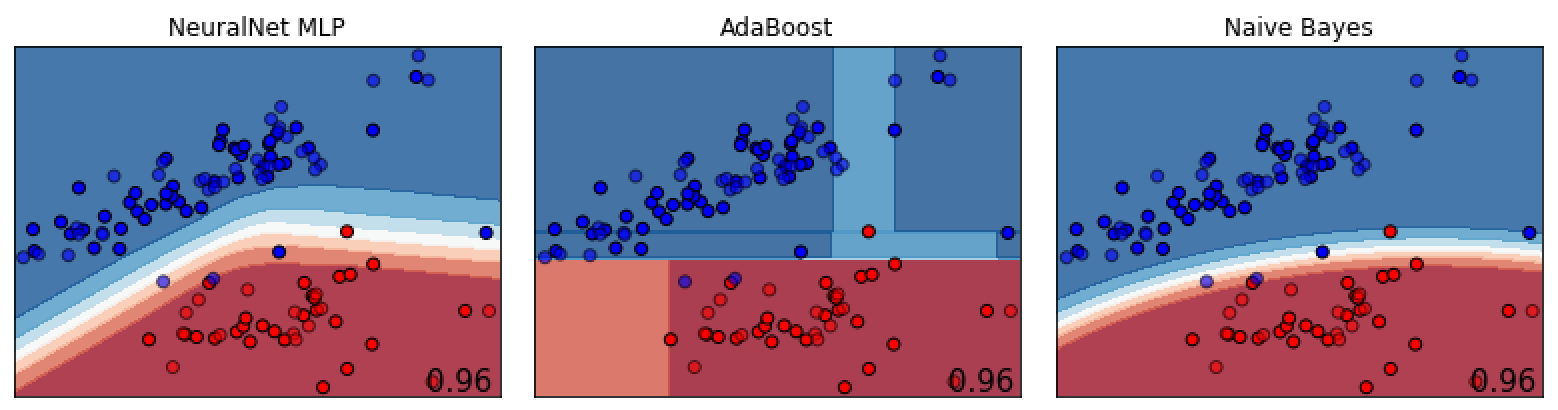
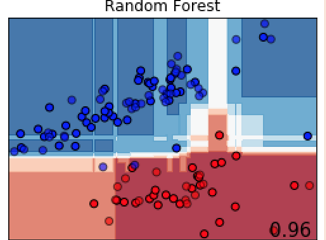
我们观察高频词汇表（左图）发现，剔除人物名字后，“笑道”这个词出现的频率最高，而且可能与情节发展有关。于是我对出现“笑道”一词的频率和对应章节名称进行了排序，得到右图。我们发现这一词出现较高频率的回数总是伴随着喜事的出现。于是我们想剔除“笑道”这个词，重新进行数据的分类和预测。

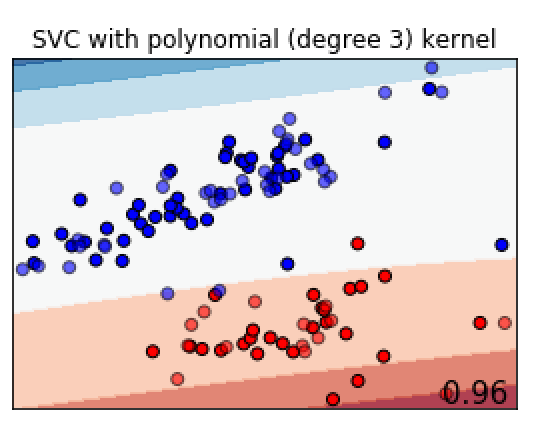
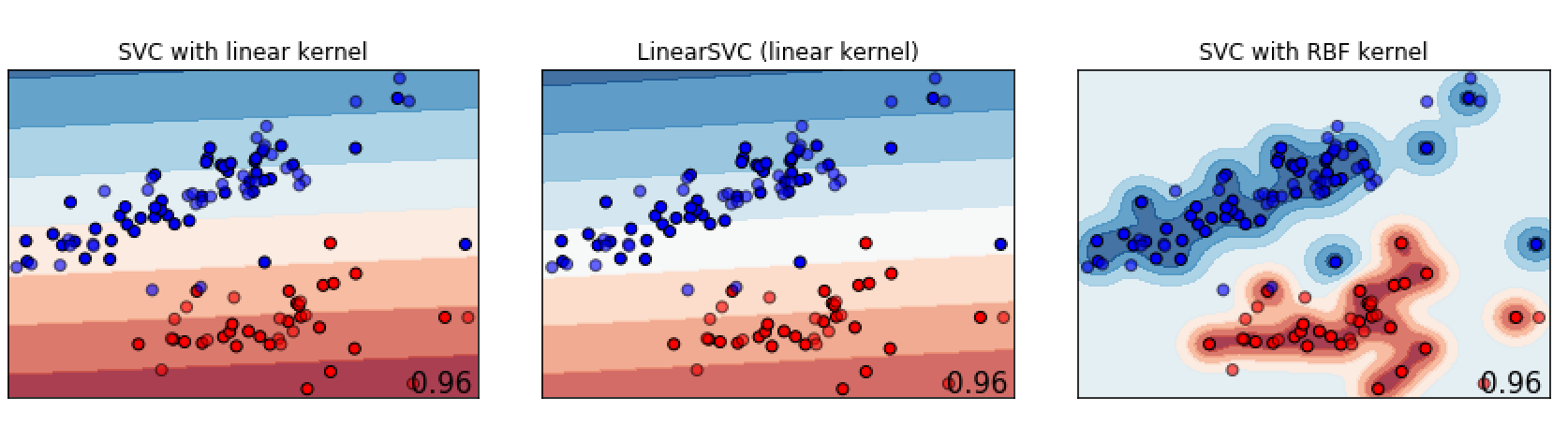
剔除“笑道”一词后，我们得到如下输入数据（下左图），对比剔除之前（下右图）反而更加泾渭分明。



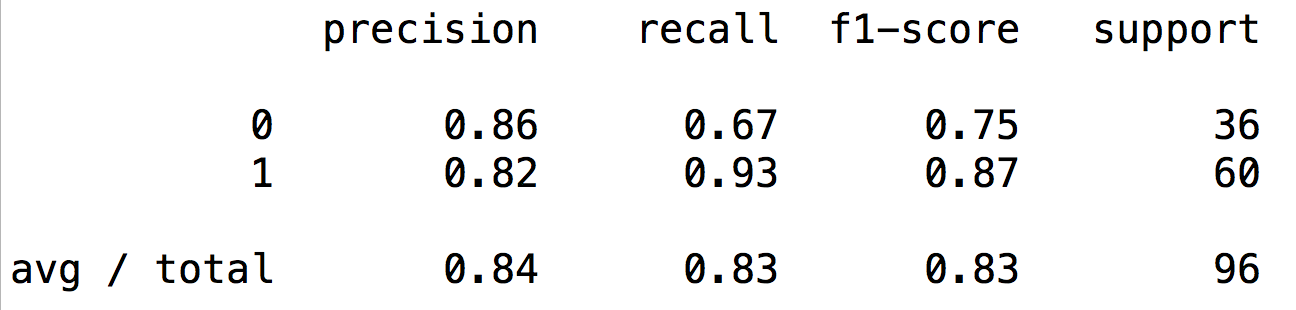
我们对剔除后的数据进行了分类训练和预测，得到的结果如下：

可以看到，使用线性内核的SVM算法达到了较高的准确率，在多类分类(文本分析)效果上较好，因此采用kernal=’linear’的SVM进行分类预测。

比较四个SVM分类器的分类结果，使用线性内核的SVC结果更好，因此使用该分类器调整参数得到预测的结果如下：



此时的结果较剔除“笑道”之前有微弱的提升（比之前高约0.1）。

# 总结

我们从作者写作习惯的角度，利用两个方面的数据：虚词和高频词汇，对现流行的120回程高本《红楼梦》进行了数据处理和分析。我们以为虚词的效果会比高频词汇更加理想，而事实恰恰相反。我们猜测这是由于高鹗在续写《红楼梦》过程中有意模仿曹雪芹所致。好在我们还有另一组数据，即高频词汇。高频词汇数据的处理比虚词复杂得多。为了得到准确的分析结果，我们进行了许多努力，比如数据的清洗，分词算法的优化（因为《红楼梦》属于半文半白的语言，没有现成的合适的词库可以进行分词），以及人名的剔除。“笑道”这一发现纯属意外，但是也为我们带来了意外的惊喜（模型准确率的提升）。

在分析完成后，基本可以得出结论（用虚词正确率在94%，用高频词汇正确率在83%），《红楼梦》的前80回和后40回来自不同的作者。许多文献资料已经证明了这一点，但是用《数据挖掘》课程学到的知识亲自去探索验证的过程如此有趣。