基于特征提取的作家风格识别与分析

孙佳仪 肖桐 郑克寒

1. 绪论

1.1. 研究背景和意义

根据文章的风格识别作者一直以来都是学术界的热门话题之一，且这一研究在实际应用上具有非常重要的意义。传统的基于文献考证的方法较为费时费力，效果也不太理想。而随着机器学习和统计学习的发展，作家风格的识别研究逐渐有了一些机器学习的模型的加入，并且取得了比较好的效果。

在《数据科学导论》的课程大作业中，我们希望能够从较为熟悉的中国近现代文本出发，对作家风格的识别与分析作出一些尝试。先前从这些文本中得到的对于作家风格的认知大多基于感性层面，如鲁迅的凝练简洁、张爱玲的泼辣犀利等，而从数据的角度出发，利用数据科学的相关知识和作家的大量文本去还原写作风格，也不失为一个可行的方案。例如，不同作家的风格往往有着鲜明的差异，若是可以运用一些机器学习方法实现作家分类，并根据分类的结果反过来探讨各种特征对于作家风格识别的重要性，便可能从中得到令人耳目一新的结论。

在本次大作业中，我们重点关注了近现代作家的作品。这些作品与当下人们的写作习惯和风格较为贴合，也因此让我们实现的作家风格分类器 App 有了更大的应用价值。此外，由于这个分类器不仅能够对已知作家的文章进行分类，还能为一些不知名的文章或者模仿的文章找到风格最相似的作家，我们还可以将其进一步运用在鉴别抄袭、鉴别争议署名和识别未署名文章等问题上。譬如，对于知乎问题“如何优雅地模仿鲁迅的文风？”，我们可以将其中的高赞回答通过风格分类器进行分类，根据是否被分类成鲁迅来评估其相似性。

1.2. 相关工作

国外学者很早便开始了关于文章风格和作者推断的研究，但早期的研究比较偏重于使用统计的方法，如对文章中字、词的使用频率进行统计，然后再从中分析规律。Melldenhall 是最早尝试用统计的方法对写作风格进行研究的学者之一。在他的研究中，单词长度是用来区分作者的重要指标。后来又有学者尝试在这一任务上使用了机器学习模型，如 Khmelev 便将隐马尔可夫模型引入到了俄语文学作品的作者识别中。他将文学作品中出现的字母序列作为特征，通过模型分析，取得了较好的效果。

然而，针对中文的作者识别研究发展得较迟，且主要集中于《红楼梦》作者的研究。由于中文与其他语言的区别较大，如何选取模型中使用的特征，以及选用什么模型，都是值得探讨的问题。在过去的研究中，孙晓明采用虚词频率作为特征进行了分析。在对虚词进行特征提取后，即便使用最简单的模型也能够取得很好的识别效果。刘莎提出了一种基于神经网络的双向长短时记忆（BiLSTM）的文章写作风格识别方法，在《红楼梦》识别和近代作家分类上都有着较高的准确率。刘明勇提取了文本中的多组风格特征，并采用隐马尔科夫模型和支持向量机作为分类模型，与旧有模型相比获得了性能上的提升。本次实验在特征提取上主要参考了最后一位学者的研究，同时尝试使用了多种机器学习模型进行风格分类，以期达到良好的效果。

2. 实验方法

2.1. 爬虫

在对不同文学网站进行比较后，我们将星月文学网和可阅文学网定为了目标网站。在爬虫过程中，我们将 rvest 包和 Chrome 的 selectgadget 插件作为主要工具，通过编写 R 语言代码，实现了对某一位作者在该网站上所有文章的爬取。两个文学网站的组织形式并不相

同，但是其爬虫的原理是相通的：首先根据输入的参数进入网站的作者页，再通过作者页获取每一本书的url，进入书本页后再获取各章节的url，从而提取文本存储下来。不过，二者在实现方式上有所区别，如可阅文学网进入书本页后仅能获取首个章节的 url，由于后续章节依次递增，我们使用 url.exists 函数判断它何时终止；星月文学网的中长篇和短篇文章的文本形式不太相同，需要分别进行爬取。文章在网站中都以 html_text 的形式呈现，因此比较容易获取。最终的文本以“书名”或“书名+章节名”为标题保存，存储在以作家名命名的文件夹中。该过程中一共爬取了 17 位作家的文章。

在爬虫结束后，我们进行了初步的数据清洗。首先，去掉了样本量过少的作家，因为我们认为样本量过少不能反映出作者真实的写作风格，会对后续实验造成负面影响；其次，去掉了字符数少于 100 的文章，我们同样认为这些文章不能反映作者的风格。最终留下的有 14 位作家的 1505 篇文章，作家分别为：胡适、老舍、鲁迅、茅盾、莫言、沈从文、汪曾祺、王小波、萧红、余华、郁达夫、张爱玲、赵树理、朱自清。

2.2. 特征工程

结合参考文献，本实验采用七类最能代表作家语言风格的特征：标点使用、平均句长与平均段落长、单双三字词语使用、词汇丰富程度、词性分布程度、词性顺序分布程度、虚词使用。

标点使用中选取了句号、逗号、顿号、分号、问号、感叹号、冒号、破折号这八个标点，统计相应标点在总标点中出现的比率。

在平均句长与平均段落长中，平均句长以句号、逗号、问号、感叹号、冒号、破折号六个标点为分隔符，将每个文章作为字符串进行分割，统计分割后字符串的总字数，取平均值。平均段落长则以空白符为分隔符，之后操作类似。

接下来的特征涉及分词和词性提取，需要文本预处理，因此我们引入了 jiebaR package 辅助工作。JiebaR 共有四种分词模式：支持最大概率法（Maximum Probability），隐式马尔科夫模型（Hidden Markov Model），索引模型（Query Segment），混合模型（Mix Segment），本实验中选择默认的最大概率法分词。

单双三字词语使用统计了每个作家每篇文章中一字词语、两字词语、三字词语的使用次数，并计算它们在该文章所有词语中出现的比率。

词汇的丰富程度由每篇文章所用的不同词语个数/该文章所用的总词语个数体现。

词性分布中，利用 jiebaR package 对文章进行词性提取，选取了名词、动词、形容词、副词、连词、语气词、成语七个词性，统计每个文章中相应词性的词语在总词语中出现的比率。

词性顺序分布中选取了最为常见的形容词+名词，副词+动词，名词+动词（即常见的主谓结构短语），动词+名词（即常见的动宾结构短语）四种，统计每个文章中相应词性顺序短语出现的频率。

虚词使用将每个虚词出现的频率做了一个词频矩阵，并且考虑到搜集的文章长短不一，将词频改为除以文章使用的词语总数的词比率。虚词表参考了《现代汉语虚词例释》，共计 738 个虚词。

至此，我们筛选出了 7 个大特征类，共 763 个特征条目。其中，部分特征没有在任何作家的任何文章中出现，即提取后的数据全为 0。这样的特征条目会稀释模型的拟合优度，因此，我们将此类零特征筛选掉，最终剩下 744 个特征条目。特征总结如表 1 所示。

特征类别	特征条目具体阐释
标点使用	每篇文章中句号、逗号、顿号、分号、问号、感叹号、冒号、破折号在总标点中出

	现比率，共 8 项
平均句长与平均段落长	每篇文章中平均句子字数、平均段落字数，共 2 项
单双三字词语使用	每篇文章中单字词语、双字词语、三字词语出现次数，共 3 项
词汇丰富程度	每篇文章中所用的不同词语个数在总词语中出现比率，共 1 项
词性分布程度	每篇文章中名词、动词、形容词、副词、连词、语气词、成语在总词语中出现比率，共 7 项
词性顺序分布程度	每篇文章中形容词+名词、副词+动词、名词+动词、动词+名词短语出现次数，共 4 项
虚词使用	每篇文章中《现代汉语虚词例释》虚词表词语出现次数，共 738 项

表 1：特征类别及特征条目具体阐释

2.3. 朴素贝叶斯分类器

朴素贝叶斯分类器（Naive Bayes Classifier）是基于贝叶斯定理以及特征条件独立假设的一种经典的文本分类方法。虽然在实际文档中，词项之间往往存在条件依赖关系，且词项在文档中出现的位置对分类的贡献也不同，导致其在概率的估计上存在一定误差，但是在最终的效果上，朴素贝叶斯经常能实现简单有效的分类，因此得到了广泛使用。

在本实验中，我们首先对这一文本分类方法进行了尝试，但是此处大量非正态的连续型变量为概率的计算带来了一定困难，而且模型的初步效果也并不理想。这可能是因为，我们的特征数量较为庞大，且相互之间有明显的相关性。因此，我们没有继续深挖朴素贝叶斯，而是尝试使用了支持向量机、弹性网回归和随机森林这三种模型。

2.4. 支持向量机

支持向量机 (Support Vector Machine，SVM) 是建立在统计学习理论和结构风险最 小化原理基础上的一种机器学习方法。它在解决小样本、非线性和高维模式识别问题中表现出了许多特有的优势，并在很大程度上克服了“维数灾难”和“过学习”等问题。此外，它具有坚实的理论基础和简单明了的数学模型，被广泛应用于文本分类问题中。

在本实验中，样本量相较特征数量并不是很大，且特征数量高达 744 个，较为符合支持向量机的优势应用场景，故在此处对该方法进行尝试。在调参部分，我们使用了 R 中的内置算法，将支持向量机中的惩罚因子设置为 3。

2.5. 弹性网回归

弹性网回归（Elastic-Net Regression）的罚函数恰好为岭回归罚函数和 Lasso 罚函数的一个凸线性组合。当 $\alpha = 0$ 时, 弹性网回归即为岭回归；当 $\alpha = 1$ 时, 弹性网回归即为 Lasso 回归。因此，弹性网回归兼有 Lasso 回归和岭回归的优点，既能达到变量选择的目的，又具有很好的群组效应。此外，它对强共线性的特征也具有更好的特征选择能力，可以有效地处理高维数据。

在实验部分，支持向量机的表现仍然存在一定的不足，可能是由于它对于特征的处理还不够精细。因此，我们进一步尝试了有更好的特征筛选功能的弹性网回归。在调参部分，我们使用了 R 中的 glmnet 包，首先通过循环选出了最佳的 alpha，再通过 cv.glmnet 函数，用五折交叉验证选出了最好的 lambda（图 1）

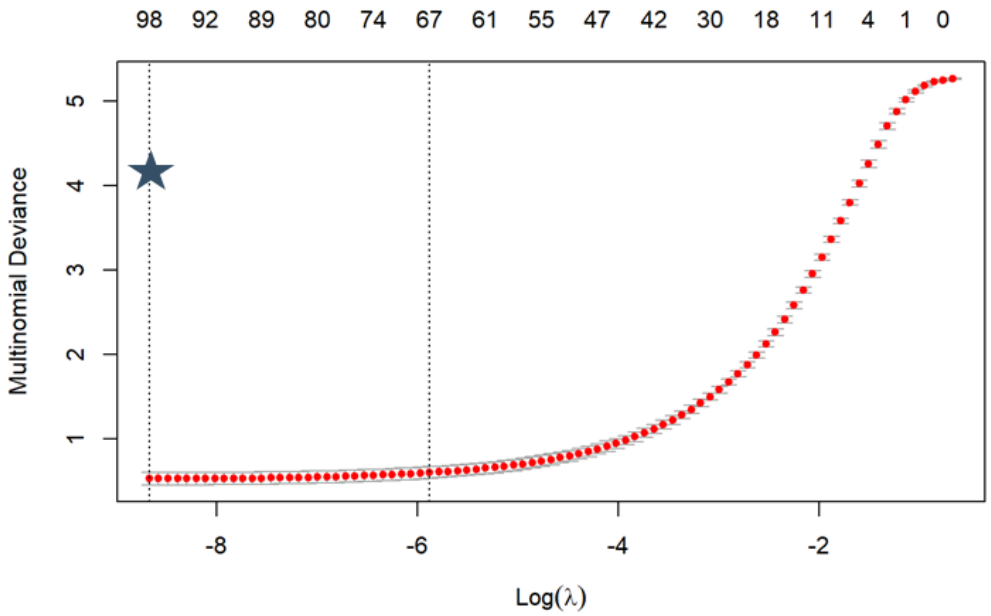

图 1：弹性网回归 lambda 筛选

2.6. 随机森林

随机森林(Random Forest)是一种基于分类树 (Classification Tree) 的算法。它利用 bootstrap 重抽样方法从原始样本中抽取多个样本，对每个样本进行决策树建模，然后组合多棵决策树的预测，通过投票得出最终预测结果。大量的理论和实证研究证明，随机森林具有很高的预测准确率，对异常值和噪声都具有很好的容忍度，且不容易出现过拟合。此外，它对于多元共线性并不敏感，可以很好地预测多达几千个解释变量的作用。因此，我们在此处对其进行尝试，以期获得更好的分类效果。

在调参部分，我们首先通过循环，挑选出了二叉树上最佳的变量个数。接着，我们通过绘图找出了最佳的树的个数。从图 2 中看出，n_tree 在 1200 附近时，模型误差较低，且较为稳定，因此，将树的个数设为 1200。

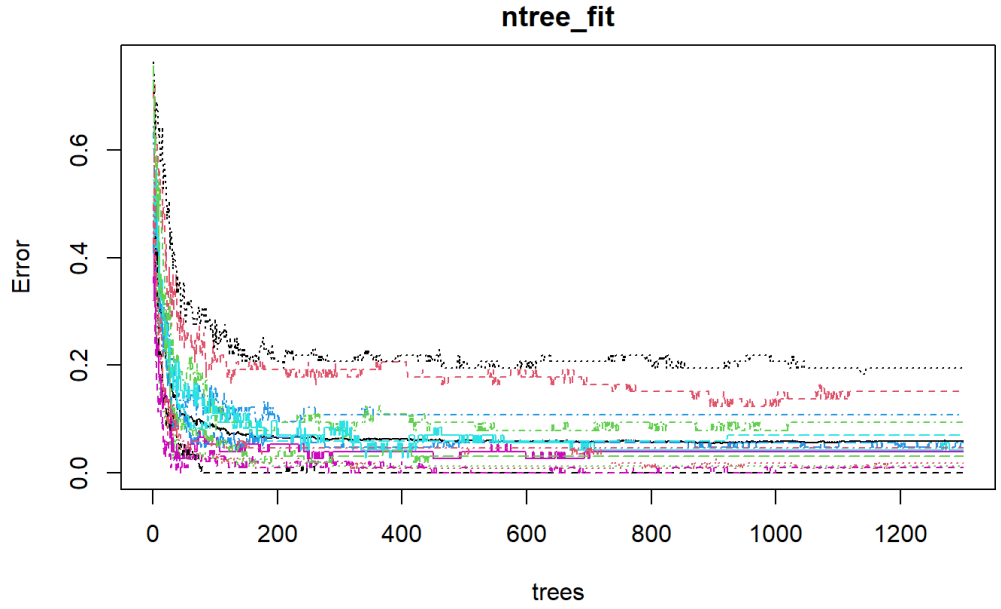

图 2：随机森林树的个数筛选

3. 实验结果

3.1. 模型预测

我们按照 8：2 的比例，将文本划分为训练集和测试集，并用 set.seed 设置随机种子，以保证结果的可重复性。所有模型均在相同的训练集和测试集上进行拟合和评估。

在测试集上，支持向量机的准确率达到了 88%，相较最初的朴素贝叶斯分类器有了明显的提升。图 3 展示了其混淆矩阵的可视化结果，从中可以看出，支持向量机的错分项大多都被错分到了“朱自清”，且这一现象在后续实际数据中更为明显。

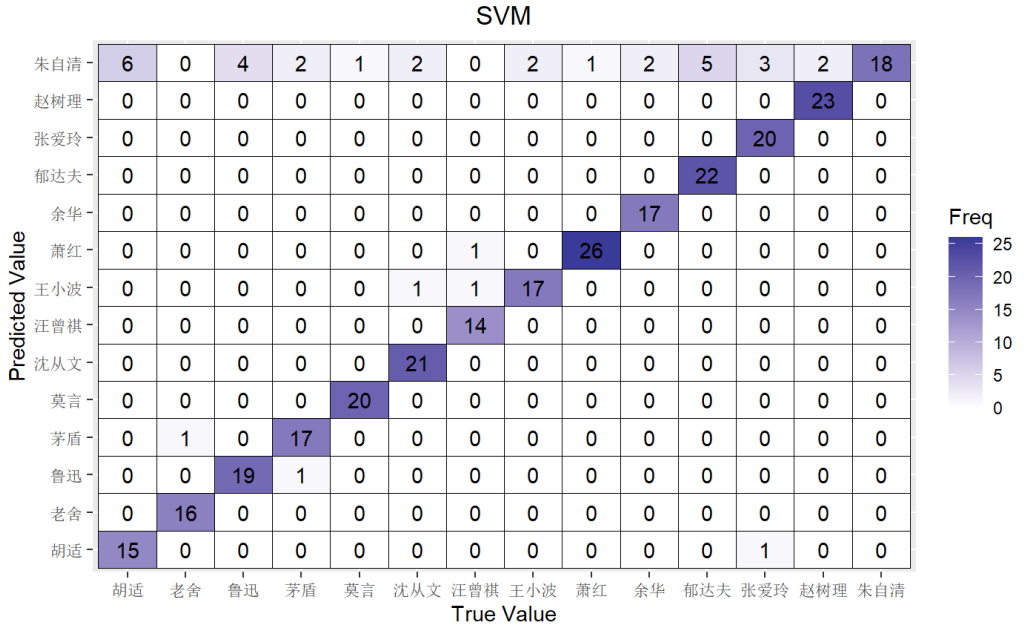

图 3：支持向量机混淆矩阵可视化

弹性网回归在测试集上的分类准确率高达 96.7%，且在这一模型下，大量错分为一类的现象不再出现。图 4 为其混淆矩阵的可视化结果。

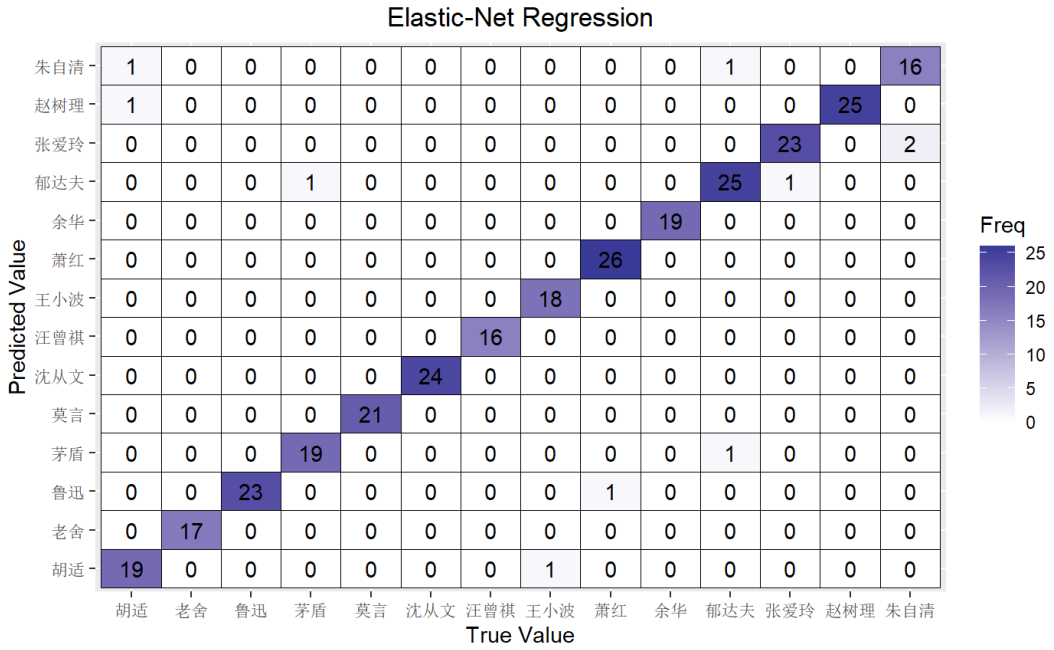

图 4 ：弹性网回归混淆矩阵可视化

随机森林在测试集上的准确率为 95.7%，相较弹性网回归并没有提升。图 5 为其混淆

矩阵可视化结果。

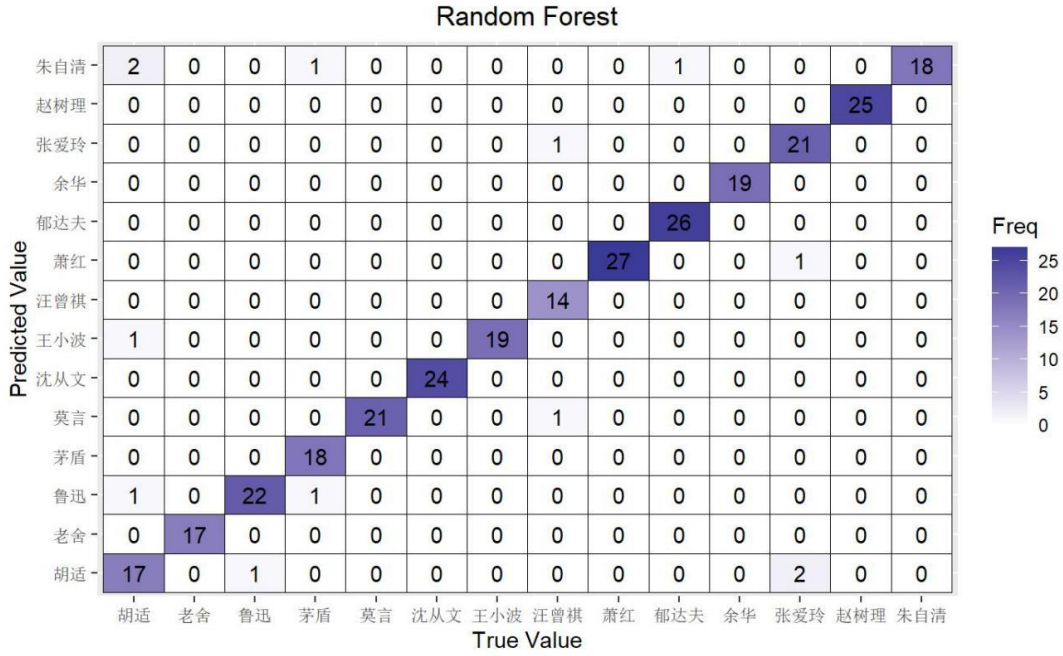

图 5：随机森林混淆矩阵可视化

3.2. Rshiny

根据搭建完的模型，我们实现了一个可交互的 Rshiny App。用户可以在输入（上传）一定量的文本后选择对应的模型，从而得到与文本风格最相似的作家预测，实现文章风格的分类。

该应用综合了先前实现的特征提取和模型训练，在初始化时完成这两个步骤。前端用户上传文本可以通过文件上传（图 6），也可以直接输入文本（图 7），分别使用了 fileInput 和 textAreaInput 两个 UI 实现交互。服务器部分使用多个 observeEvent 函数实现了对不同事件的处理，具体细节实现可以参考相应代码。

图 6 风格分类器演示（上传文件）

图 7 风格分类器演示（输入文本）

4. 结果讨论

4.1. 结果解读

在搭建完模型之后，我们进一步对重要的特征进行了可视化，并且对结果进行了一定的解读。

图 8 是通过随机森林得到的特征重要性排序，可以看出重要的特征都主要集中在标点的使用、不同词性的词的使用，以及"但"，"假如"等常见虚词的使用上。

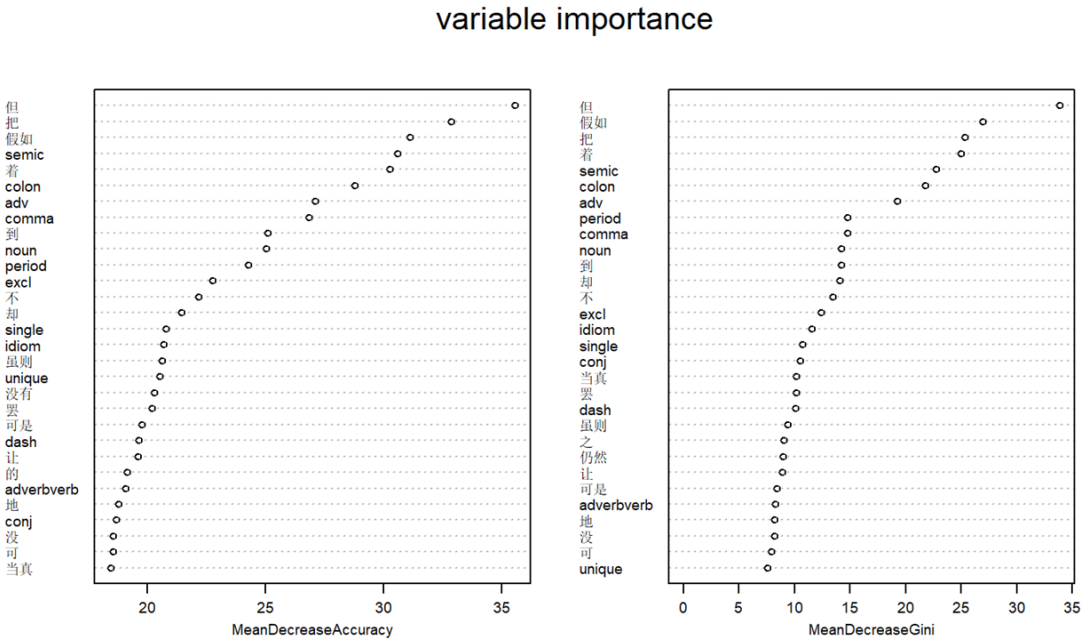

图 8：特征重要性排序

进一步，我们更加精细地对上面提到的一些重要特征进行了可视化。值得注意的是，我们在选用文本的时候，除了不考虑诗歌这种较为特殊的文体之外，并没有对其余文体做出严格区分，而且所有作家的文本都是来源于多个文集。因此，筛选出的这些特征还是能比较真实地反映一个作家整体的写作风格。

图 9 是对于标点使用的可视化。从图 中可以看出，重要性排名比较靠前的有冒号、分号和破折号的使用。比如我们拿破折号 dash 来举例，对它使用频率最高的是作家汪曾祺的作品，而这也是学者们公认的一个汪曾祺小说中很有个性的现象。这一现象与他口语化的

表达风格和他喜欢详尽叙述的写作习惯是分不开的。当用到比较深奥的词语，或者在人物、情节出现变化时, 他便常常要在文字中，通过破折号插入通俗的注释，从而让读者更好地理解文章内容。

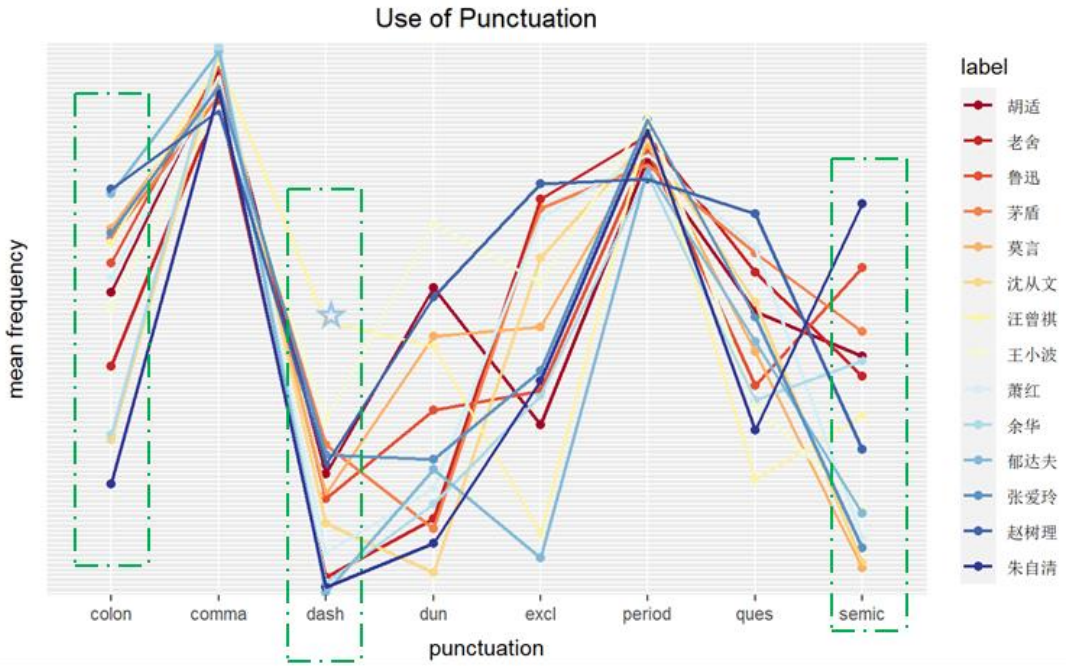

图 9：标点符号使用

图 10 是对于不同词性的词的使用。比如我们可以看到，朱自清先生对于形容词、副词和动词的使用频率都是比较高的，而这一特点与他追求真挚美、绘画美和理趣美的散文风格也是相吻合的。

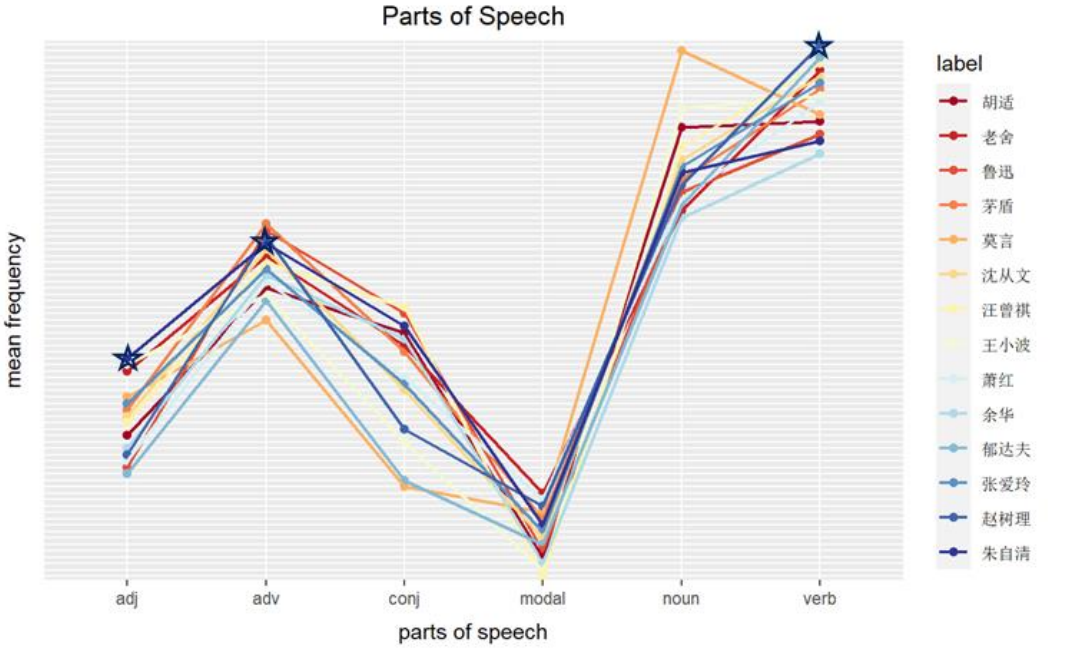

图 10：不同词性词语的使用

图 11 包含了四个虚词的使用情况，其中，“但”、“假如”和“却”都是连词。通过画红色

三角和蓝色圆圈的地方可以看出，作家们对于某一类虚词的使用还是能呈现出较为规律的偏好，比如鲁迅往往会比老舍更喜欢用连词。而绿色框住的部分则从另一个角度说明，即便“却”和“但”都是表示转折的连词，作家们对它们的使用也会有一定的差别，而这种对单个词的偏好则很大程度体现了个人语言风格和写作习惯的差异。

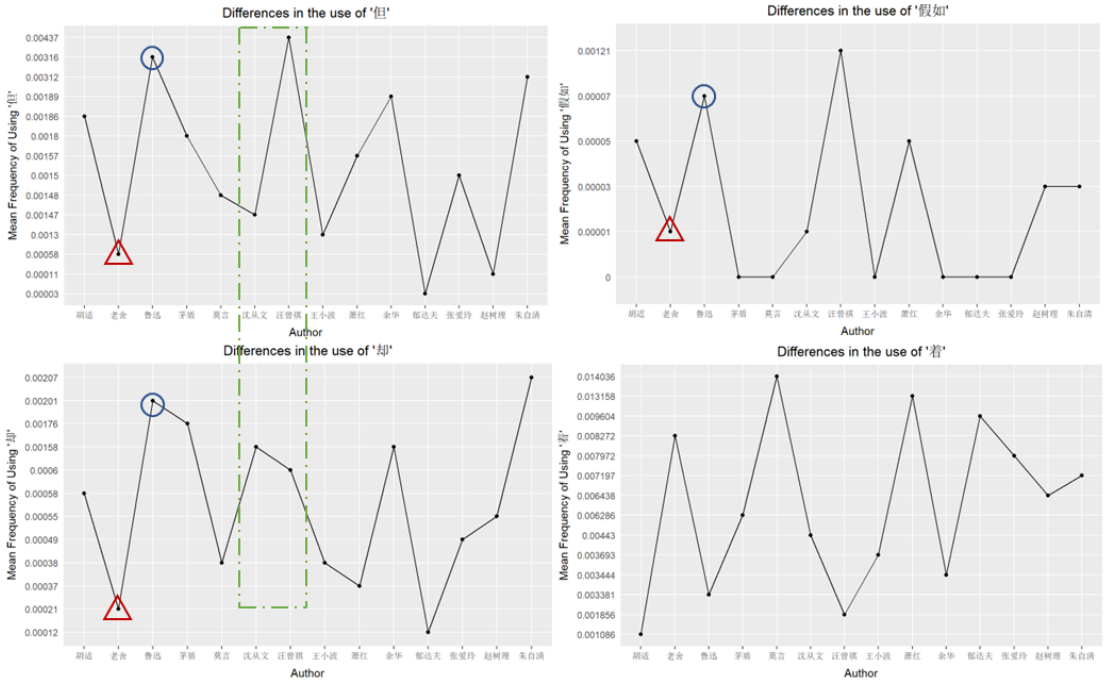

图 11：部分虚词使用

当然，我们已经在做大作业的过程中充分体会到，风格分析是一个非常精细的任务，因此，我们当然不能直接通过这些有限的定量的数据下结论。但是，这样的分析思路和方法能够在计量风格学中发挥重要的作用，也可以用来作证学者们对于作家风格的定性分析，从而进一步加深人们对中国作家和他们作品的理解。

4.2. 不足

一、实验的特征工程完全凭借专业知识人工筛选，因此难以覆盖全面。未来在特征选择方面，我们可以进行更为充分的文献调研，参考前人的经验以及建议，进行多次尝试，从而更加充分地评估特征的重要性。

二、在特征选择部分只选取了一些通过计算机程序可以自动抽取的风格特征，而这样的做法实际上损失了大量更为复杂的语义、句法的信息。在后续研究中，可以考虑加入知识图谱学习上下文特征，通过语义网等语义学习模型来解决语义句法特征的学习问题。

5. 总结

根据文章写作风格鉴定作者是一项历史悠久的研究，从中可以衍生出多种应用方向，例如鉴别存在争议的作者、鉴定作品是否存在抄袭、以及国内相当关注的《红楼梦》80 章前后的作者问题等等。

本实验对写作风格鉴别这一任务进行了较为充分的探索，首先通过爬虫对作家文本进行了爬取，接着针对标点符号使用、虚词使用等六个大类特征进行了特征提取，并运用多种机器学习算法（朴素贝叶斯、支持向量机、弹性网回归和随机森林）进行了作家文本的十四分类，在测试集上达到了 96.7%的分类准确率。此外，我们对筛选出的重要特征进行

了可视化和解读，并用 Rshiny 实现了一个可交互的小应用。最后，我们对本实验的贡献和不足之处进行了讨论和总结。

6. 实验分工

孙佳仪：特征提取、弹性网回归、Rshiny 框架搭建、改进与不足
肖桐：特征提取、朴素贝叶斯、弹性网回归、随机森林、结果可视化与解读
郑克寒：特征提取、爬虫、支持向量机、Rshiny 细节实现

7. 参考文献

1. Mendenhall T C．The characteristic curves of composition[J]．Science，1887：237～249
2. Khrnelev D V，Tweedy F J．Using Markov chains for identification of Writers[J]. Literary and Linguistic Computing，2001，16(4)：299～307
3. 孙晓明,清华大学计算机科学与技术系(北京),马少平. 基于写作风格的作者识别[C]. //中国中文信息学会二十周年学术会议论文集.北京:中国中文信息学会, 2001:198-204.
4. 刘莎,陈艳平. 基于 BiLSTM 写作风格识别方法研究[J]. 计算机与数字工程,2021,49(9):1829-1833.
5. 刘明勇. 基于写作风格学的作者识别技术研究[D]. 浙江:浙江大学,2013.

8. 代码

https://cloud.tsinghua.edu.cn/d/4f61e7dcc76c477197fd/