统计软件 HW9

邵智轩　　1400012141　　物理学院

```
HLM.url <- "http://www.xiexingcun.net/hongloumeng/"
index.url <- paste(HLM.url, "index.html", sep = "")
chapters.url <- paste(HLM.url, c(formatC(1:9, width = 2, flag = 0), 10:120),
    ".htm", sep = "")


outfcon <- file("hongloumeng.txt", open = "wt")  # 输出的 txt 文件

chchar.substr <- function(c) {
    # 找出汉字子串
    find.Chinese <- gregexpr("[^\\x00-\\xff]|[,!?:;]", c, perl = TRUE)[[1]]
    substr(c, start = find.Chinese[1], stop = find.Chinese[length(find.Chinese)])
}

subpunc <- function(c) {
    # 替换标点符号
    c <- gsub("&quot;", "\"", c, fixed = T)
    c <- gsub("[ ;]?&lt;*[ ]?", "<", c, perl = T)
    c <- gsub("[ ;]?&gt;*[ ]?", ">", c, perl = T)
    c
}


# 利用 index 页制作 txt 的 Contents 页
makeContents <- function(index.url = index.url) {
```

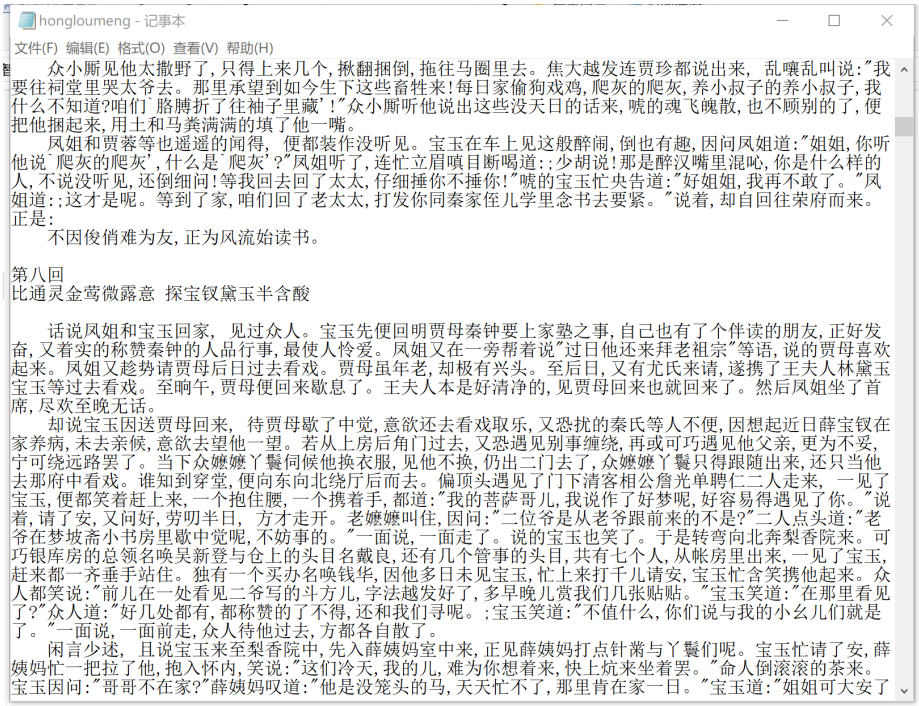
```
    index.text <- readLines(url(index.url))
    Contents <- grep("\\d+[.]htm", index.text, perl = TRUE, value = TRUE)
    Contents <- vapply(Contents, chchar.substr, "s")
    Contents <- c(" 红楼梦", " 清 曹雪芹 著", Contents)
    Contents
}

# 读取单章
addChapter <- function(chapter.url) {
    chapter <- readLines(url(chapter.url))
    # 找出含汉字的行
    chchar <- grep("[^\\x00-\\xff]", chapter, perl = TRUE, value = TRUE)
    # 去掉汉字行中不需要的行
    chchar <- chchar[-grep("<title>|htm| 作者: 曹雪芹", chchar, perl = TRUE)]
    # chchar<-vapply(chchar,chchar.substr,'s') 先处理标题第几回和题目
    chapnum <- chchar.substr(chchar[1])
    chapname <- chchar.substr(chchar[2])

    # 删掉回目，下面处理正文
    chchar <- chchar[-(1:2)]

    # 没有<td> 或<p> 或<br> 的段落连接到上一行
    newlines.index <- grep("    ", chchar, perl = TRUE)  # 重启一行的
    chchar <- vapply(chchar, chchar.substr, "s")

    chnew <- c()
    for (i in (1:length(chchar))) {
        if (i %in% newlines.index) {
            chnew <- c(chnew, chchar[i])
        } else {
            chnew[length(chnew)] <- paste(chnew[length(chnew)], chchar[i], sep = "")
        }
    }
```

```
    # 替换掉标点符号
    chnew <- subpunc(chnew)

    # 加上回目，以及回目和正文间的空行
    c("", chapnum, chapname, "", chnew)
}

time.start <- proc.time()

writeLines(makeContents(index.url), con = outfcon)
for (chapter.url in chapters.url) {
    writeLines(addChapter(chapter.url), con = outfcon)
}

proc.time() - time.start
```

```
##    user  system elapsed
##    9.53    2.78   18.07
```

```
close(outfcon)
```

效果图：


hongloumeng - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

众小厮见他太撒野了，只得上来几个，揪翻捆倒，拖往马圈里去。焦大越发连贾珍都说出来，乱嚷乱叫说："我要往祠堂里哭太爷去。那里承望到如今生下这些畜牲来！每日家偷狗戏鸡，爬灰的爬灰，养小叔子的养小叔子，我什么不知道？咱们`胳膊折了往袖子里藏'！"众小厮听他说出这些没天日的话来，唬的魂飞魄散，也不顾别的了，便把他捆起来，用土和马粪满满的填了他一嘴。

凤姐和贾蓉等也遥遥的闻得，便都装作没听见。宝玉在车上见这般醉闹，倒也有趣，因问凤姐道："姐姐，你听他说`爬灰的爬灰'，什么是`爬灰'？"凤姐听了，连忙立眉嗔目断喝道：；少胡说！那是醉汉嘴里混吣，你是什么样的人，不说没听见，还倒细问！等我回去回了太太，仔细捶你不捶你！"唬的宝玉忙央告道："好姐姐，我再不敢了。"凤姐道：；这才是呢。等到了家，咱们回了老太太，打发你同秦家侄儿学里念书去要紧。"说着，却自回往荣府而来。正是：

不因俊俏难为友，正为风流始读书。

第八回
比通灵金莺微露意　探宝钗黛玉半含酸

话说凤姐和宝玉回家，见过众人。宝玉先便回明贾母秦钟要上家塾之事，自己也有了个伴读的朋友，正好发奋，又着实的称赞秦钟的人品行事，最使人怜爱。凤姐又在一旁帮着说"过日他还来拜老祖宗"等语，说的贾母喜欢起来。凤姐又趁势请贾母后日过去看戏。贾母虽年老，却极有兴头。至后日，又有尤氏来请，遂携了王夫人林黛玉宝玉等过去看戏。至晌午，贾母便回来歇息了。王夫人本是好清净的，见贾母回来也就回来了。然后凤姐坐了首席，尽欢至晚无话。

却说宝玉因送贾母回来，待贾母歇了中觉，意欲还去看戏取乐，又恐扰的秦氏等人不便，因想起近日薛宝钗在家养病，未去亲候，意欲去望他一望。若从上房后角门过去，又恐遇见别事缠绕，再或可巧遇见他父亲，更为不妥，宁可绕远路罢了。当下众嬷嬷丫鬟伺候他换衣服，见他不换，仍出二门去了，众嬷嬷丫鬟只得跟随出来，还只当他去那府中看戏。谁知到穿堂，便向东向北绕厅后而去。偏顶头遇见了门下清客相公詹光单聘仁二人走来，一见了宝玉，便都笑着赶上来，一个抱住腰，一个携着手，都道："我的菩萨哥儿，我说作了好梦呢，好容易得遇见了你。"说着，请了安，又问好，劳叨半日，方才走开。老嬷嬷叫住，因问："二位爷是从老爷跟前来的不是？"二人点头道："老爷在梦坡斋小书房里歇中觉呢，不妨事的。"一面说，一面走了。说的宝玉也笑了。于是转弯向北奔梨香院来。可巧银库房的总领名唤吴新登与仓上的头目名戴良，还有几个管事的头目，共有七个人，从帐房里出来，一见了宝玉，赶来都一齐垂手站住。独有一个买办名唤钱华，因他多日未见宝玉，忙上来打千儿请安，宝玉忙含笑携他起来。众人都笑说："前儿在一处看见二爷写的斗方儿，字法越发好了，多早晚儿赏我们几张贴贴。"宝玉笑道："在那里看见了？"众人道："好几处都有，都称赞的了不得，还和我们寻呢。；宝玉笑道："不值什么，你们说与我的小幺儿们就是了。"一面说，一面前走，众人待他过去，方都各自散了。

闲言少述，且说宝玉来至梨香院中，先入薛姨妈室中来，正见薛姨妈打点针黹与丫鬟们呢。宝玉忙请了安，薛姨妈忙一把拉了他，抱入怀内，笑说："这们冷天，我的儿，难为你想着来，快上炕来坐着罢。"命人倒滚滚的茶来。宝玉因问："哥哥不在家？"薛姨妈叹道："他是没笼头的马，天天忙不了，那里肯在家一日。"宝玉道："姐姐可大安了