Homework 8 *Biostatistics 203A* *Yenlin Lai*

Due date: Friday December 10th @ 12:01 PM

1. Text mining contains with three steps in the process: (a) establish the Corpus, (b) create the Term Document Matrix, and(c) extract knowledge. Use an article of your interest and run a text mining approach. The recommended R libraries to used are pdftools, stringr, rvest, tm, and SnowballC.

```
library(tm)
library(SnowballC)
library(wordcloud)
library(RColorBrewer)

filepath <- "https://pauladaunt.com/books/Golding,%20William%20-
%20Lord%20of%20the%20Flies%20v1.0.txt"
text <- readLines(filepath)
lof <- Corpus(VectorSource(text))
toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
lof <- tm_map(lof, toSpace, "/")
lof <- tm_map(lof, toSpace, "@")
lof <- tm_map(lof, toSpace, "\\|")
# Convert the text to lower case
lof <- tm_map(lof, content_transformer(tolower))
# Remove numbers
lof <- tm_map(lof, removeNumbers)
# Remove english common stopwords
lof <- tm_map(lof, removeWords, stopwords(kind="english"))
# Remove punctuations
lof <- tm_map(lof, removePunctuation)
# Eliminate extra white spaces
lof <- tm_map(lof, stripWhitespace)
# Text stemming
lof <- tm_map(lof, stemDocument)
inspect(lof)
```

```
dtm <- TermDocumentMatrix(lof)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d,10)
```

Description: df [10 × 2]

	word <chr>	freq <dbl>
ralph	ralph	793
piggi	piggi	411
jack	jack	374
said	said	262
look	look	259
fire	fire	212
boy	boy	212
one	one	204
back	back	204
like	like	196

1–10 of 10 rows

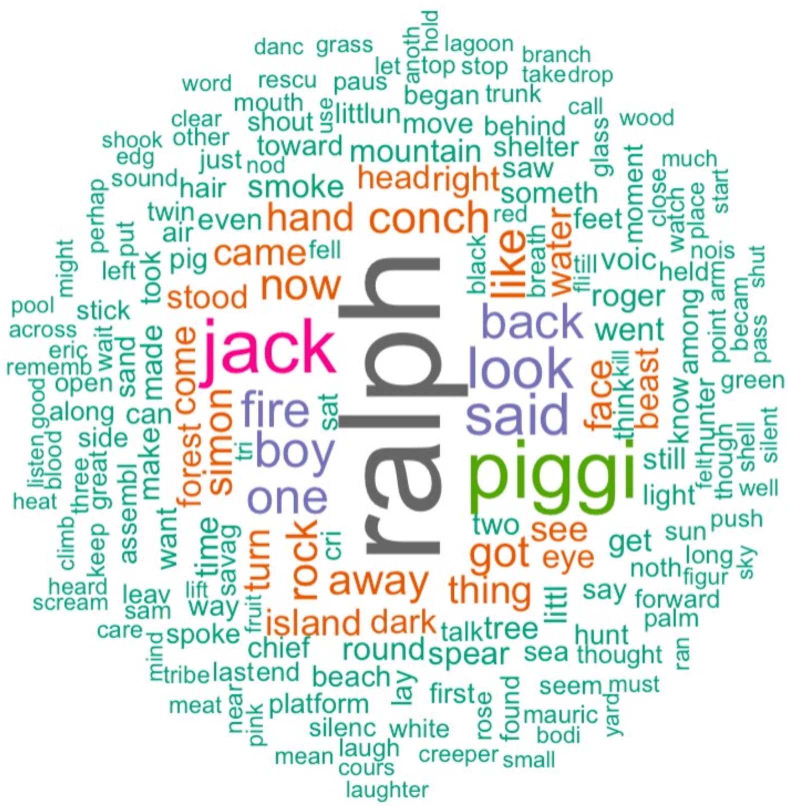
1.1. Report your frequency of Document_Term_Matrix using **str** command in R

```
str(d)
```

```
'data.frame':   3918 obs. of  2 variables:
 $ word: chr  "ralph" "piggi" "jack" "said" ...
 $ freq: num  793 411 374 262 259 212 212 204 204 196 ...
```

1.2. Show your word cloud plot that represents your article of interest using the R library of **wordcloud** and **RColorBrewer**

```
wordcloud(words = d$word, freq = d$freq, min.freq = 1,max.words=200,
random.order=FALSE, rot.per=0.35,colors=brewer.pal(8, "Dark2"))
```



1.3. Report the correlations of the words that are higher than a certain threshold. This threshold can be determined in the ranges from 0.15 to 0.50 if any.

```
wordralph <- findAssocs(dtm, "ralph", corlimit=0.15)
wordralph
```

```
$ralph
   piggi     look    stood      one     turn      sat     hand     came     jack    water
    0.25     0.20     0.20     0.19     0.18     0.18     0.17     0.17     0.17     0.16
    like    smile     away     sand     step constrain   diffid  dismiss   grassi   intrud
    0.16     0.16     0.16     0.16     0.16     0.16     0.16     0.16     0.16     0.16
  stroll     palm platform    spear
    0.16     0.15     0.15     0.15
```