老友记剧本文本挖掘文本报告

数据挖掘文字报告

工业工程 52 2015010846　唐静雯

摘要：老友记作为经久不衰的经典美剧，上映以来就一直处于美剧口碑榜顶端，这与这部剧的精彩的剧本是有很大关系的。本次数据挖掘任务是想通过对老友记十季的剧本进行文本挖掘，从数据当中对该剧得到进一步的了解。

前言：由于十季剧本数据量并不是很大，因此不是大数据挖掘，只是进行了文本数据挖掘并作出相应的分析。

研发现状：在网上并未搜到有关的前人的研究或者是论文。

目录

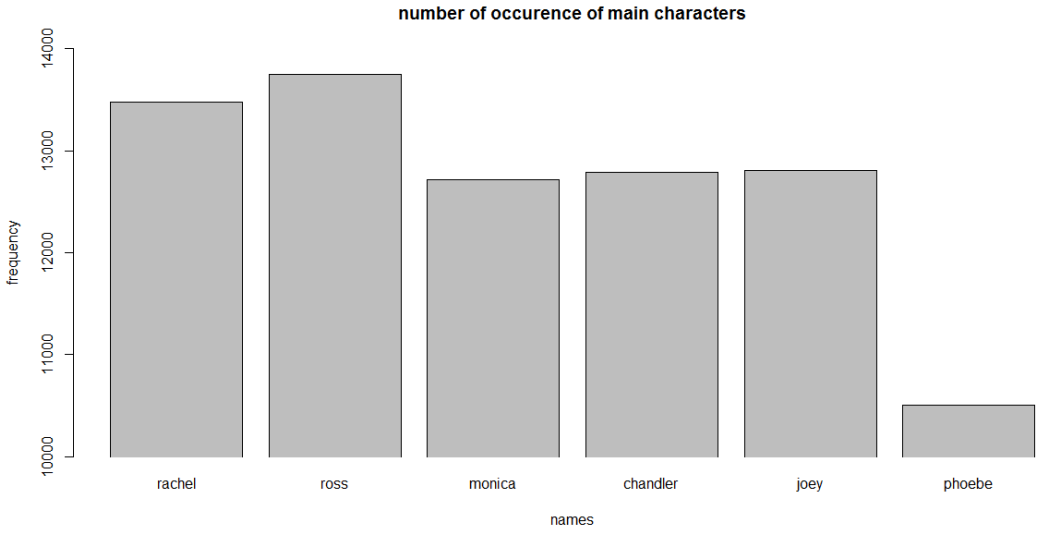
1. 设计和应用实现：

数据介绍

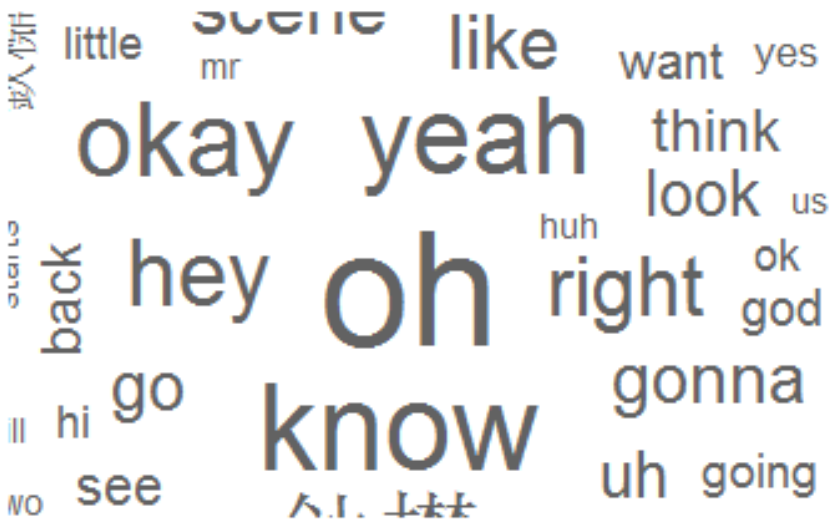
十季剧本
获取方法：网上下载
分析方法：对于每一集（有剧本的 225 集）建立词频矩阵，每一集的剧本作为一个文档。最

终得到一个 19141*255 的词项文档矩阵，下面的分析基本是基于这个矩阵。
十季每一集的 IMDB 评分
获取方法：网络爬虫从 IMDB 上面爬取。
分析方法：和上面的 255 个文档的词频关系建立联系进行分析。

初步分析



纵坐标是每个人物在十季的剧本中出现的总次数。可以看出 Ross 是当之无愧的主角，一共出现了近 14000 次，其次是搭档 Rachel，一万三千多次，二者（“美国甜心”）要比其余四位主演出现的次数多很多。
最少的是 Phoebe，只一共出现了一万一千次不到。但是 Phoebe 对于整个老友记的影响是十分关键的（后面的分析会解释。）



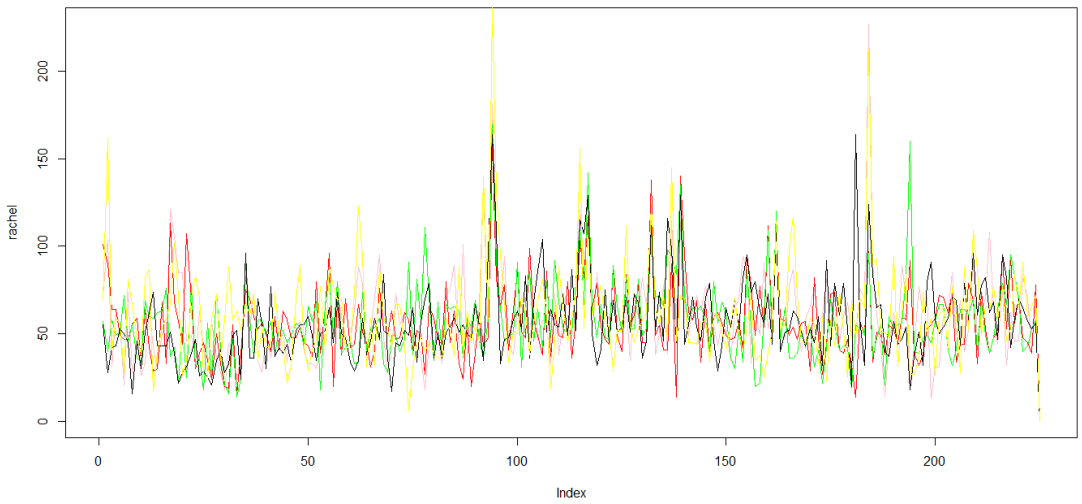
在去除了六大主要人物的名字之后，做出的词云如图所示。出现最多的是感叹词，和本身情

景喜剧的身份比较符合。

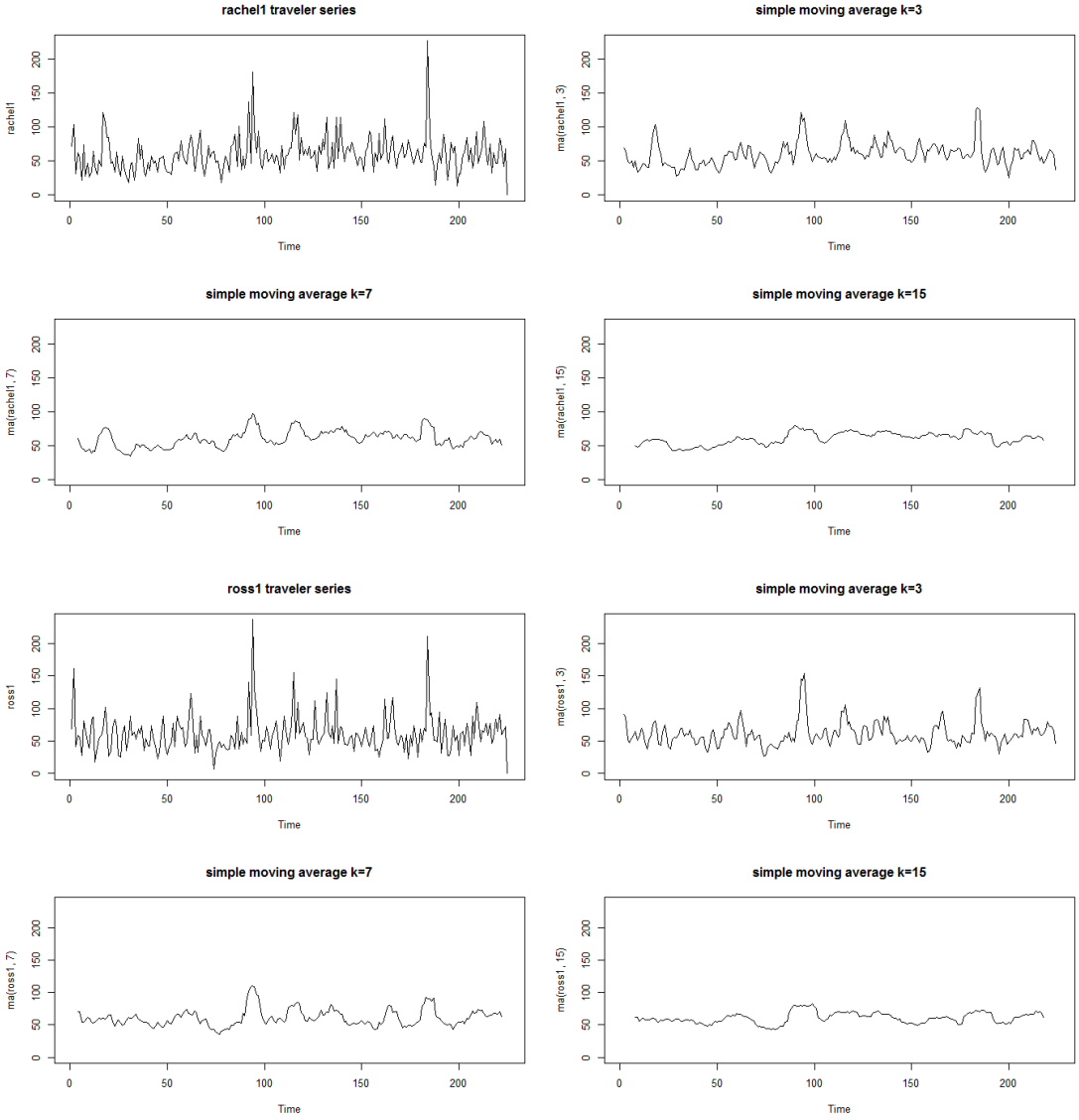
```
> findAssocs(myTdm,"rachel",0.6)
$rachel
  oh ross
0.66 0.64

> findAssocs(myTdm,"ross",0.6)
$ross
rachel
  0.64

> findAssocs(myTdm,"phoebe",0.5)
$phoebe
okay
0.53

> findAssocs(myTdm,"joey",0.5)
$joey
 hey
0.53
```

```
> findAssocs(myTdm,"monica",0.5)
$monica
chandler
     0.6

> findAssocs(myTdm,"chandler",0.5)
$chandler
monica   back
   0.6    0.5
```

首先来看一下两位绝对主演的关系。
数据也证明了二者的紧密联系：即在所有的出现的词中，对于 Rachel 来说，相关系数最高的是“Ross”和一个感叹词“Oh”。除去 Rachel 个人的用语习惯，发现有 Rachel 的地方，有 64%的可能性会有 Ross。
对于 Ross 来说，与之联系最为密切的词就是 Rachel。☺
接下来看一下 Monica 和 chandler。同样的是二者互为关系最为密切的项。但是 chandler 有趣的一点是，他和“back”这个词的相关系数也很高。跟 Rachel 爱说“oh”的情况不一样，chandler 这个现象很容易让人联想到他懦弱没有主张的性格。
至于 Phoebe 最为紧密的词是“okay”，也是 Phoebe 的最为明显的口头禅。而 Joey 更是以一句阴阳怪气的“Hey, how are you doing”闻名。
很巧合的是，二者的相关系数竟然是一样的，可以想到编剧们在操刀的时候是将这二位主演放在一个平等的构思的角度，用同样的力度，通过口头禅的角度来塑造他们的形象的。



从人物出现次数图可以看出，人物出场次数基本是平稳的。



从上面两张（时序差分图）可以看出，以 Rachel 和 Ross 的出现次数为例，是平稳的，即每一集的人物重心并没有产生很大的变化。对于其他人结果也是类似的。

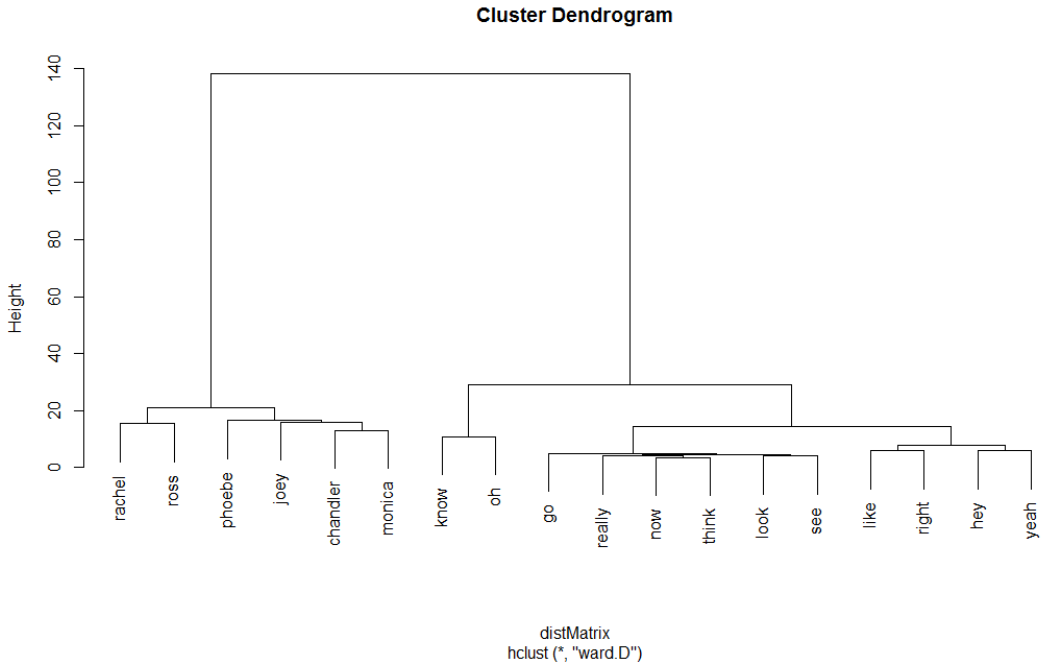
人物关系

```
> cor(t(maincha))
            rachel      ross    monica  chandler      joey    phoebe
rachel   1.0000000 0.6426702 0.4980909 0.2973018 0.4280749 0.3992156
ross     0.6426702 1.0000000 0.3930893 0.2849654 0.3658235 0.3230413
monica   0.4980909 0.3930893 1.0000000 0.6001832 0.3563623 0.4187551
chandler 0.2973018 0.2849654 0.6001832 1.0000000 0.4752550 0.4472833
joey     0.4280749 0.3658235 0.3563623 0.4752550 1.0000000 0.4464116
phoebe   0.3992156 0.3230413 0.4187551 0.4472833 0.4464116 1.0000000
```

上面是六位主要人物之间的相关系数图。是基于他们在每一集中的出现次数来计算的相关系数。

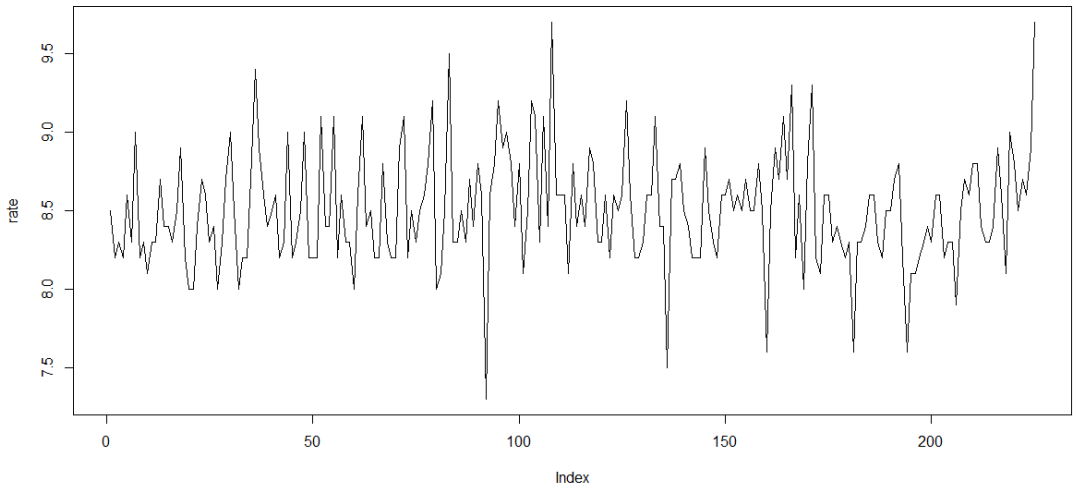
从中可以看出，六人之中，Rachel 和 Ross 的关系最为紧密，达到了 0.64，其次是 Monica 和

chandler，继两对情侣之后，再其次是 Monica 和 Rachel 这对闺蜜，chandler 和 Joey 这对好基友，然后就是 Phoebe 和 chandler 了。
值得注意的是，Ross 在与其他五人的关系上除了 Rachel 很高，其余都较低（低于 0.4）。其中最为疏远 的一对是 Ross 和 chandler。回想起来他们两人之间似乎并无直接情节相连，唯有后面 chandler 和 Monica 在一起之后因为 Monica 是 Ross 的妹妹因此会有一点交集。



词项聚类十分精确地体现了人物之间的关系：
最左边，六位好朋友划分成了一类：RR 和 MC 分别占一支，Joey 和 Phoebe 十分正确地没有分为一类（因为不是情侣亲密关系，而是通过 chandler 和 Joey 联系起来，Phoebe 呈现出通过 Joey 和其余人联系起来。）
之后就是 know 和 oh 划为一支（因为最常用的口语：Iknow 和 oh，其中 I 作为停词表中的单词被忽略。）
之后就是和内容相关的一些单词，常用的比如 go look see think 等等。
最右侧的一支十分有趣，是语气环境十分欢乐的词成了一类：like　right　hey yeah.

人物对评分的影响



最低分：第四季 21 集 7.3 分
"421 The One With The Invitation“
讲述了 Ross 和 Julie 秀恩爱而不是 Rachel（很可能是 RR 党打了巨低分）
最高分：最后一集 9.7 分
“1018 The Last One”
最后一集。

```
> cor(t(r))
              rate     rachel      ross    monica   chandler      joey    phoebe
rate      1.00000000 0.05474488 0.1627399 0.1599269 0.07735386 0.1087462 0.1184954
rachel    0.05474488 1.00000000 0.6426702 0.4980909 0.29730179 0.4280749 0.3992156
ross      0.16273993 0.64267020 1.0000000 0.3930893 0.28496543 0.3658235 0.3230413
monica    0.15992694 0.49809085 0.3930893 1.0000000 0.60018316 0.3563623 0.4187551
chandler  0.07735386 0.29730179 0.2849654 0.6001832 1.00000000 0.4752550 0.4472833
joey      0.10874624 0.42807485 0.3658235 0.3563623 0.47525499 1.0000000 0.4464116
phoebe    0.11849541 0.39921564 0.3230413 0.4187551 0.44728329 0.4464116 1.0000000
```

上面是每一集的评分和每一集人物出现次数的相关系数矩阵。从表中可以看出，对分数产证最多正面影响的人物排序分别是：Ross， Monica ，Phoebe， Joey ，chandler 和 Rachel。结果比较出乎意料，因为老友记中最受欢迎的人气王“Rachel”竟然对评分的贡献在六人中是最小的。
下面进行回归分析，找出比较严谨的真的影响分数的主要因素。

```
Call:
lm(formula = rate ~ rachel + ross + monica + chandler + joey +
    phoebe, data = as.data.frame(t(r)))

Residuals:
     Min       1Q   Median       3Q      Max
-1.12206 -0.23151 -0.02707  0.17658  1.38552

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.314480   0.079887 104.078   <2e-16 ***
rachel      -0.002549   0.001298  -1.963   0.0509 .
ross         0.002342   0.001064   2.201   0.0288 *
monica       0.002870   0.001446   1.985   0.0484 *
chandler    -0.001405   0.001368  -1.027   0.3055
joey         0.001028   0.001227   0.838   0.4031
phoebe       0.001183   0.001424   0.831   0.4072
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3515 on 218 degrees of freedom
Multiple R-squared:  0.05804,   Adjusted R-squared:  0.03211
F-statistic: 2.239 on 6 and 218 DF,  p-value: 0.04067
```

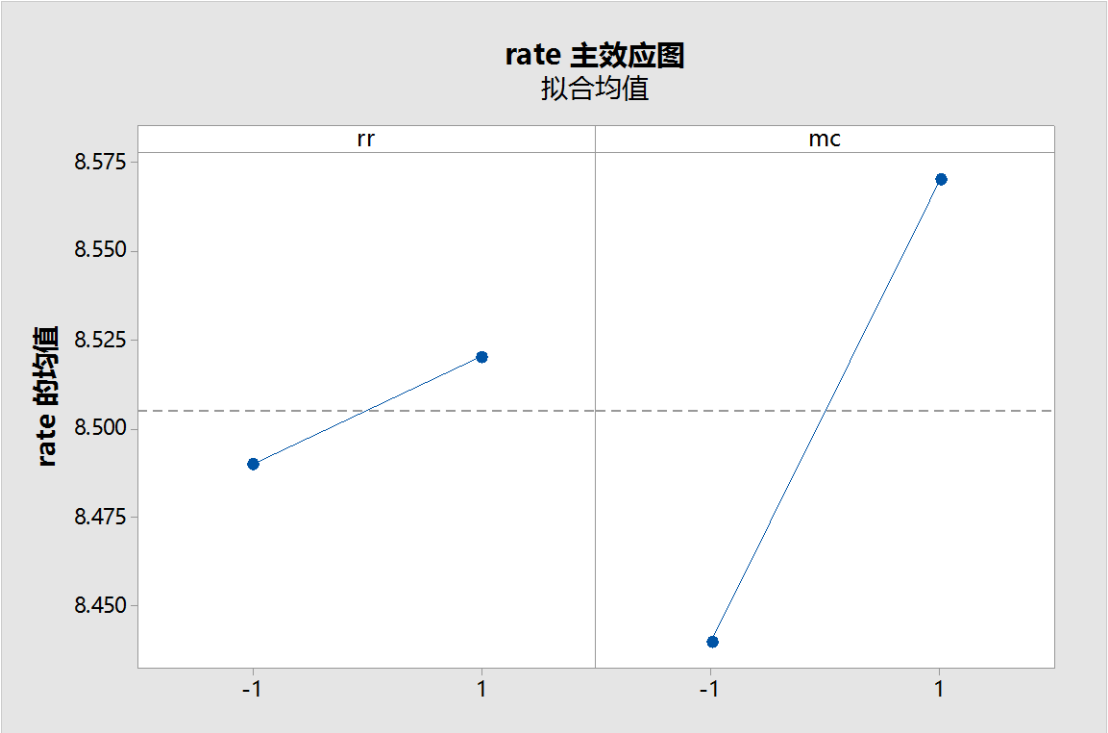
首先建立一个只有一阶项的模型，发现和上面的相关系数矩阵结果一致。
最终显著的是 Ross 和 Monica 两兄妹。
在此分析中，Rachel 和 chandler 甚至对于评分有着负面的影响。

```
> step.f$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
rate ~ 1

Final Model:
rate ~ ross + monica + rachel


     Step Df  Deviance Resid. Df Resid. Dev       AIC
1                           224   28.59982 -462.1075
2   + ross  1 0.7574458       223   27.84238 -466.1468
3 + monica  1 0.3114586       222   27.53092 -466.6779
4 + rachel  1 0.3348238       221   27.19609 -467.4311
~
```

```
Call:
lm(formula = rate ~ ross + monica + rachel, data = as.data.frame(t(r)))

Residuals:
     Min       1Q   Median       3Q      Max
-1.19951 -0.22201 -0.04417  0.17594  1.36453

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.335469   0.070795 117.740   <2e-16 ***
ross         0.002419   0.001053   2.296   0.0226 *
monica       0.002492   0.001206   2.066   0.0400 *
rachel      -0.002056   0.001247  -1.649   0.1005
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3508 on 221 degrees of freedom
Multiple R-squared:  0.04908,   Adjusted R-squared:  0.03617
F-statistic: 3.802 on 3 and 221 DF,  p-value: 0.01096
```

采用逐步遍历模型的方法，最终确定了最佳拟合模型是：
Rate = 8.33 + 0.002419*Ross + 0.002492*Monica – 0.002056*Rachel
系数都可以看作是显著的。
因此已知了一集中的 Ross， Monica 和 Rachel 的出现次数，基本就可以为这一集老友记的 IMDB 的评分大致上做出一个准确的预估☺

人物关系变化及对评分的影响



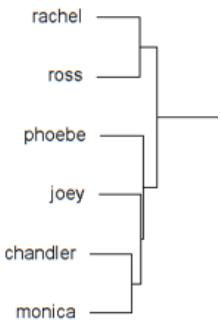
根据 RR（Rachel 和 Ross）和 MC（Monica 和 chandler）的关系随季数的变化。
因此选取第 1,3,7,10 季做了因子分析，得到了左边的因子效应图。发现 Monica 和 chandler 在一起的剧情对整个季的评分有着显著提高，而 Ross 和 Rachel 的感情似乎没有什么影响（相比之下比较小）。

```
> rrr
          rr
rr 1.0000000 0.2198628
   0.2198628 1.0000000
> mcr
          mc
mc 1.0000000 0.8350492
   0.8350492 1.0000000
```

上面是二者的相关系数分析。发现 MC 对评分的影响要远大于 RR 对。可能是因为 RR 对的争议较大，两位主人公的性格相比较于 Monica 和 chandler 更为复杂，经历地情节也更多。因此观众对于这份感情的看法也会更多的不一致。

2. 总结：

1. 人物关系：



Ross 和 Rachel 是绝对主角，
Monica 和 chandler 次之，
Phoebe 和 Joey 次之。
两对情侣，以及一对好基友，还有和众人都很玩得来的 Phoebe 构成了六人行。

2. IMDB 上的评分 Rate = 8.33 + 0.002419*Ross + 0.002492*Monica – 0.002056*Rachel（其中人名代表在该集剧本中出现的次数）
3.Monica 和 chandler 的感情对于评分比 Rachel 和 Ross 的作用大得多。