**基于主成分分析法的《红楼梦》前后作者研究**

**摘要**

对《红楼梦》的作者的研究是“红学”的一个重要方向。前八十回为曹雪芹本著，后四十回为高鹗续著是目前公认的版本。本文以主成分分析法、相对熵(KL散度)为理论基础，对《红楼梦》前八十回和后四十回是否写自两个作者进行了探究。

针对问题一，我们将55个常见的虚字作为特征字，统计了它们在每一章回出现的频率。建立主成分分析模型，对经过预处理的数据进行主成分分析，将55维的数据降至3维，同时最大程度上保留了原始信息。再用MATLAB绘制三维散点图，直观地展现出前八十回和后四十回点的分布呈现出的差别。为了进行更加严谨的证明，我们还对不同的数据预处理方法得出的结果进行了比较，发现均能较明显地体现出前八十回和后四十回的差别。

针对问题二，我们从词量和词频两个方面进行分析。首先建立词量分析模型，从出现次数较高并且有意义的词语中随机选择部分词语，得到50个随机样本。统计这些词语在每一章节中出现的次数，利用问题一的主成分分析模型，将数据降至三维。用MATLAB绘制三维散点图，分析点的分布，可以观察到前八十回和后四十回点的分布存在的明显的差异。再建立词频分析模型，首先计算出小说中每个词语的词频，在高频词语中删去无意义以及和情节关系紧密的词语，得到50个特征词。统计特征词在每个章回中出现的频率，利用主成分分析模型降维，绘制散点图，观察到前后部分的点几乎分布在互不重合的两个区域。结合两个方面可以得出，前八十回和后四十回属于两个不同的作者。

针对问题三，我们从语句和语义两个方面分析。由于不同作者写作时的平均语句长度存在一定的区别，因此从平均句长入手，建立语句分析模型。首先对小说文本进行分句处理，得到语句数量相关数据。再将小说分为前四十回、中四十回和后四十回三部分，计算得每一部分的平均句长分别为6.9925字、7.07923字和7.6544字。由此可见，前四十回合中四十回之间平均句长相近，而后四十回与前两部分的平均句长存在较大的差异。

由于虚字能够较好地体现作者的语义表达习惯，我们从虚字入手分析前后语义之间的差别。将每十个章回看做一个单位，利用Matlab统计各虚字相对于所有虚字在每一部分中出现的概率，得到12个随机变量。求每个随机变量和最后四个随机变量之间的相对熵（KL散度），并画出柱状图。分析可知，前8个单位和后4个单位之间的相对熵普遍大于后4部分之间的相对熵，因此《红楼梦》前八十回和后四十回之间的语义存在较大的差别。结合两个方面可得，前八十回和后四十回属于两个不同的作者。

**关键词：** 《红楼梦》作者 主成分分析法 相对熵

# 问题的重述

《红楼梦》是清代作家曹雪芹创作的长篇章回体小说，是中国古典四大名著之一，中国古典小说的巅峰之作。但在流传过程中，由于某些历史原因，出现了缺失，目前公认的版本中前80章回为曹雪芹本著，后40章回为高鹗续著。用数学建模的方式来证明前80章回和后40章回的作者不同，是研究“红学”的一个重要课题。

要求通过建模分析，解决以下关于《红楼梦》的问题。

（1）通过利用统计《红楼梦》120 章回书中的“红”、“玉”等字出现的频率，建模来佐证，前 80 章回和后 40 章回的作者不同。

（2）通过建模分析，《红楼梦》前80章回和后40章回中的词量和词频，来说明80章回和后40章回的作者不同。

（3）通过建模分析，《红楼梦》前80章回和后40章回中的语句和语义，来说明80章回和后40章回的作者不同

# 问题的分析

问题一：根据题目的提示，需要找到若干特征字，计算出它们在每一章回出现的频率，通过比较前八十回和后四十回特征词出现频率的差别，佐证前后为两个作者。在特征词的选取上，它们需要满足受情节影响小，能体现不同作者写作风格的要求。如果数据的维度过大，会对分析造成困难，因此可以用主成分分析法对数据进行降维，同时最大程度地保留数据包含的信息。最后可以用MATLAB绘制散点图，将前八十回和后四十回的区别直观的反映出来。

问题二，从词量和词频两个方面进行分析。针对词量，从小说中出现次数较多的词语中随机选取样本词语，删去其中无意义的词汇。再统计出每个词语在每一章回中出现的次数，利用成分分析法对数据进行降维，绘制三维散点图，体现前后的区别。针对词频，在高频词中删去无意义的以及和情节紧密相关的词语，得到特征词。统计出特征词在每一章回出现的频率，对数据进行成分分析，绘制散点图，观察前后部分之间的区别。

问题三，是通过《红楼梦》前 80 章回和后 40 章回中的语句和语义来说明前 80章回和后40章回的作者不同。首先我们通过语句来分析，根据作者写作方式和语言习惯不同导致语句长短不同，利用MATLAB工具分别计算出各章回的语句数，以及每个语句的字数，通过分析比较1-40回，41-80回，81-120回平均句长，得出前 80章回和后40章回的作者是否不同的结论。我们再用语义来分析，虚词是一种很好的体现作者写作风格和语义特点的表现方式，因此我们从55个虚词入手进行分析。我们引进了相对熵（KL散度）的概念，相对熵越小，文本间随机分布的差别越小，因此通过分别计算每个虚词在每十章回中出现的频率，利用相对熵的计算方式，可以得出每十回合对后四十回合相对熵，作图分析得出1-40回、41-80回、81-120回之间的关系。

# 模型的假设

（1）各字词之间相互独立。

（2）同一作者的写作风格基本保持稳定。

（3）平均句长不受小说情节的影响。

（4）第二问中样本是随机选取的所以具有普遍性。

（5）特征词的选取上忽略一词多义对结果的影响。

# 符号说明

|  |  |
| --- | --- |
| 符号 | 说明 |
|  | 样本的数量 |
|  | 特征字的总量 |
|  | 频率原始数据 |
|  | 经过预处理数据 |
|  | 协方差矩阵 |
|  | 协方差 |
|  | 矩阵的特征根 |
|  | 特征向量 |
|  | 主成分的组成个数 |
|  | 第i个样本的第j个主成分 |
|  | 第i个主成分的方差贡献率 |
|  | 第i部分语句数量 |
|  | 第i部分总字数 |
|  | 第i部分平均句长 |
|  | 各虚字在每一单位出现的次数 |
|  | 每一单位所有虚字出现次数之和 |
|  | 出现的频率 |
|  | 相对熵 |

# 模型的建立和求解

## 问题一模型的建立与求解

### 模型的简介

首先选取若干特征字，计算每一章回中各个特征字出现的频率，再建立主成分分析模型对数据进行降维，同时最大程度上保留原始数据包含的信息，再通过散点图直观地分析前八十回合后四十回之间写作风格的差别。

### 特征字的选取

由于虚字的使用受小说情节的影响较小，与作者地域、经历等因素有关，在不同作者之间呈现出较为明显的差别，而对于同一作者则比较稳定。因此虚字能够较为合理地体现出不同作者的写作风格。

### 主成分分析模型的建立

假定选取了 个样本， 个特征字，可以构成一个乘以阶的特征词频率原始数据矩阵，



（1）数据预处理

为了消除量纲不同，自身变异或者数值相差较大产生的误差，首先对得到的原始数据进行预处理：

方法一，Min-Max标准化



方法二，零均值化处理



其中，



得到处理后的矩阵。

（2）主成分的得出

首先求矩阵的协方差矩阵



其中， 

解协方差矩阵的特征方程，得到矩阵的特征根，通过解方程组**得矩阵的特征向量。

将变量转化为主成分，

表示第i个样本的第j个主成分，从而得到主成分矩阵。

运用数学知识可以证明，协方差矩阵的特征根等于对应主成分的方差。主成分的方差贡献率体现了主成分的重要程度，



选取方差贡献率最大的前三个主成分，通过绘制三维散点图，比较前八十回和后四十回写作风格的差异。

### 模型的求解

由于《红楼梦》同时运用了文言和白话进行写作，因此本文选取了文言文中的18个虚字和常见的白话虚字，共55个，作为特征字，如表 5‑1所示。

表 5‑1选取的55个特征字

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 于 | 者 | 若 | 亦 | 呀 | 向 | 很 | 在 |
| 也 | 何 | 之 | 方 | 吗 | 往 | 偏 | 别 |
| 而 | 为 | 焉 | 即 | 咧 | 就 | 了 | 好 |
| 乎 | 因 | 然 | 皆 | 罢 | 但 | 的 | 是 |
| 以 | 乃 | 矣 | 仍 | 么 | 越 | 着 | 便 |
| 所 | 则 | 且 | 故 | 呢 | 再 | 不 | 比 |
| 与 | 其 | 或 | 尚 | 让 | 更 | 把 |  |

将文本进行处理，去除标点符号、去无意义文字后，利用MATLAB统计出各个特征字在每一章回出现的频率，部分数据如表5‑1所示。

表 5‑2各章回特征字频率部分数据

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 回数 特征字 | 不 | 把 | 罢 | 便 | 别 | 的 | … |
| 第1回 | 0.01694 | 0.000873 | 0.001222 | 0.006287 | 0.000349 | 0.006287 | … |
| 第2回 | 0.019135 | 0.00104 | 0.000416 | 0.004784 | 0.001248 | 0.013727 | … |
| 第3回 | 0.014523 | 0.000296 | 0.001482 | 0.005335 | 0.002816 | 0.012893 | … |
| 第4回 | 0.020009 | 0.00043 | 0.000215 | 0.004303 | 0.000645 | 0.018287 | … |
| 第5回 | 0.012073 | 0.001127 | 0.000483 | 0.004507 | 0.000966 | 0.009659 | … |
| 第6回 | 0.016748 | 0.001005 | 0.003015 | 0.004522 | 0.001172 | 0.031988 | … |
| 第7回 | 0.016496 | 0.00202 | 0.002188 | 0.006396 | 0.00202 | 0.028783 | … |
| 第8回 | 0.014205 | 0.001136 | 0.00322 | 0.005303 | 0.00322 | 0.021023 | … |
| 第9回 | 0.023547 | 0.001296 | 0.00108 | 0.003889 | 0.001728 | 0.023331 | … |
| 第10回 | 0.022722 | 0.000957 | 0.001674 | 0.000478 | 0.001435 | 0.027984 | … |
| 第11回 | 0.016288 | 0.000643 | 0.004072 | 0.000429 | 0.002357 | 0.021217 | … |
| 第12回 | 0.019275 | 0.001438 | 0.001726 | 0.002014 | 0.003452 | 0.013234 | … |
| 第13回 | 0.019184 | 0.000486 | 0.0017 | 0.004371 | 0.001943 | 0.013113 | … |
| … | … | … | … | … | … | … | … |

将原始数据矩阵用 Min-Max标准化进行处理，再利用MATLAB实现主成分分析法。选取前三个主成分，绘制三维散点图，如图 5‑1所示，其中蓝色“\*”表示前八十回的点，红色“+”表示后四十回的点。

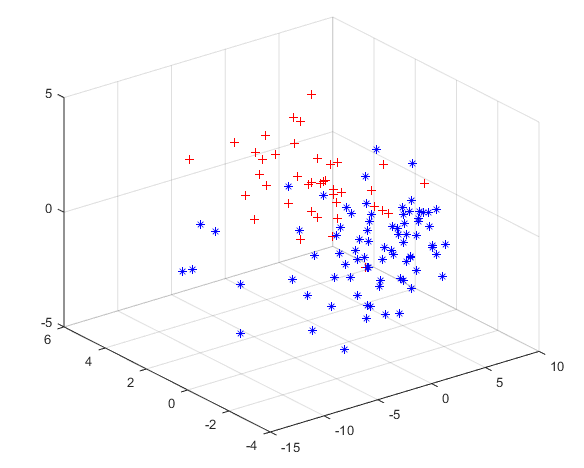


图 5‑1Min-Max标准化处理散点图

通过三维散点图，可以看出前八十回和后四十回的点出现了分层，位于两个几乎不重叠的区域。能够佐证，《红楼梦》前八十回和后四十回的写作风格差别较大，属于两个不同的作者。

但是用Min-Max标准化的方法处理原始数据，得到的前三个主成分对原始信息的保留率欠理想。主成分的方差贡献率能够反映主成分对原始数据信息的保留完整度。将各主成分方差的贡献率和累计贡献率反映在图表上，得到用Min-Max标准化处理后数据的pareto图，如图 5‑2所示，

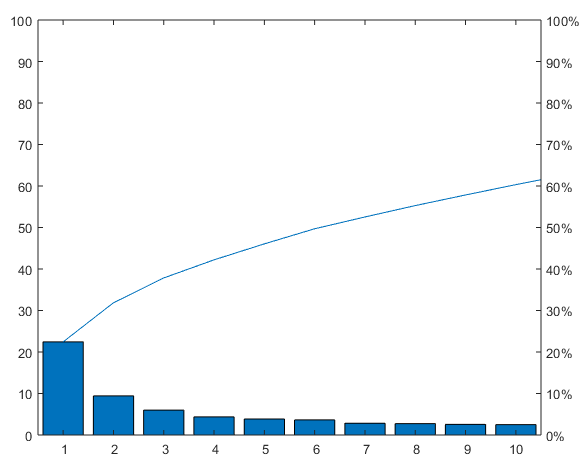


图 5‑2标准化处理数据pareto图

从图中可以看出，标准化处理后得出的三个主成分的方差累计贡献率并不理想，仅为37.85%，这表明前三个主成分对原始数据中的信息保留率不高。虽然这部分信息已经能够反映出前八十回和后四十回写作风格之间的区别，但是为了使证明更加严谨，下面换用零均值化的方法对数据进行预处理，进行主成分分析，得到三维散点图和pareto图，如下所示，

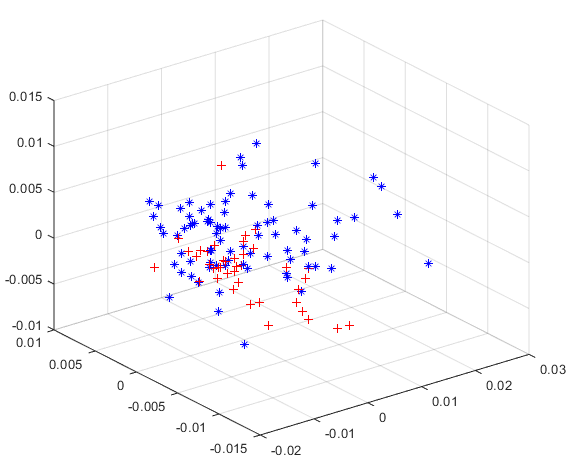


图 5‑3零均值化处理散点图

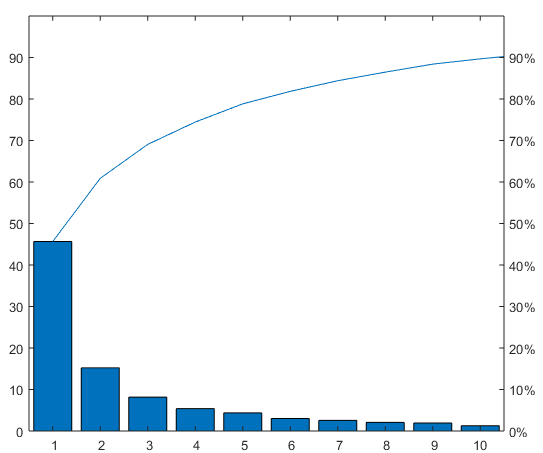


图 5‑4零均值化处理pareto图

从上图可以看出，数据经过零均值化预处理后，前三个主成分的累计方差贡献率较高（达到近70%），较好地保留了原始数据中的信息。三维散点图中前八十回和后四十回点的分布同样呈现出较大差别。

综上所述，《红楼梦》前八十回和后四十回的写作风格呈现出较为明显的差异，可以佐证前八十回和后四十回的作者不是同一人。

## 问题二模型的建立与求解

### 模型的建立

建立词量、词频分析模型。首先利用分词工具统计《红楼梦》的总词量以及每一章回的词量。

（1）词量分析模型的建立

由于全部词语的数量较难进行分析，因此在出现次数较高的词语中，删去无意义的词语，随机选取50个样本词语进行词量的分析。

选出样本后，统计出这些词语在每一章回中出现的次数，利用问题一种建立的主成分分析模型对数据进行降维，选取方差贡献率最高的三个主成分绘制三维散点图，分析前八十回和后四十回点的分布。

（2）词频分析模型的建立

·特征词的选取

由于分词工具统计出的部分词语并不是具有实际意义的词汇，且部分词语受情节的影响较为明显，并不能作为特征词区分不同作者的写作风格。例如，出现次数最多的前二十个词语如表 5‑3所示，

表 5‑3出现次数最多的前二十个词语

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 词语 | 次数 | 词语 | 次数 | 词语 | 次数 | 词语 | 次数 |
| 宝玉 | 3353 | 贾母 | 1106 | 知道 | 883 | 出来 | 836 |
| 什么 | 1501 | 我们 | 1092 | 你们 | 879 | 姑娘 | 813 |
| 一个 | 1199 | 说道 | 933 | 王夫人 | 874 | 众人 | 810 |
| 凤姐 | 1128 | 如今 | 903 | 起来 | 868 | 太太 | 790 |
| 那里 | 1125 | 这里 | 891 | 老太太 | 840 | 怎么 | 789 |

其中“凤姐”“贾母”“王夫人”“老太太”等表示人名或者称呼的词语与小说的情节发展有着紧密的联系，无法作为评判《红楼梦》前后作者不同的标准。

因此在出现频率较高的词语中，删去没有实际意义，以及受情节影响较大的词语，得到50个特征词，如表 5‑3所示，

表 5‑4选取的50个特征词

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 词语 | 次数 | 词语 | 次数 | 词语 | 次数 | 词语 | 次数 |
| 宝玉 | 3353 | 他们 | 751 | 告诉 | 558 | 过来 | 401 |
| 什么 | 1501 | 只见 | 749 | 回来 | 537 | 心里 | 398 |
| 一个 | 1199 | 没有 | 741 | 只得 | 510 | 一时 | 376 |
| 那里 | 1125 | 不是 | 704 | 只是 | 497 | 不能 | 372 |
| 我们 | 1092 | 听见 | 698 | 大家 | 488 | 答应 | 355 |
| 说道 | 933 | 两个 | 676 | 不敢 | 472 | 今日 | 350 |
| 如今 | 903 | 不知 | 674 | 这些 | 467 | 几个 | 340 |
| 这里 | 891 | 这个 | 665 | 这样 | 466 | 那边 | 323 |
| 知道 | 883 | 就是 | 634 | 所以 | 454 | 只管 | 322 |
| 你们 | 879 | 进来 | 577 | 出去 | 428 | 还有 | 322 |
| 起来 | 868 | 一面 | 566 | 这么 | 426 | 说话 | 321 |
| 出来 | 836 | 咱们 | 562 | 不好 | 415 |  |  |
| 自己 | 767 | 东西 | 559 | 不过 | 410 |  |  |

·主成分分析

用MATLAB统计出特征词在每一章回中出现的频率，得到原始数据的矩阵。用问题一中建立的主成分分析模型，将维数较高的数据降维，得到方差贡献率最大的前三个主成分。绘制三维散点图，比较前八十回合后四十回点的分布。

### 模型的求解

（1）词量分析模型的求解

在出现次数较高的词语中随机选取样本，删去其中没有意义的词汇，得到50个随机样本词语。对其进行主成分分析后，选取方差贡献率最高的三个主成分，绘制三维散点图，如图 5‑5所示，

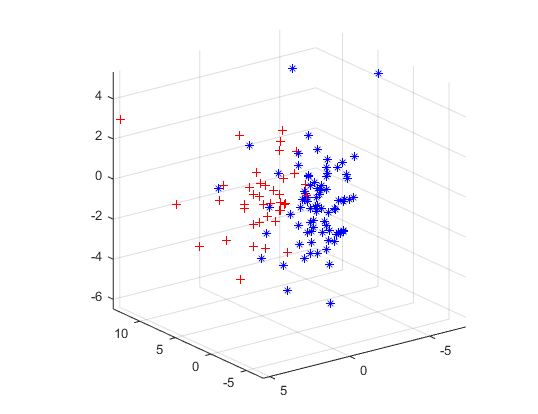


图 5‑5词量分析散点图

从散点图中可以看出，前八十回和后四十回的点大致分布在两个几乎不重合但又彼此接近的区域，较为明显地体现出了《红楼梦》前后风格在词量上的区别。

主成分的pareto图如图 5‑6所示，

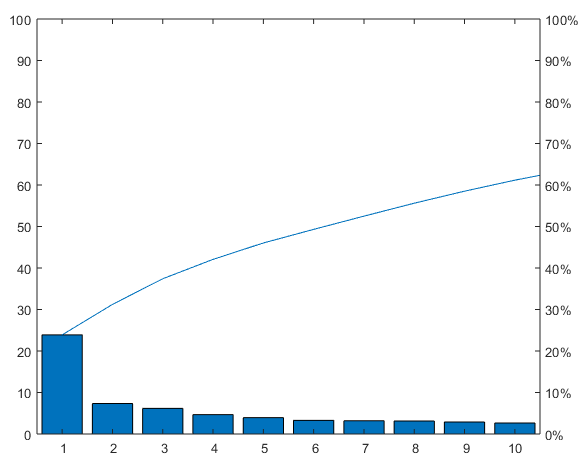


图 5‑6词量分析pareto图

尽管前三个主成分的累计方差贡献率并不高，但是它们蕴含的信息已经能够较为明显地体现出前八十回和后四十回在词量上的区别。因此，在对词量的分析上体现出前后作者并非同一个人。

（2）词频分析模型的求解

首先，用MATLAB统计出50个特征词在每一章回中出现的频率，部分数据如表 5‑5所示，

表 5‑5特征词频率部分数据

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 回数 特征词 | 起来 | 知道 | 不知 | 出来 | 如今 | 自己 | … |
| 第1回 | 0.001006 | 0.000503 | 0.004525 | 0.001508 | 0.002011 | 0.004525 | … |
| 第2回 | 0.002514 | 0.001257 | 0.003771 | 0.002514 | 0.005657 | 0.001257 | … |
| 第3回 | 0.001316 | 0.001755 | 0.002194 | 0.001316 | 0.002194 | 0.001755 | … |
| 第4回 | 0.000654 | 0.001962 | 0.001308 | 0.000654 | 0.00654 | 0.000654 | … |
| 第5回 | 0.000467 | 0.000467 | 0.005605 | 0.001401 | 0.002335 | 0.000467 | … |
| 第6回 | 0.0028 | 0.005599 | 0.003359 | 0.003359 | 0.007279 | 0.00224 | … |
| 第7回 | 0.002292 | 0.004011 | 0.004011 | 0.005158 | 0.002865 | 0.004011 | … |
| 第8回 | 0.002614 | 0.003268 | 0.001961 | 0.003268 | 0.001307 | 0.003268 | … |
| 第9回 | 0.006307 | 0.001402 | 0.002102 | 0.004205 | 0.004905 | 0.005606 | … |
| 第10回 | 0.001597 | 0.001597 | 0.003994 | 0.003195 | 0.00639 | 0.000799 | … |
| … | … | … | … | … | … | … | … |

利用主成分分析法，50维的数据降为3维，三维散点图以及主成分的pareto图，如下所示，

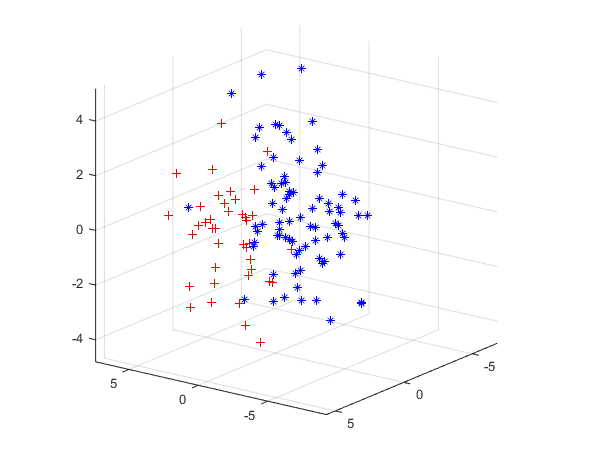


图 5‑7词频分析散点图

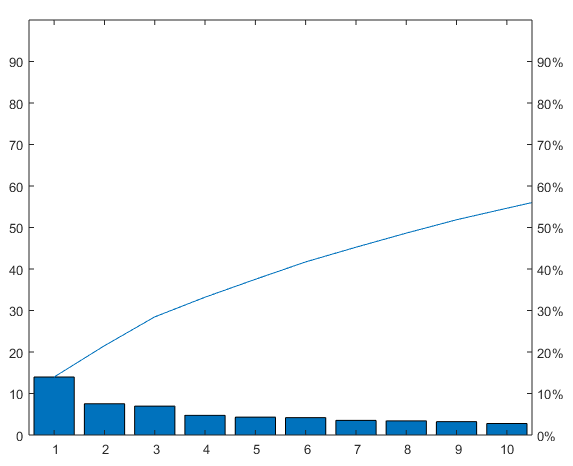


图 5‑8词频分析pareto图

从散点图中可以看出前八十回和后四十回特征词出现的频率有着较为明显的差别，红色和蓝色的点仍旧分布在两个几乎不相重叠的区域。尽管前三个主成分的方差累计贡献率不高，但是这部分信息已经足够体现出前八十回和后四十回在词频上的区别。

结合对词量和词频的分析，《红楼梦》前八十回和后四十回存在这较为明显的区别，能够推测是写自两个不同的作者。

## 问题三模型的建立与求解

### 模型的建立

从语句和语义两方面对《红楼梦》前八十章回和后四十章回进行建模分析。

（1）语句分析模型的建立

·文本的处理

不同作者写作时，句子的平均长度之间会存在这一定的差异。为了便于语句数量的统计，我们首先对文本进行处理。首先将标点符号分为两类：分句的标点，包括句号、问号、感叹号、句号、破折号、分号、冒号；不分句的标点，包括顿号、书名号、双引号、单引号。删去文本中所有不分句的标点，多余的行以及与内容无关的文字，将分句的标点替换成空格。再把文本导入Excel表格，以空格作为分隔符将每一短句分别放在一个单元格内。

·平均句长的计算

将小说分为前四十回、中四十回和后四十回，分别计算出三个部分的语句数量 ，以及字数 ，得到每一部分的平均句长。



绘制图表，比较三个部分的平均句长。

（2）语义分析模型的建立

由于虚字能够较好地体现作者的写作风格，因此仍然选用第一问中的55个虚字对《红楼梦》的语义进行分析。

·虚字出现频率的计算

为了简化计算方便比较，将每十回看做一个单位，得到12个随机变量。统计各虚字在每一单位中出现的次数，和每一单位所有虚词的总次数之和



计算某虚字在某一单位，相对于所有虚字出现的频率，即随机变量取第j个值的概率，

·相对熵的计算与比较

相对熵（K-L散度）可以量化两种概率分布P与Q之间的差异，其表达式为，



相对熵越大，表明两个随机变量之间的差别越明显。

分别计算每一部分和最后四部分（后四十回）之间的相对熵，得出前八十回和后四十回，以及后四十回之间的差异。

### 模型的求解

（1）语句分析模型的求解

利用Excel统计出前四十回、中四十回和后四十回的句子数量如表 5‑6所示，

表 5‑6前、中、后三部分句子数量

|  |  |  |  |
| --- | --- | --- | --- |
| 回数 | 1~40回 | 41~80回 | 81~120回 |
| 句子数量 | 31333 | 36390 | 31820 |

再对字数进行统计，得到三部分的字数如表 5‑7所示，

表 5‑7前、中、后三部分字数

|  |  |  |  |
| --- | --- | --- | --- |
| 回数 | 1~40回 | 41~80回 | 81~120回 |
| 字数 | 219096 | 257613 | 243563 |

用三部分句子数量、字数的数据求得前、中、后三部分的平均句长如表 5‑8所示，

表 5‑8前、中、后三部分平均句长

|  |  |  |  |
| --- | --- | --- | --- |
| 回数 | 1~40回 | 41~80回 | 81~120回 |
| 平均句长 | 6.9925 | 7.079225 | 7.6544 |

将表格中的数据绘制成柱状图，如表 5‑8所示，

图 5‑9前、中、后三部分平均句长柱状图

从柱状图中可以直观地看出，1~40回和41~80回的平均句长较为接近，而81~120回的平均句长和前两部分有着较为明显的差别。

比较前四十回、后四十回相对于中四十回的差别，分别为0.0123和0.0812，后者是前者的6.6倍，说明前四十回和中四十回的平均句长比较接近，它们和后四十回之间的差异较为明显。这证明了《红楼梦》前八十回和后四十回属于两个不同的作者。

（2）语义分析模型的求解

以每十个章回为一个单位，计算可得每个单位各虚字出现频率的部分数据如表 5‑9所示，

表 5‑9每个单位各虚字出现的频率（部分）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 章回  频率  虚字 | 01~10 | 11~20 | 21~30 | 31~40 | 41~50 | … |
| 不 | 0.108246 | 0.118314 | 0.122055 | 0.11497 | 0.120534 | … |
| 把 | 0.006388 | 0.005335 | 0.012084 | 0.011776 | 0.010503 | … |
| 罢 | 0.00964 | 0.011471 | 0.016755 | 0.016467 | 0.015901 | … |
| 便 | 0.029733 | 0.022276 | 0.033611 | 0.026148 | 0.021398 | … |
| 别 | 0.009988 | 0.013872 | 0.012388 | 0.012774 | 0.011091 | … |
| 的 | 0.119861 | 0.105776 | 0.117486 | 0.127146 | 0.134079 | … |
| … | … | … | … | … | … | … |

计算每一部分和最后四部分之间的相对熵，得到数据如表 5‑10所示，

表 5‑10每十回对后四十回KL相对熵

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 后四十回  KL值  回合数 | 81~90 | 91~100 | 101~110 | 111~120 |
| 01~10 | 0.149564 | 0.145603 | 0.155369 | 0.142997 |
| 11~20 | 0.102445 | 0.099769 | 0.111045 | 0.117174 |
| 21~30 | 0.019738 | 0.038022 | 0.043636 | 0.06554 |
| 31~40 | 0.023028 | 0.033701 | 0.042418 | 0.065193 |
| 41~50 | 0.025041 | 0.035408 | 0.04492 | 0.061645 |
| 51~60 | 0.046195 | 0.046881 | 0.053553 | 0.069538 |
| 61~70 | 0.035713 | 0.036542 | 0.044629 | 0.052198 |
| 71~80 | 0.073811 | 0.063604 | 0.080189 | 0.081651 |
| 81~90 | 0 | 0.022371 | 0.025725 | 0.04548 |
| 91~100 | 0.022346 | 0 | 0.011792 | 0.020665 |
| 101~110 | 0.027563 | 0.012239 | 0 | 0.021846 |
| 111~120 | 0.046034 | 0.023474 | 0.023694 | 0 |

将上表数据反映在柱状图上，如图 5‑10所示，

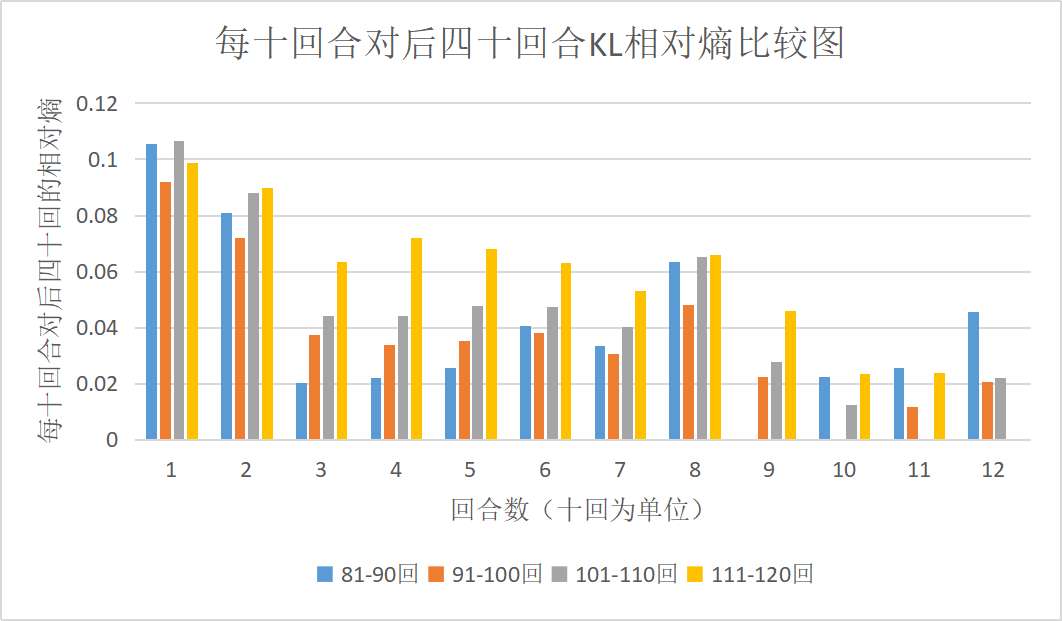


图 5‑10每十回合对后四十回合KL相对熵比较图

从上表数据以及柱状图中可以发现，前八十回与后四十回之间的相对熵普遍高于后四十回之间的相对熵。因此，可以得出前八十回和后四十回之间的差异较大，而后四十回之间的差异较小。能够说明前八十回和后四十回属于不同作者。

# 模型的评价与推广

## 模型的优点

（1）将虚字作为特征字衡量前八十回和后四十回之间的区别，具有不受情节影响、较好地体现作者写作风格的优点。

（2）在问题二特征词的选取上删去了与情节紧密相关的人名、称呼、地名等，使结果能够更加客观地展现前后两部分在词频上的区别，

（3）使用了主成分分析法，将多维的数据进行降维，使之便于分析，同时最大程度地保留了数据包含的原始信息。

（4）由于不同作者写作时在语句长度上存在差异，问题三中将平均句长作为比较前后语句差异的指标，较为合理。

（5）利用相对熵将语义在前后两部分之间的差异量化

## 模型的缺点

（1）特征字和特征词选取的数量较少，没有更为全面地展现出前八十回和后四十回之间的差异。

（2）第二问中三个主成分的累计方差贡献率较小，对原始数据信息的还原度欠理想。

（3）第三问平均句长在一定程度上还会受到情节的影响。

## 模型的推广

本模型从数学上对《红楼梦》前八十回和后四十回写自两个作者进行了证明，并从字频、词量、词频、语句、语义等方面对差异进行了具体的分析。对《红楼梦》作者的研究具有一定的参考意义，为前八十回和后四十回属于两个不同的作者这一学说，提供了数据上的支持。

# 参考文献

[1]李贤平.《红楼梦》成书新说[J].复旦学报(社会科学版),1987(05):3-16.

[2]陈大康.从数理语言学看后四十回的作者——与陈炳藻先生商榷[J].红楼梦学刊,1987(01):293-318.

[3]包辰瑶.《红楼梦》前八十回和后四十回词频的对比研究[J].湖北科技学院学报,2013,33(09):61-62+74.

[4]孙刘平,钱吴永.基于主成分分析法的综合评价方法的改进[J].数学的实践与认识,2009,39(18):15-20.

# 附录

附录一：

data1=xlsread('C:\Users\leafpigbirds\Desktop\database2.xlsx',2);

N=55+1;

data=data1(2:N,:)';%Ç°80»ØÊý¾Ý¶ÁÈ¡ 80\*55

Z=zscore(data);%±ê×¼»¯´¦Àí

% Z=data-mean(data);

cov\_M=cov(Z);%Ð­·½²î¾ØÕó

[Vectors,Values]=eig(cov\_M);%ÌØÕ÷ÏòÁ¿ ÌØÕ÷¸ù£¨Öµ£©

D=diag(Values);%ÌáÈ¡ÌØÕ÷¸ù

eig1=sort(D,'descend');%¹±Ï×ÂÊ´ÓÐ¡µ½´ó

V=fliplr(Vectors);%ÒÀÕÕDÖØÐÂÅÅÁÐ

S=0;

i=0;

while S/sum(eig1)<0.85

i=i+1;

S=S+eig1(i);

end

NEW=Z\*V(:,1:i);

figure(2);

W=100\*eig1/sum(eig1);

pareto(W);

figure(1);

% plot(NEW(1:40,1),NEW(1:40,2),'b\*');

scatter3(NEW(1:40,1),NEW(1:40,2),NEW(1:40,3),'b\*');

hold on;

% plot(NEW(81:120,1),NEW(81:120,2),'r+');

scatter3(NEW(41:80,1),NEW(41:80,2),NEW(41:80,3),'b\*');

hold on;

scatter3(NEW(81:120,1),NEW(81:120,2),NEW(81:120,3),'r+');

附录二：

# -\*- coding: utf-8 -\*-

# coding=utf-8

import collections

f1 = open("sum\_OUT",'a',encoding='utf8')

for i in range(121)

name=str(i)+".txt"

f1.write(name,":")

f = open(name, 'r', encoding='utf8') # 打开文件，并读取要处理的大段文字

f1.write()

txt1 = f.read()

txt1 = txt1.replace('\n', '') # 删掉换行符

txt1 = txt1.replace('，', '') # 删掉逗号

txt1 = txt1.replace('。', '') # 删掉句号

mylist = list(txt1)

mycount = collections.Counter(mylist)

for key, val in mycount.most\_common(10000): # 有序（返回前10个）

f1.write(key, val)

del mycount

del mylist

del txt1

附录三：

#!/usr/bin/python

# -\*- coding:utf-8 -\*-

import sys

reload(sys)

sys.setdefaultencoding('utf-8')

import jieba

import jieba.analyse

import xlwt #写入Excel表的库

if \_\_name\_\_=="\_\_main\_\_":

wbk = xlwt.Workbook(encoding = 'ascii')

sheet = wbk.add\_sheet("wordCount")#Excel单元格名字

word\_lst = []

key\_list=[]

for line in open('T2.txt'):#1.txt是需要分词统计的文档

item = line.strip('\n\r').split('\t') #制表格切分

# print item

tags = jieba.analyse.extract\_tags(item[0]) #jieba分词

for t in tags:

word\_lst.append(t)

word\_dict= {}

with open("wordCount.txt",'w') as wf2: #打开文件

for item in word\_lst:

if item not in word\_dict: #统计数量

word\_dict[item] = 1

else:

word\_dict[item] += 1

orderList=list(word\_dict.values())

orderList.sort(reverse=True)

# print orderList

for i in range(len(orderList)):

for key in word\_dict:

if word\_dict[key]==orderList[i]:

wf2.write(key+' '+str(word\_dict[key])+'\n') #写入txt文档

key\_list.append(key)

word\_dict[key]=0

for i in range(len(key\_list)):

sheet.write(i, 1, label = orderList[i])

sheet.write(i, 0, label = key\_list[i])

wbk.save('wordCount.xls') #保存为 wordCount.xls文件