测试：《破冰行动》文本分析

张笑竹 / 201618070114

2019年10月20日

今年5月，凭借着在垂直题材上的直接经验，爱奇艺深度参与制作的《破冰行动》，以网台联动、央视播出的模式，在内容热度、收视率和口碑上连连走高，成为5月霸屏神剧，《破冰行动》被网友比喻为缉毒版《人民的名义》。本案例试图通过文本分析的方法，分析改编自网络剧《破冰行动》的同名小说中的人物关系，及其人物塑造，和复杂人物关系中的故事线，探究其引爆口碑、获得高收视的主要原因，为未来这类题材的类型剧创作提供参考。

0 准备工作

首先，清空工作环境，设置工作目录。

然后，加载需要的R包。

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
## Loading required package: jiebaRD
```

```
##
## Attaching package: 'igraph'
```

```
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
```

```
## The following object is masked from 'package:base':
##
##     union
```

1 任务一：数据预处理

1.1 空行、标题、空格的处理

首先，导入小说文本。

```
pobing <- readLines("破冰行动.txt")
```

然后，逐段判别，删去段落中的空行。

```
## 删去段落中的空行
pobing1 <- c()
for(i in 1:length(pobing))
{
  if(pobing[i]!="")
  {
    pobing1 <- c(pobing1,pobing[i])
  }
}
```

再次，通过grep()函数，删去段落中的标题行。注意到，每一个标题都带有“正文”二字。因此，将grep()函数的关键字设置为“正文”。

```
## 删去段落中的标题行
title_index=grep("正文",pobing1)
pobing1 <- pobing1[-title_index]
pobing1 <- pobing1[-1]
```

最后，删去每个段落开始的空格。利用str_trim()函数逐段处理。

```
## 删去段落中的空格
for(i in 1:length(pobing1))
{
  pobing1[i]=str_trim(pobing1[i],'left')
}
```

1.2 预览

下面，展示处理后数据的前6条。

```
head(pobing1)
```

```
## [1] "今年天气反常，还没到六月呢，就热得天怒人怨。在太阳地上站一会儿，就被烤得脸皮都疼。"
## [2] "大中午，路边没什么阴凉地儿，李飞两道剑眉拧得死紧，大步跨进东山市公安局禁毒大队的办公楼，攥着手机的手背青筋暴起，正在给始终“号码无法接通”的搭档拨打第四次电话。"
## [3] "他从热浪糊脸的街道钻进冷气十足的办公楼，满脑门的汗，自己倒是感觉不到热，只觉得那头汗是急出来的。"
## [4] "好在这一次，电话通了。"
## [5] "“宋杨？你在哪儿，为什么不接电话？你前女友来找我，说不知道谁P了张她和别人的不雅照，把你给引走了，你现在是什么情况！”前几天扫毒行动刚扫了个竹篮打水，嫌疑人被保回家，转头就吊死在自家房梁上，以此向他们整个禁毒大队鸣冤示威。这么个节骨眼，有人把宋杨引走了，他的电话还死活打不通，李飞害怕当天跟着自己一起去“探底”的宋杨出什么事儿，差点就要跟上面打报告查定位了。"
## [6] "电话通了，李飞把上楼的脚步又收了回来。电话信号不好，宋杨的声音断断续续，依然精神抖擞皮得很，“跟你说了陈珂不是我‘前’女友，我会把她再追回来的！—我找到照片里的‘男主角’了，看着就不是好鸟，没想到一逼问，拔出萝卜带出泥的，还有大发现！”"
```

2 任务二：小说主要人物词频统计

2.1 设置分词器

首先，利用提供的词典、停用词，对分词器的参数进行设定。

```
## 导入词典
dictpath <- "./pbxd_user.txt"
## 导入停用词词库
stoppath <- "./stopword.txt"
## 设置分词器
cutter <- worker(bylines = TRUE, user = dictpath, stop_word=stoppath,"tag")
```

然后，对文本进行分词，并转换为合适的格式。

```
## 进行分词
pobing_fenci <- cutter[pobing1]
## 将pobing_fenci从list转换为matrix格式
pobing_fenci <- lapply(pobing_fenci,as.matrix)
```

2.2 统计人物出现频率

首先，写入小说中出现的主要任务姓名。注意到，在不同语境下，对同一个人的称呼有可能不同。因此，在提取人物信息时，一定要做到不重不漏。

```
roles <- c("(李维民)|(维民)","(马云波)|(云波)","(李飞)","(陈文泽)|(文泽)","(赵嘉良)|(嘉良)",
           "(马雯)","(蔡永强)|(永强)","(陈光荣)|(光荣)","(林耀东)|(耀东)","(林水伯)|(水伯)",
           "(陈珂)","(宋杨)","(蔡杰)","(蔡军)","(林宗辉)|(宗辉)","(林耀华)|(耀华)",
           "(林胜武)|(胜武)","(林胜文)|(胜文)","(周琳林)","(方天逸)|(天逸)","(陆童)","(何瑞龙)|(瑞龙)",
           "(苏建国)|(建国)","(苏康)","(左兰)","(陈岩)","(王志雄)|(志雄)")
```

然后，构建bool逻辑矩阵，表示人物出现的段落。矩阵的每一行代表一段，每一列代表每一个人物角色。具体而言，

$$x_{ij} = \begin{cases} 0, & \text{人物}j\text{在段落}i\text{中未出现} \\ 1, & \text{人物}j\text{在段落}i\text{中出现} \end{cases}$$

```
## 找到人物出现的段落
role_para <- sapply(roles, grepl, pobing1)
## role_para列名的更改
role_name = c("李维民","马云波","李飞","陈文泽","赵嘉良",
              "马雯","蔡永强","陈光荣","林耀东","林水伯",
              "陈珂","宋杨","蔡杰","蔡军","林宗辉","林耀华",
              "林胜武","林胜文","周琳林","方天逸","陆童","何瑞龙",
              "苏建国","苏康","左兰","陈岩","王志雄")
colnames(role_para) <- role_name
```

随后，构建数据框，统计人物出现的频率。

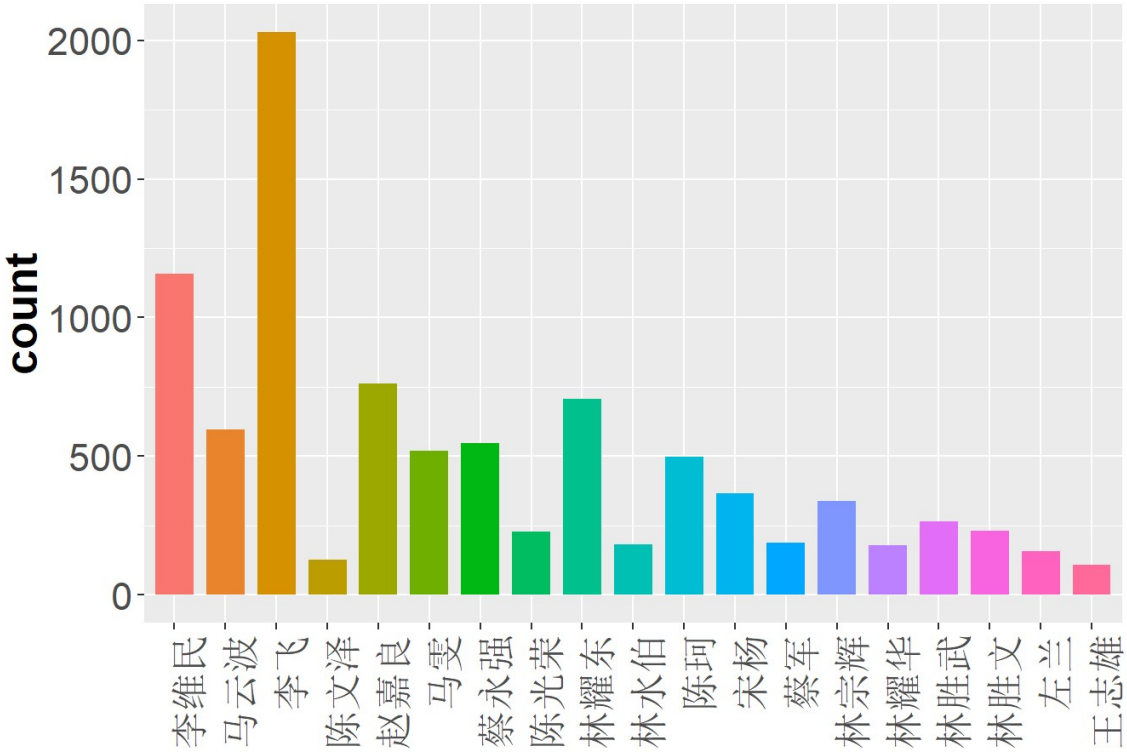
```
## 构建数据框，最后一列统计人物在全书出现的次数
role_count = data.frame(role = factor(colnames(role_para),
                                      levels = role_name),
                        count = colSums(role_para))
head(role_count)
```

```
##          role count
## 李维民 李维民  1159
## 马云波 马云波   596
## 李飞     李飞  2029
## 陈文泽 陈文泽   127
## 赵嘉良 赵嘉良   763
## 马雯     马雯   521
```

2.3 词频柱状图

最后，对出现频率不小于100的人物，绘制词频柱状图。

```
role_count_100 <- role_count[which(role_count$count>=100),]

ggplot(role_count_100, aes(x = role, y = count, fill = role)) +
  geom_bar(stat = "identity", width = 0.75) +
  xlab("") +
  theme(axis.text=element_text(size=17),
        axis.title=element_text(size=20,face="bold"),
        axis.text.x = element_text(angle = 90, hjust = 1),
        legend.position="none")
```



观察上图，不难发现，李飞是小说的主人公，而李维民、马云波、赵嘉良、林耀东也是高频人物。因此对于这部小说，分析与上述5人有关的情节发展、心理活动，是十分关键的。

3 任务三：人物关系网络分析

3.1 构造“节点-节点-权重”数据结构

利用上面得到的role_para矩阵，可以构造网络中的“节点-节点-权重”数据结构。

当同一段中出现2个人名时，毋庸置疑，这两个主人公连接，构成1个边（edge），即仅存在1种“节点-节点”的组合。

当同一段中出现3个人名时，根据排列数$C(3,2)=3$，需要分别构造3种“节点-节点”的组合。

当同一段中出现出现4个人名时，根据排列数$C(4,2)=6$，需要分别构造4种“节点-节点”的组合。

...

而当同一段中出现出现n个人名时，根据排列数$C(n,2)=\frac{n\cdot(n-1)}{2}$，需要分别构造$\frac{n\cdot(n-1)}{2}$种“节点-节点”的组合。

事实上，对于绝大多数段落，人名个数都在3个以及3个之内。因此，为方便起见，这里只考虑$n \leq 3$的情况。

```
NET=c()
for(i in 1: nrow(role_para))
{
  rr=c()
  for(j in 1:ncol(role_para))
  {
    if(role_para[i,j]==TRUE)
    {
      rr=c(rr,role_name[j])
    }
  }
  k=length(rr)
  if(k==2)
  {
    NET=rbind(NET,rr)
  }
  if(k==3)
  {
    NET=rbind(NET,c(rr[1],rr[2]))
    NET=rbind(NET,c(rr[1],rr[3]))
    NET=rbind(NET,c(rr[2],rr[3]))
  }
}
colnames(NET) <- c("from", "to")
```

然后，对“节点-节点-权重”的组合进行频率统计。

```
## 进行频率统计
NET.freq = unique(NET)
freq=rep(0,nrow(NET.freq))
for(i in 1:nrow(NET))
{
  for(j in 1:nrow(NET.freq))
  {
    if(NET[i,1]==NET.freq[j,1] && NET[i,2]==NET.freq[j,2])
    freq[j]=freq[j]+1
  }
}
NET.freq = cbind(NET.freq,freq)
head(NET.freq)
```

```
##    from     to       freq
## rr "李飞"   "宋杨"   "164"
##    "李飞"   "陈珂"   "144"
##    "陈珂"   "宋杨"   "56"
##    "李飞"   "蔡永强" "165"
##    "蔡永强" "宋杨"   "16"
## rr "李飞"   "林水伯" "67"
```

上面输出了“节点-节点-权重”数据结构的前6行。

```
## 人物（节点）数量
length(role_name)
```

```
## [1] 27
```

```
## 组合数量
nrow(NET.freq)
```

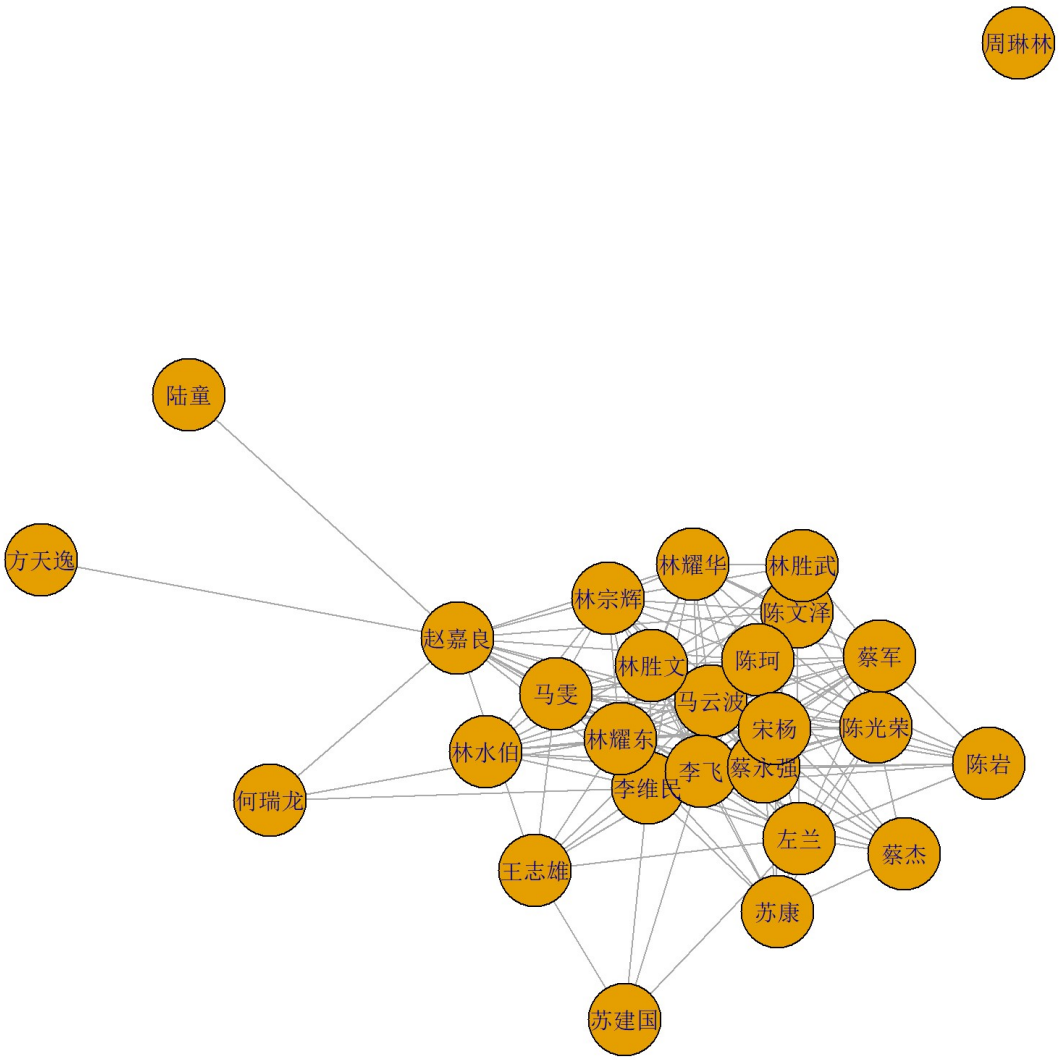
```
## [1] 170
```

根据统计结果，在27个小说人物中，共记录了170对人物的相互关系。

3.2 绘制网络关系图

最后，画出人物网络关系图。

```
g <- graph_from_data_frame(NET.freq, directed=FALSE, vertices=role_name)
plot(g)
```



可以看出，人物关系是围绕着李飞、李维民、林耀东、马云波等几人为中心展开的。而周琳林则几乎是“孤立点”，可被视为小说的“龙套”人物。

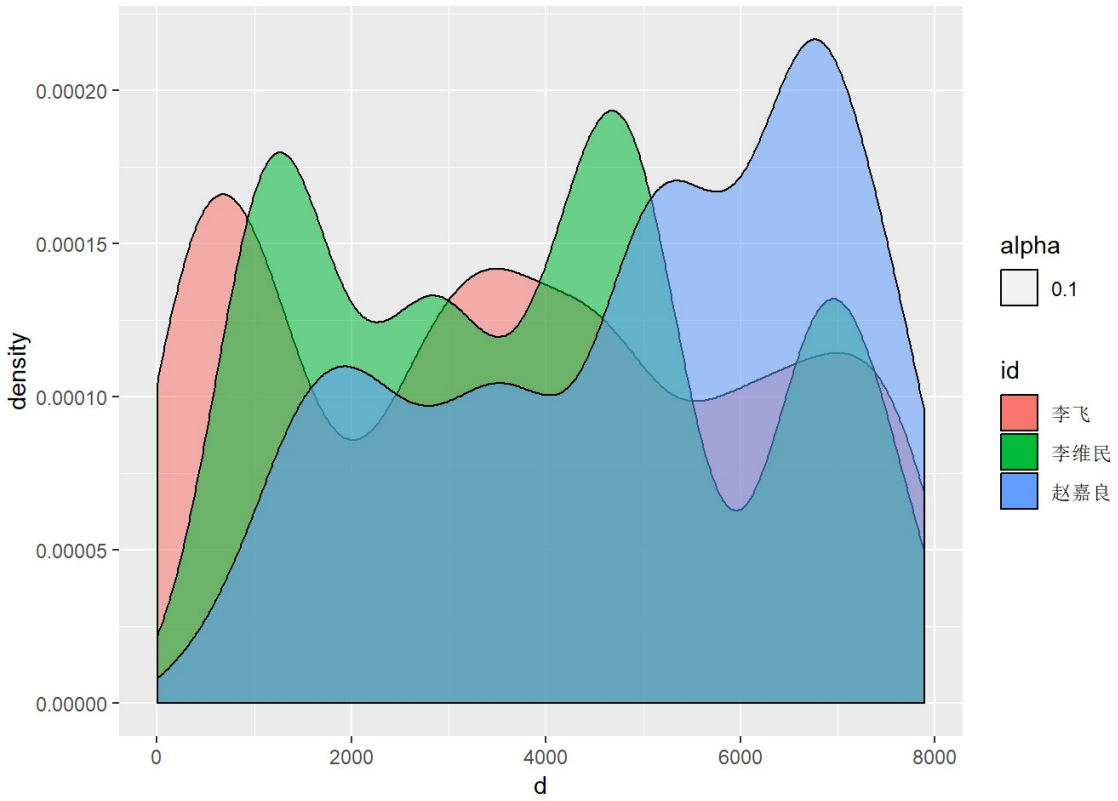
4 任务四：李飞、李维民、赵嘉的出镜密度

仍然根据之前得到的role_para矩阵，分别提取李飞、李维民、赵嘉的出镜段落。

```
## 找出对应角色出现的段落序号（role_para即可）
para <- 1:nrow(role_para)

para.lifei <- para[role_para[,3]]
n1 <- length(para.lifei)

para.liweimin <- para[role_para[,1]]
n2 <- length(para.liweimin)

para.zhaojialiang <- para[role_para[,5]]
n3 <- length(para.zhaojialiang)

d <- c(para.lifei, para.liweimin, para.zhaojialiang)
id <- c(rep("李飞",n1),rep("李维民",n2),rep("赵嘉良",n3))
id <- as.factor(id)

dd <- data.frame(d,id)
```

然后，绘制三人的出镜曲线密度图。

```
## 绘制密度曲线图
ggplot(dd, aes(x = d, fill = id,alpha = 0.1)) +
  geom_density()
```



可以看出，李飞的出镜密度略称下降趋势；李维民在小说中部的出镜频次较高；而关于赵嘉良的情节则主要集中于后

半段。

5 任务五：林耀东、林宗辉共同出现段落词频分析 5.1 提取段落中的词汇

首先，对二人（林耀东、林宗辉）共同出现的段落进行分词提取。这里将“共同出现段落”的定义放宽，将其延伸至“在每50段中同时出现即视为1次同时出现”；因此，应提取该50段全部文本进行分析。

通过每次50段的循环，将林耀东和林宗辉共同出现时的词全部存入Ciyun向量。

```
Ciyun=c()
j1=1
j2=50
while(j2<=nrow(role_para))
{
  if(sum(role_para[j1:j2,'林耀东'])!=0 && sum(role_para[j1:j2,'林宗辉'])!=0)
  {
    for(k in j1:j2)
    {
      Ciyun=c(Ciyun,as.vector(pobing_fenci[[k]]))
    }
  }
  j1=j1+50
  j2=j2+50
}
```

5.2 词频的统计和调整

下面，对Ciyun向量进行频率统计，并按降序排列。

```
freq = as.data.frame(table(Ciyun),stringsAsFactors = F)
freq = freq[order(-freq[,2]),]
```

然后，考虑到一些语气词，如“么”、“来”、“们”等对分析并无实际意义，因此将它们去除。

```
freq <- subset(freq,str_detect(freq$Ciyun,"一|么|有|以|来|去|这|那|我|们|不|是|在|的|开|关|为|然|几乎|最|里|外|上|下|了|只")==F)
freq <- subset(freq,str_detect(freq$Ciyun,"己|已|着|个|道|刘|完|看|于|事|见|马|出|进|能|觉|定|今|天|之|前|后|时|如|说|声|眼睛|应|目|大|瞬间|刚|希望|就|连忙|到")==F)
```

此外，还应注意到，上面给出的例子远远不能涵盖文本内所有的“无意义”单字。因此，索性将所有“字符串长度”小于2的单字全部去除。

```
freq <- freq[nchar(freq$Ciyun)>=2,]
head(freq,30)
```

```
##        Ciyun Freq
## 4718     李飞  546
## 5007  林宗辉  326
## 5004  林耀东  285
## 9821  赵嘉良  207
## 4751  李维民  205
## 2015     东山  161
## 1940     电话  150
## 1048     陈珂  148
## 7074     手机  146
## 4994  林胜武  145
## 792   蔡小玲  135
## 4957     林灿  119
## 801   蔡永强  113
## 5005  林耀华  105
## 783      蔡军   99
## 7489     塔寨   90
## 4984     林兰   68
## 2711     告诉   67
## 6558     三宝   62
## 3071     行动   60
## 10024    证据   58
## 4470     口气   51
## 2236     儿子   50
## 4063     禁毒   49
## 10093    直接   47
## 4106     警察   45
## 1930     点头   44
## 5922     片刻   44
## 8926     摇头   42
## 8131     问题   41
```

最终得到了上述的词频统计。

5.3 词云图

下面，画出词云图。

```
wordcloud2(freq[1:200,],size = 1, minRotation = -pi/3, maxRotation = pi/3,
           rotateRatio = 0.8,fontFamily = "微软雅黑",
           color = "random-light")
```

词云图中的最显眼词汇是各个主人公的名字，从而为我们提供了一场主要人物之间错综复杂、勾心斗加的大戏。然而，除了主人公之间的人物关系，我们还希望能够通过词云获知一些关于情节的其他信息。为此，将前20个词去除，得到新的词云。

```
## 去掉人物
freq1 <- freq[20:200,]
wordcloud2(freq1, size = 0.4, minRotation = -pi/3, maxRotation = pi/3,
           rotateRatio = 0.8,fontFamily = "微软雅黑",
           color = "random-light")
```

根据新的词云，基本可以推测，《破冰行动》是一部有关缉毒、犯罪类型的作品，各种反派要素十分完备，具有相当强的吸引力。