PS3

Hangyu Huang

26 September 2017

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document.

```
getwd()
```

```
## [1] "C:/Users/crypress/Desktop/243/ps3"
```

```
setwd("C:/Users/crypress/Desktop/243/ps3")
url <- "http://www.gutenberg.org/cache/epub/100/pg100.txt"
download.file(url, "shakespeare.txt")
```

```
library(stringr)
## loading required package: method

spdata <- readLines("shakespeare.txt")
spdata <- spdata[spdata != ""]
## find the location of each play start with year and end with "THE END"

location <- data.frame(beg=which(grepl("^[[:digit:]]{4}",spdata)), end=which(grepl("THE END",spdata)))
## extract the text from the beginning loaction to the end location

complworks <- lapply(seq_along(location$beg), function(i){
  spdata[seq(from=location$beg[i],
                  to = location$end[i],
                  by=1)]
  })


## remove the frist and the last sonnet
playsonly <- complworks[-c(1, 38)]
```

```
# body of each play
begin1 <- data.frame(begin=grep(regex("SCENE:"), spdata))
begin2 <- data.frame(begin=grep(regex("Scene:"), spdata))
begin3 <- data.frame(begin=grep(regex("SCENE.-"), spdata))
begin <- rbind.data.frame(begin1,begin2,begin3)
begin <- sort(begin$begin)
end <- data.frame(end=grep("THE END", spdata))
end <- end[-c(1, 38),]
bodyloc <- cbind.data.frame(begin,end)
body <- sapply(seq_along(bodyloc$begin), FUN = function(i) {
  spdata[seq(from=bodyloc$begin[i],
                  to = bodyloc$end[i],
                  by=1)]
```

```
})

# year of plays
year <- data.frame(year = sapply(playsonly, '[[', 1))

#name of plays
name <- data.frame(name = sapply(playsonly, '[[', 2))

#number of scene for each play
numofscene <- str_count(body, "SCENE | Scene") - 1
numofscene <- data.frame(numofscene = numofscene)

# number of ACT for each play
numofACT1<- matrix(str_count(body, fixed("SCENE I.")),nrow =36,ncol =1)
numofACT2<- matrix(str_count(body, fixed("SCENE 1")), nrow =36,ncol =1)
numofACT3<- matrix(str_count(body, fixed("Scene I.")), nrow =36,ncol =1)
numofACT <- numofACT1 + numofACT2 + numofACT3
numofACT <- data.frame(numofACT = numofACT)

#metadata of Shake Sperea's plays

metadata <- cbind(name,year,numofACT, numofscene)
```

```
#functin to find the chunks in one play
searchchunks <- function(x){
play1 <- unlist(body[x])
chunkbeg <- matrix(grep("  [A-Z]{4,}. ", play1), ncol=1)
chunkend <- matrix(chunkbeg[-1]-1,nrow = (nrow(chunkbeg)-1))
chunkbeg <- matrix(chunkbeg[-nrow(chunkbeg)],nrow = (nrow(chunkbeg)-1))
chunkloc <- data.frame(begin=chunkbeg,end=chunkend)
chunk1 <- (sapply(seq_along(chunkloc$begin), function(i){
  paste(play1[seq(from=chunkloc$begin[i],
                  to = chunkloc$end[i],
                  by=1)])
}))
return(chunk1)
}


#list of chunks for each play
index <-matrix(c(1:36),ncol = 1)
chunks <- sapply(index, FUN = function(x) searchchunks(x))
```

```
# number of sentence per chunk
Searchsentence <- function(x){
   play1 <- chunks[[x]]
  index <- matrix(c(1:length(play1)),ncol=1)
  sentences <- sapply(index, function(i){
  str_extract_all(play1[i], "\\. [A-Z]{1,}")
})
  numofsentence <- matrix(lengths(sentences), ncol=1)
  return(numofsentence)
}
numofsentences <- sapply(index, FUN=function(x)Searchsentence(x))
```

```
# number of word per chunk
Searchwords <- function(x){
   play1 <- chunks[[x]]
  index <- matrix(c(1:length(play1)),ncol=1)
  wordperchunk <- sapply(index, function(i){
  str_extract_all(play1[i], " [a-z]{1,} ")
})
  wordsperchunk <- matrix(lengths(wordperchunk), ncol=1)
  return(wordsperchunk)
}
wordsperchunk <- sapply(index, FUN=function(x)Searchwords(x))

#average of word per chunk

avewordperchunk <-sapply(index, function(x){
  as.integer(mean(wordsperchunk[[x]], trim=0))
       })

#number of speakers per play
speakers <- sapply(index, function(x){
  str_extract_all(body[x], "  [A-Z]{4,}.")
})
speakers <- sapply(speakers,unique)
numofspeakers <- matrix(lengths(speakers), ncol=1)

#number of chunks per play
numofchunks <- matrix(lengths(chunks), ncol=1)


#number of unique word
words <- sapply(index, function(x){
  str_extract_all(body[x], " [a-z]{1,} ")
})
words <- sapply(words,unique)
numofwords <- matrix(lengths(words), ncol=1)
```

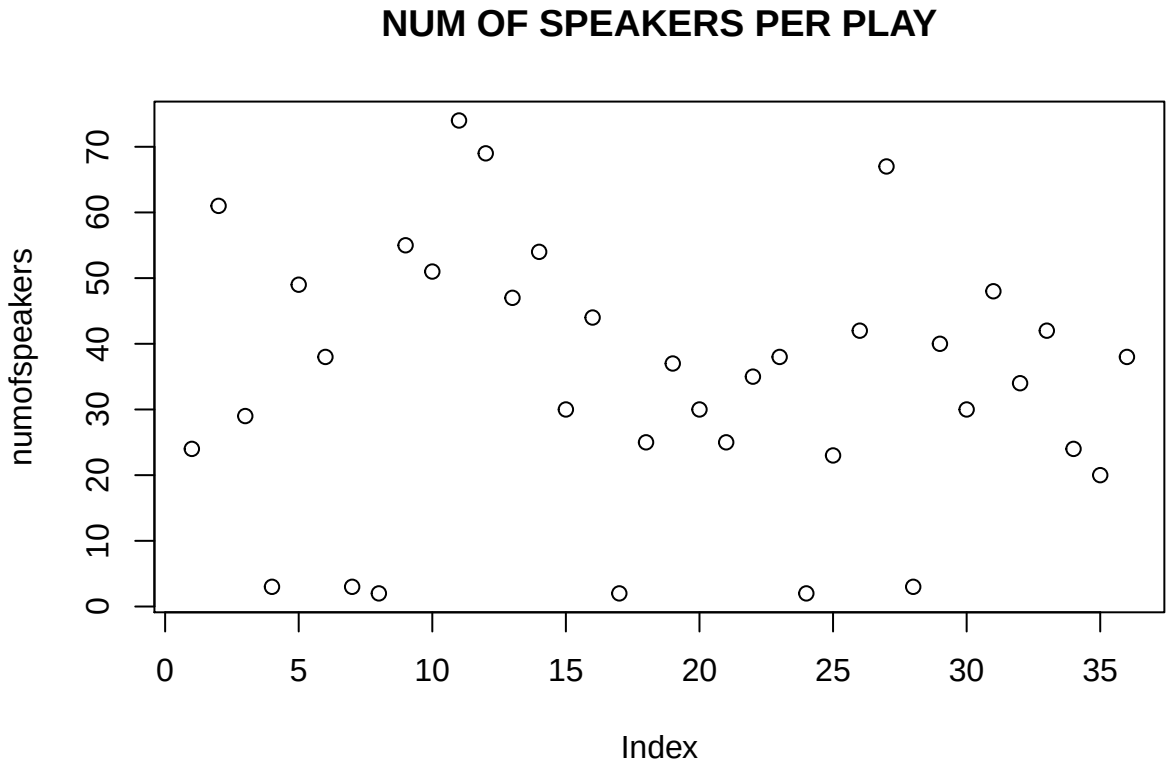
```
#metadata
metadata <- cbind(name,year,numofACT, numofscene, numofchunks,numofspeakers)
print(metadata)
```

```
##                                       name year numofACT numofscene
## 1                   ALLS WELL THAT ENDS WELL 1603        5         22
## 2       THE TRAGEDY OF ANTONY AND CLEOPATRA 1607        5         41
## 3                             AS YOU LIKE IT 1601        5         21
## 4                        THE COMEDY OF ERRORS 1593        5         10
## 5                  THE TRAGEDY OF CORIOLANUS 1608        5         28
## 6                                  CYMBELINE 1609        5         26
## 7  THE TRAGEDY OF HAMLET, PRINCE OF DENMARK 1604        5          4
## 8   THE FIRST PART OF KING HENRY THE FOURTH 1598        5          4
## 9               SECOND PART OF KING HENRY IV 1598        5         18
## 10         THE LIFE OF KING HENRY THE FIFTH 1599        5         22
## 11        THE FIRST PART OF HENRY THE SIXTH 1592        5         26
## 12  THE SECOND PART OF KING HENRY THE SIXTH 1591        5         23
## 13   THE THIRD PART OF KING HENRY THE SIXTH 1591        5         27
```

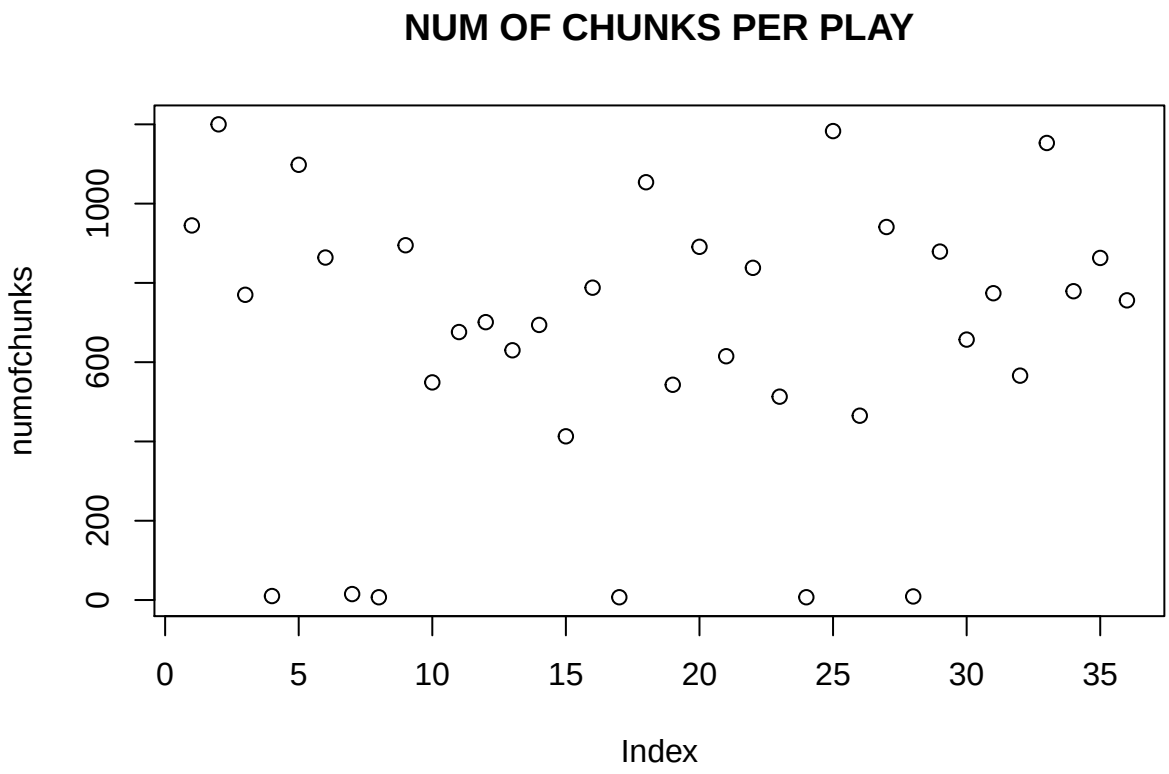
```
## 14                    KING HENRY THE EIGHTH 1611        5         16
## 15                                KING JOHN 1597        5         15
## 16             THE TRAGEDY OF JULIUS CAESAR 1599        5         17
## 17                 THE TRAGEDY OF KING LEAR 1606        5          4
## 18                     LOVE'S LABOUR'S LOST 1595        4          8
## 19                   THE TRAGEDY OF MACBETH 1606        5         28
## 20                     MEASURE FOR MEASURE 1605        5         15
## 21                   THE MERCHANT OF VENICE 1597        5         19
## 22                THE MERRY WIVES OF WINDSOR 1601        5         22
## 23                A MIDSUMMER NIGHT'S DREAM 1596        5          8
## 24                    MUCH ADO ABOUT NOTHING 1599        5          4
## 25   THE TRAGEDY OF OTHELLO, MOOR OF VENICE 1605        5         14
## 26                  KING RICHARD THE SECOND 1596        5         18
## 27                       KING RICHARD III 1593        5         24
## 28           THE TRAGEDY OF ROMEO AND JULIET 1595        5          4
## 29                  THE TAMING OF THE SHREW 1594        6         13
## 30                              THE TEMPEST 1612        5          8
## 31               THE LIFE OF TIMON OF ATHENS 1608        5         16
## 32           THE TRAGEDY OF TITUS ANDRONICUS 1594        5         13
## 33       THE HISTORY OF TROILUS AND CRESSIDA 1602        6         23
## 34          TWELFTH NIGHT; OR, WHAT YOU WILL 1602        5         17
## 35               THE TWO GENTLEMEN OF VERONA 1595        5         19
## 36                       THE WINTER'S TALE 1611        5         14
##    numofchunks numofspeakers
## 1          945            24
## 2         1200            61
## 3          770            29
## 4           10             3
## 5         1098            49
## 6          864            38
## 7           15             3
## 8            7             2
## 9          895            55
## 10         549            51
## 11         676            74
## 12         701            69
## 13         630            47
## 14         694            54
## 15         413            30
## 16         788            44
## 17           7             2
## 18        1054            25
## 19         543            37
## 20         891            30
## 21         615            25
## 22         838            35
## 23         513            38
## 24           7             2
## 25        1183            23
## 26         465            42
## 27         941            67
## 28           9             3
## 29         879            40
## 30         657            30
```

```
## 31         774            48
## 32         566            34
## 33        1153            42
## 34         779            24
## 35         863            20
## 36         756            38
```

```r
#number of speakers for each play
plot(numofspeakers, main = "NUM OF SPEAKERS PER PLAY")
```




```r
#number of chunks for each play
plot(numofchunks, main = "NUM OF CHUNKS PER PLAY")
```



Fields of class: 1. “name”, a character list contains the name of each play 2. “year”, a numeric list contains the year of each play 3. “speakers”, a character list contains the speakers of each play 4. “body”, a character list conatains the chunks of each play

Methods to process the text of the plays to produce the field 1. getname() # to get a matrix of names for each play from the original txt to produce ‘name’ field 2. getyear() # to get a matrix of year for each play from the original txt to produce the ‘year’ field 3. getspeaker() # to get a matrix of speakers for each play from the original txt to produce the ‘speakers’ field 4. getbody() # to get a matrix of main body for each play from the original txt to produce the ‘body’ field

Methods to provide information to a user 1. metadata() # to get a table of all the information including year, name, number of scene, number of speakers by combing related data. 2. getgraph() # to provide the graph use plot() function of the summary anlysis results to users. 3. print() # to print out the play user wanna see