Text Mining in R

巨量資料與統計分析

Hsiao Ling, Hsu

Nov. 30, 2016

R 語言簡介

- R語言自1993年問世，用於統計分析、繪圖、資料採礦、矩陣運算與機器學習等多個面向
- 兩大特色：免費下載、開放原始碼。
- 套件：ggmap, ggplot
- R Studio/R Pubs/GitHub
- 軟體安裝教學(R & SQL)

字串處理

常用字串處理function

- nchar
- grep
- grepl
- regexpr
- sub
- gsub
- substr
- paste
- strsplit

example

```
> address <- c("臺北市文山區指南路二段91~120號"
+          ,"臺北市大同區重慶北路一段61~90號"
+          ,"臺北市文山區指南路三段1~30號"
+          ,"臺北市文山區指南路二段45巷31~60號"
+          ,"臺北市內湖區民權東路六段90巷6弄1~30號"
+          ,"臺北市文山區興隆路四段1~30號")
```

```
> class(address)
```

```
[1] "character"
```

nchar

計算字串長度

```
> address
```

```
[1] "臺北市文山區指南路二段91~120號"             "臺北市大同區重慶北路一段61~90號"
[3] "臺北市文山區指南路三段1~30號"               "臺北市文山區指南路二段45巷31~60號"
[5] "臺北市內湖區民權東路六段90巷6弄1~30號" "臺北市文山區興隆路四段1~30號"
```

```
> nchar(address)
```

```
[1] 18 18 16 20 22 16
```

grep

grep第一個參數是pattern, 第二個是data,

output為有符合這個pattern是第幾筆資料

```
> grep(pattern = "文山區", x = address)
```

```
[1] 1 3 4 6
```

```
> grep("指南路", address)
```

```
[1] 1 3 4
```

grep (2)

```
> address
```

```
[1] "臺北市文山區指南路二段91~120號"           "臺北市大同區重慶北路一段61~90號"
[3] "臺北市文山區指南路三段1~30號"             "臺北市文山區指南路二段45巷31~60號"
[5] "臺北市內湖區民權東路六段90巷6弄1~30號" "臺北市文山區興隆路四段1~30號"
```

```
> (a <- grep("指南路二段", address))
```

```
[1] 1 4
```

```
> address[a]
```

```
[1] "臺北市文山區指南路二段91~120號"     "臺北市文山區指南路二段45巷31~60號"
```

grepl

與grep用法相似,差異在於其output是TRUE/FALSE

```
> grepl(pattern = "文山區", x = address)
```

```
[1]  TRUE FALSE  TRUE  TRUE FALSE  TRUE
```

```
> grepl("指南路", address)
```

```
[1]  TRUE FALSE  TRUE  TRUE FALSE FALSE
```

grepl (2)

```
> address
```

```
[1] "臺北市文山區指南路二段91~120號"           "臺北市大同區重慶北路一段61~90號"
[3] "臺北市文山區指南路三段1~30號"             "臺北市文山區指南路二段45巷31~60號"
[5] "臺北市內湖區民權東路六段90巷6弄1~30號" "臺北市文山區興隆路四段1~30號"
```

```
> (a <- grepl("指南路二段", address))
```

```
[1]  TRUE FALSE FALSE  TRUE FALSE FALSE
```

```
> address[a]
```

```
[1] "臺北市文山區指南路二段91~120號"     "臺北市文山區指南路二段45巷31~60號"
```

regexpr

找出**第一個**符合pattern的字串在哪個位置及長度,

如果不符合pattern,會顯示-1

```
> address
```

```
[1] "臺北市文山區指南路二段91~120號"          "臺北市大同區重慶北路一段61~90號"
[3] "臺北市文山區指南路三段1~30號"            "臺北市文山區指南路二段45巷31~60號"
[5] "臺北市內湖區民權東路六段90巷6弄1~30號" "臺北市文山區興隆路四段1~30號"
```

```
> regexpr(pattern = "指", text = address)
```

```
[1]  7 -1  7  7 -1 -1
attr(,"match.length")
[1]  1 -1  1  1 -1 -1
```

regexpr (2)

```
> regexpr(pattern = "指南路", text = address)
```

```
[1]  7 -1  7  7 -1 -1
attr(,"match.length")
[1]  3 -1  3  3 -1 -1
```

```
> regexpr("指南路二段", address)
```

```
[1]  7 -1 -1  7 -1 -1
attr(,"match.length")
[1]  5 -1 -1  5 -1 -1
```

regexpr (3)

```
> regexpr("號", address)
```

```
[1] 18 18 16 20 22 16
attr(,"match.length")
[1] 1 1 1 1 1 1
```

```
> regexpr("[0-9]", address)
```

```
[1] 12 13 12 12 13 12
attr(,"match.length")
[1] 1 1 1 1 1 1
```

gregexpr

找出所有符合pattern的字串在哪個位置及長度

```
> address
```

```
[1] "臺北市文山區指南路二段91~120號"           "臺北市大同區重慶北路一段61~90號"
[3] "臺北市文山區指南路三段1~30號"             "臺北市文山區指南路二段45巷31~60號"
[5] "臺北市內湖區民權東路六段90巷6弄1~30號" "臺北市文山區興隆路四段1~30號"
```

先回憶一下剛剛regexpr的結果

```
> regexpr("[0-9]", address)
```

```
[1] 12 13 12 12 13 12
attr(,"match.length")
[1] 1 1 1 1 1 1
```

```
> gregexpr("[0-9]", address)
```

```
[[1]]
[1] 12 13 15 16 17
attr(,"match.length")
[1] 1 1 1 1 1

[[2]]
[1] 13 14 16 17
attr(,"match.length")
[1] 1 1 1 1

[[3]]
[1] 12 14 15
attr(,"match.length")
[1] 1 1 1

[[4]]
[1] 12 13 15 16 18 19
attr(,"match.length")
[1] 1 1 1 1 1 1

[[5]]
```

sub

sub指substitute,

把每個字串中**第一個**符合pattern的內容取代

```
> sub(pattern = "指南路", replacement = "AAA", x = address)
```

```
[1] "臺北市文山區AAA二段91~120號"              "臺北市大同區重慶北路一段61~90號"
[3] "臺北市文山區AAA三段1~30號"                "臺北市文山區AAA二段45巷31~60號"
[5] "臺北市內湖區民權東路六段90巷6弄1~30號" "臺北市文山區興隆路四段1~30號"
```

```
> sub("[0-9]", "X", address)
```

```
[1] "臺北市文山區指南路二段X1~120號"           "臺北市大同區重慶北路一段X1~90號"
[3] "臺北市文山區指南路三段X~30號"             "臺北市文山區指南路二段X5巷31~60號"
[5] "臺北市內湖區民權東路六段X0巷6弄1~30號" "臺北市文山區興隆路四段X~30號"
```

gsub

把每個字串中**所有**符合pattern的內容取代

```
> gsub(pattern = "指南路", replacement = "AAA", x = address)
```

```
[1] "臺北市文山區AAA二段91~120號"              "臺北市大同區重慶北路一段61~90號"
[3] "臺北市文山區AAA三段1~30號"                "臺北市文山區AAA二段45巷31~60號"
[5] "臺北市內湖區民權東路六段90巷6弄1~30號" "臺北市文山區興隆路四段1~30號"
```

```
> gsub("[0-9]","X",address)
```

```
[1] "臺北市文山區指南路二段XX~XXX號"          "臺北市大同區重慶北路一段XX~XX號"
[3] "臺北市文山區指南路三段X~XX號"            "臺北市文山區指南路二段XX巷XX~XX號"
[5] "臺北市內湖區民權東路六段XX巷X弄X~XX號" "臺北市文山區興隆路四段X~XX號"
```

substr

```
> address
```

```
[1] "臺北市文山區指南路二段91~120號"            "臺北市大同區重慶北路一段61~90號"
[3] "臺北市文山區指南路三段1~30號"              "臺北市文山區指南路二段45巷31~60號"
[5] "臺北市內湖區民權東路六段90巷6弄1~30號" "臺北市文山區興隆路四段1~30號"
```

- 如果我只想要地址中的[行政區 + 路段]

```
> substr(address, start = 4, stop = 12)
```

```
[1] "文山區指南路二段9"  "大同區重慶北路一段" "文山區指南路三段1"  "文山區指南路二段4"
[5] "內湖區民權東路六段" "文山區興隆路四段1"
```

substr (2)

結合regexpr

```
> (a1 = regexpr("[0-9]",address))
```

```
[1] 12 13 12 12 13 12
attr(,"match.length")
[1] 1 1 1 1 1 1
```

```
> (a2 = regexpr("號",address))
```

```
[1] 18 18 16 20 22 16
attr(,"match.length")
[1] 1 1 1 1 1 1
```

substr (3)

```
> substr(x = address, start = a1, stop = a2)
```

```
[1] "91~120號"       "61~90號"        "1~30號"          "45巷31~60號"    "90巷6弄1~30號" "1~30號"
```

```
> substr(address, 4, a1)
```

```
[1] "文山區指南路二段9"    "大同區重慶北路一段6" "文山區指南路三段1"    "文山區指南路二段4"
[5] "內湖區民權東路六段9" "文山區興隆路四段1"
```

```
> substr(address, 4, a1-1)
```

```
[1] "文山區指南路二段"    "大同區重慶北路一段" "文山區指南路三段"    "文山區指南路二段"
[5] "內湖區民權東路六段" "文山區興隆路四段"
```

paste

字串的剪貼

```
> paste("臺北市", substr(address, 4, a1-1), substr(x = address, start = a1, stop = a2))
```

```
[1] "臺北市 文山區指南路二段 91~120號"          "臺北市 大同區重慶北路一段 61~90號"
[3] "臺北市 文山區指南路三段 1~30號"            "臺北市 文山區指南路二段 45巷31~60號"
[5] "臺北市 內湖區民權東路六段 90巷6弄1~30號" "臺北市 文山區興隆路四段 1~30號"
```

```
> paste("臺北市", substr(address, 4, a1-1),
+       substr(x = address, start = a1, stop = a2),sep = "")
```

```
[1] "臺北市文山區指南路二段91~120號"          "臺北市大同區重慶北路一段61~90號"
[3] "臺北市文山區指南路三段1~30號"            "臺北市文山區指南路二段45巷31~60號"
[5] "臺北市內湖區民權東路六段90巷6弄1~30號" "臺北市文山區興隆路四段1~30號"
```

```
> # paste(..., sep="") 相當於 paste0(...)
```

strsplit

字串的切割

```
> strsplit(address,"市")
```

```
[[1]]
[1] "臺北"                          "文山區指南路二段91~120號"

[[2]]
[1] "臺北"                          "大同區重慶北路一段61~90號"

[[3]]
[1] "臺北"                          "文山區指南路三段1~30號"

[[4]]
[1] "臺北"                          "文山區指南路二段45巷31~60號"

[[5]]
[1] "臺北"                          "內湖區民權東路六段90巷6弄1~30號"
```

Reference

- 正規表示式 Regular Expression
- Stringr套件

文字處理

讀取資料

- readLines()
- 試讀蔡英文總統就職演說中文全文

```
> readLines("Tsai.txt")
```

```
 [1] "各位友邦的元首與貴賓、各國駐台使節及代表、現場的好朋友，全體國人同胞，大家好"
 [2] ""
 [3] "感謝與承擔"
 [4] ""
 [5] "就在剛剛，我和陳建仁已經在總統府裡面，正式宣誓就任中華民國第十四任總統與副總統。我們要感謝這塊土地對我
 [6] ""
 [7] "台灣，再一次用行動告訴世界，作為一群民主人與自由人，我們有堅定的信念，去捍衛民主自由的生活方式。這段旅
 [8] ""
 [9] "我要告訴大家，對於一月十六日的選舉結果，我從來沒有其他的解讀方式。人民選擇了新總統、新政府，所期待的就
[10] ""
[11] "我也要告訴大家，眼前的種種難關，需要我們誠實面對，需要我們共同承擔。所以，這個演說是一個邀請，我要邀請
[12] ""
[13] "國家不會因為領導人而偉大；全體國民的共同奮鬥，才讓這個國家偉大。總統該團結的不只是支持者，總統該團結的
[14] ""
[15] "在我們共同奮鬥的過程中，身為總統，我要向全國人民宣示，未來我和新政府，將領導這個國家的改革，展現決心，
```

資料處理

- 安裝套件

```
> install.packages("tm")
```

- 載入套件

```
> library("tm")
```

- 刪除標點符號與數字

```
> removePunctuation()
> removeNumbers()
```

```
> # 讀入資料
> (G <- readLines("Tsai.txt"))
```

```
 [1] "各位友邦的元首與貴賓、各國駐台使節及代表、現場的好朋友，全體國人同胞，大家好"
 [2] ""
 [3] "感謝與承擔"
 [4] ""
 [5] "就在剛剛，我和陳建仁已經在總統府裡面，正式宣誓就任中華民國第十四任總統與副總統。我們要感謝這塊土地對我
 [6] ""
 [7] "台灣，再一次用行動告訴世界，作為一群民主人與自由人，我們有堅定的信念，去捍衛民主自由的生活方式。這段旅
 [8] ""
 [9] "我要告訴大家，對於一月十六日的選舉結果，我從來沒有其他的解讀方式。人民選擇了新總統、新政府，所期待的就
[10] ""
[11] "我也要告訴大家，眼前的種種難關，需要我們誠實面對，需要我們共同承擔。所以，這個演說是一個邀請，我要邀請
[12] ""
[13] "國家不會因為領導人而偉大；全體國民的共同奮鬥，才讓這個國家偉大。總統該團結的不只是支持者，總統該團結的
[14] ""
[15] "在我們共同奮鬥的過程中，身為總統，我要向全國人民宣示，未來我和新政府，將領導這個國家的改革，展現決心，
[16] ""
[17] "為年輕人打造一個更好的國家"
[18] ""
[19] "未來的路並不好走，台灣需要一個正面迎向一切挑戰的新政府，我的責任就是領導這個新政府。"
[20] ""
```

```
> # 刪除標點符號
> (G1 <- removePunctuation(G))
```

```
 [1] "各位友邦的元首與貴賓各國駐台使節及代表現場的好朋友全體國人同胞大家好"
 [2] ""
 [3] "感謝與承擔"
 [4] ""
 [5] "就在剛剛我和陳建仁已經在總統府裡面正式宣誓就任中華民國第十四任總統與副總統我們要感謝這塊土地對我們的栽
 [6] ""
 [7] "台灣再一次用行動告訴世界作為一群民主人與自由人我們有堅定的信念去捍衛民主自由的生活方式這段旅程我們每一
 [8] ""
 [9] "我要告訴大家對於一月十六日的選舉結果我從來沒有其他的解讀方式人民選擇了新總統新政府所期待的就是四個字解
[10] ""
[11] "我也要告訴大家眼前的種種難關需要我們誠實面對需要我們共同承擔所以這個演說是一個邀請我要邀請全體國人同胞
[12] ""
[13] "國家不會因為領導人而偉大全體國民的共同奮鬥才讓這個國家偉大總統該團結的不只是支持者總統該團結的是整個國
[14] ""
[15] "在我們共同奮鬥的過程中身為總統我要向全國人民宣示未來我和新政府將領導這個國家的改革展現決心絕不退縮"
[16] ""
[17] "為年輕人打造一個更好的國家"
[18] ""
[19] "未來的路並不好走台灣需要一個正面迎向一切挑戰的新政府我的責任就是領導這個新政府"
[20] ""
```

```
> # 刪除數字
> (G1 <- removeNumbers(G1))
```

```
 [1] "各位友邦的元首與貴賓各國駐台使節及代表現場的好朋友全體國人同胞大家好"
 [2] ""
 [3] "感謝與承擔"
 [4] ""
 [5] "就在剛剛我和陳建仁已經在總統府裡面正式宣誓就任中華民國第十四任總統與副總統我們要感謝這塊土地對我們的栽
 [6] ""
 [7] "台灣再一次用行動告訴世界作為一群民主人與自由人我們有堅定的信念去捍衛民主自由的生活方式這段旅程我們每一
 [8] ""
 [9] "我要告訴大家對於一月十六日的選舉結果我從來沒有其他的解讀方式人民選擇了新總統新政府所期待的就是四個字解
[10] ""
[11] "我也要告訴大家眼前的種種難關需要我們誠實面對需要我們共同承擔所以這個演說是一個邀請我要邀請全體國人同胞
[12] ""
[13] "國家不會因為領導人而偉大全體國民的共同奮鬥才讓這個國家偉大總統該團結的不只是支持者總統該團結的是整個國
[14] ""
[15] "在我們共同奮鬥的過程中身為總統我要向全國人民宣示未來我和新政府將領導這個國家的改革展現決心絕不退縮"
[16] ""
[17] "為年輕人打造一個更好的國家"
[18] ""
[19] "未來的路並不好走台灣需要一個正面迎向一切挑戰的新政府我的責任就是領導這個新政府"
[20] ""
```

```
> # 計算列數
> (row = length(G1))
```

```
[1] 141
```

```
> # 將所有段落連結在一起
> (Gf <-  paste(G1[1:row],collapse = ""))
```

```
[1] "各位友邦的元首與貴賓各國駐台使節及代表現場的好朋友全體國人同胞大家好感謝與承擔就在剛剛我和陳建仁已經在總
```

```
> # 將所有的空白格移除掉
> (Gfinal <- gsub(" ","",Gf))
```

```
[1] "各位友邦的元首與貴賓各國駐台使節及代表現場的好朋友全體國人同胞大家好感謝與承擔就在剛剛我和陳建仁已經在總
```

```
> # 先宣告一個空向量
> word = NULL
```

```
> # 計算就職演說總字數
> (n = nchar(Gfinal))
```

```
[1] 5356
```

```
> # 每兩個字兩個字做字串剪貼（12,23,34,45,...）
> for(i in 1:n-1){
+   word <- c(word, substr(Gfinal,i,i+1))
+ }
> word
```

```
 [1] "各"   "各位" "位友" "友邦" "邦的" "的元" "元首" "首與" "與貴" "貴賓" "賓各" "各國" "國駐"
[14] "駐台" "台使" "使節" "節及" "及代" "代表" "表現" "現場" "場的" "的好" "好朋" "朋友" "友全"
[27] "全體" "體國" "國人" "人同" "同胞" "胞大" "大家" "家好" "好感" "感謝" "謝與" "與承" "承擔"
[40] "擔就" "就在" "在剛" "剛剛" "剛我" "我和" "和陳" "陳建" "建仁" "仁已" "已經" "經在" "在總"
[53] "總統" "統府" "府裡" "裡面" "面正" "正式" "式宣" "宣誓" "誓就" "就任" "任中" "中華" "華民"
[66] "民國" "國第" "第十" "十四" "四任" "任總" "總統" "統與" "與副" "副總" "總統" "統我" "我們"
[79] "們要" "要感" "感謝" "謝這" "這塊" "塊土" "土地" "地對" "對我" "我們" "們的" "的栽" "栽培"
```

```
> wordtable_i <- table(word)
> wordtable_i[1:10]
```

```
word
愛的 安教 安矛 安全 安心 安與 岸的 岸都 岸關 岸過
   2    1    1    8    1    1    1    1    5    1
```

```
> wordtable <- sort(wordtable_i,decreasing = TRUE)
> wordtable[1:10]
```

```
word
我們 台灣 政府 國家 一個 經濟 新政 這個 民主 社會
  86   41   37   32   29   27   27   25   24   22
```

文字雲

- 安裝套件

```
> install.packages("wordcloud")
```

- 載入套件

```
> library(wordcloud)
```

- 建立data.frame

```
> (d <- data.frame(word = names(wordtable), freq = as.numeric(wordtable)))
```

```
     word freq
1    我們   86
2    台灣   41
3    政府   37
4    國家   32
5    一個   29
6    經濟   27
```

Word Cloud

- `?wordcloud`

```
> wordcloud(words,freq,scale=c(4,.5),min.freq=3,max.words=Inf,
+   random.order=TRUE, random.color=FALSE, rot.per=.1,
+   colors="black",ordered.colors=FALSE,use.r.layout=FALSE,
+   fixed.asp=TRUE, ...)
```

- words : 文字
- freq : 出現次數
- min.freq : 最小出現次數，若低於此值，則不會畫在圖上
- max.words : 最多畫幾組文字
- random.order : 是不是要隨機順序來畫圖
- colors : 顏色選取ColorBrewer
- … 其餘參數設定請參考 `?wordcloud` or `help(wordcloud)`

```
> par(family = 'STHeiti') # MAC使用者需此行程式碼才能顯示中文字
> wordcloud(d$word, d$freq, scale=c(8,.2),min.freq=3,
+           max.words=Inf, random.order=FALSE,
+           colors=c("#7F7F7F", "#5F9EA0", "#FF8C69"))
```

Thank you.