大數據資料處理實務

李水彬

2023-09-01

Chapter 05-2 集合與字串的運算

集合比對

- %in% 運算子

```
x<-c(3,4,5)
y<-c(3,5,6,7,8)
```

```
x%in%y #比對每個x的元素是否屬於y
```

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

```
y%in%x #比對每個y的元素是否屬於x
```

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

集合比對

```
which(x%in%y) # 哪幾個x的元素屬於y
```

```
## [1] 1 3
```

```
which(y%in%x) # 哪幾個y的元素屬於x
```

```
## [1] 1 2
```

課堂練習11

```
x<-c(5, 11,  4, 10,  7, 13, 10,  6, 14,  6,  9,  5, 14,  9, 8,  9, 10,  9,  6,  8)
```

```
y<-1:10
```

找出那些 x 落在 y 內?

字串比較

== 比較兩字串是否相同

```
x<-c("白酒蛤蜊義大利麵", "青醬義大利麵", "香炸雞腿排", "香草烤雞飯", "泰式三杯雞飯", "椰香咖哩飯",
     "韓式石鍋飯","和風照燒豬排","清蒸檸檬魚","雲南椒麻雞")
```

```
x=="和風照燒豬排"
```

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

字串內文筆對

- grepl() 函數: 判定是否含有給定字串

```
x<-c("白酒蛤蜊義大利麵", "青醬義大利麵", "香炸雞腿排", "香草烤雞飯", "泰式三杯雞飯", "椰香咖哩飯",
     "韓式石鍋飯","和風照燒豬排","清蒸檸檬魚","雲南椒麻雞")
```

- 找出麵類主餐

```
grepl("麵",x)
```

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

字串內文筆對

- 找出雞肉主餐

```
grepl("雞",x)
```

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

兩種條件的比對

```
nd.flag<-grepl("麵",x)
ck.flag<-grepl("雞",x)
nd.flag|ck.flag # 或
```

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

```
nd.flag&ck.flag # 且
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
!(nd.flag) # 非麵食
```

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

兩種條件的比對

```
!(ck.flag) # 非雞肉
```

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

```
!(nd.flag)&!(ck.flag) #沒有雞肉也沒有麵
```

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

顯示含比對文字的內容

```
x<-c("白酒蛤蜊義大利麵", "青醬義大利麵", "香炸雞腿排", "香草烤雞飯", "泰式三杯雞飯", "椰香咖哩飯",
     "韓式石鍋飯","和風照燒豬排","清蒸檸檬魚","雲南椒麻雞")
```

- 是否含有雞肉

```
ck.flag<-grepl("雞",x)
ck.flag
```

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

- 含有雞肉的資料位置

```
ck.idx<-which(ck.flag)
ck.idx
```

```
## [1]  3  4  5 10
```

顯示含比對文字的內容

- 顯示主餐名稱

```
#ck.idx 含有雞肉的資料位置
x[ck.idx]
```

```
## [1] "香炸雞腿排"     "香草烤雞飯"     "泰式三杯雞飯" "雲南椒麻雞"
```

課堂練習 12

```
x<-c("白酒蛤蜊義大利麵", "青醬義大利麵", "香炸雞腿排", "香草烤雞飯", "泰式三杯雞飯", "椰香咖哩飯",
     "韓式石鍋飯","和風照燒豬排","清蒸檸檬魚","雲南椒麻雞")
```

- 飯類主餐名稱。
- 主餐不是麵的名稱

grep() 函數

傳回有該字串的位置

```
x<-c("白酒蛤蜊義大利麵", "青醬義大利麵", "香炸雞腿排", "香草烤雞飯", "泰式三杯雞飯", "椰香咖哩飯",
     "韓式石鍋飯","和風照燒豬排","清蒸檸檬魚","雲南椒麻雞")
```

- 傳回字串的位置

```
grep("義大利",x,value=FALSE)
```

```
## [1] 1 2
```

- 傳回字串

```
grep("義大利",x,value=TRUE)
```

```
## [1] "白酒蛤蜊義大利麵" "青醬義大利麵"
```

regexpr() 函數

- 正數(positive): 有比對字串，傳回比對字串在受比對字串的第幾個字元出現。

```
x<-c("白酒蛤蜊義大利麵", "青醬義大利麵", "香炸雞腿排", "香草烤雞飯", "泰式三杯雞飯", "椰香咖哩飯",
     "韓式石鍋飯","和風照燒豬排","清蒸檸檬魚","雲南椒麻雞")
```

```
regexpr("義大利",x)
```

```
##  [1]  5  3 -1 -1 -1 -1 -1 -1 -1 -1
## attr(,"match.length")
##  [1]  3  3 -1 -1 -1 -1 -1 -1 -1 -1
```

- -1: 受比對字串中沒有比對字串

蔡英文總統國慶演說

大會主席游錫堃院長、諾魯共和國昆洛斯總統伉儷、聖克里斯多福及尼維斯聯邦萊柏總督、聖文森及格瑞那丁朵根總督，現場的貴賓、好朋友，還有我們剛從亞運歸來的臺灣之光，以及收看電視和網路直播的國人同胞：大家好！今天是中華民國112年的國慶日。闊別三年，我們終於脫下口罩、齊聚在此，共度國家的慶典。現場有許多來自全球各地的僑胞，還有許多睽違三年，再次遠道而來的國際友人，我要代表臺灣人民，向大家致上最真摯的感謝。 回首三年來辛苦的防疫之路，彷彿那段日子已經很遙遠。然而，有另一條艱辛的路，我們走了三十年。就在上個月底，「潛艦國造」的第一艘原型艦下水了。在完成後續的測試後，這艘「海鯤軍艦」預計在2025年正式服役。 潛艦國造是歷經三十年，不同政黨的總統，都想實現的夢想。現在，我們做到了！從無到有，踏出這一步需要無比的勇氣。要扛住壓力，要突破瓶頸，要頂住流言蜚語，只要稍有猶豫，就會失敗。但是，我們終於做到了！我們的國防自主再跨出一大步，國軍不對稱戰力再向上提升；我們更再次展現， 守護中華民國臺灣的決心。我相信，全世界更會認同，海鯤軍艦是為了守護區域和平穩定而破浪前行。這正是中華民國立足臺灣七十四年來，之所以屹立不搖的精神。面對特殊的國際處境和瞬息萬變的挑戰，我們不前進就會倒退；不奮進努力，就無法掌握自己的未來和命運。

蔡英文總統國慶演說

- 有無提到台灣?

```
grepl("臺灣",x)
```

```
## [1] TRUE
```

- 有無提到中華民國?

```
grepl("中華民國",x)
```

```
## [1] TRUE
```

蔡英文總統國慶演說

- 哪裡提到台灣?

```
regexpr("臺灣",x)
```

```
## [1] 76
## attr(,"match.length")
## [1] 2
```

第一出現在第76個字，長度2。

蔡英文總統國慶演說

- 哪裡提到中華民國?

```
regexpr("中華民國",x)
```

```
## [1] 107
## attr(,"match.length")
## [1] 4
```

第一次出現在第107個字，字的長度4。

多次重複

- gregexpr()
- 哪裡提到台灣?

```
gregexpr("臺灣",x)
```

```
## [[1]]
## [1]  76 188 450 500
## attr(,"match.length")
## [1] 2 2 2 2
```

多次重複

- 哪裡提到中華民國?

```
gregexpr("中華民國",x)
```

```
## [[1]]
## [1] 107 446 494
## attr(,"match.length")
## [1] 4 4 4
```

結果

大會主席游錫堃院長、諾魯共和國昆洛斯總統伉儷、聖克里斯多福及尼維斯聯邦萊柏總督、聖文森及格瑞那丁朵根總督，現場的貴賓、好朋友，還有我們剛從亞運歸來的**臺灣**之光，以及收看電視和網路直播的國人同胞：大家好！今天是 **中華民國**112年的國慶日。闊別三年，我們終於脫下口罩、齊聚在此，共度國家的慶典。現場有許多來自全球各地的僑胞，還有許多睽違三年，再次遠道而來的國際友人，我要代表 **臺灣**人民，向大家致上最真摯的感謝。 回首三年來辛苦的防疫之路，彷彿那段日子已經很遙遠。然而，有另一條艱辛的路，我們走了三十年。就在上個月底，「潛艦國造」的第一艘原型艦下水了。在完成後續的測試後，這艘「海鯤軍艦」預計在2025年正式服役。 潛艦國造是歷經三十年，不同政黨的總統，都想實現的夢想。現在，我們做到了！從無到有，踏出這一步需要無比的勇氣。要扛住壓力，要突破瓶頸，要頂住流言蜚語，只要稍有猶豫，就會失敗。但是，我們終於做到了！我們的國防自主再跨出一大步，國軍不對稱戰力再向上提升；我們更再次展現， 守護 **中華民國** **臺灣**的決心。我相信，全世界更會認同，海鯤軍艦是為了守護區域和平穩定而破浪前行。這正是 **中華民國**立足 **臺灣**七十四年來，之所以屹立不搖的精神。面對特殊的國際處境和瞬息萬變的挑戰，我們不前進就會倒退；不奮進努力，就無法掌握自己的未來和命運。

課堂練習 13

- 找出以下這段話，那些地方出現**大數據** 這三個字?

大數據，台灣又稱巨量資料，指的是傳統資料處理應用軟體不足以處理的大或複雜的資料集的術語。 大數據也可以定義為來自各種來源的大量非結構化或結構化資料。從學術角度而言，大數據的出現促成廣泛主題的新穎研究。這也導致各種大數據統計方法的發展。大數據並沒有統計學的抽樣方法；它只是觀察和追蹤發生的事情。因此，大數據通常包含的資料大小超出傳統軟體在可接受的時間內處理的能力。由於近期的技術進步，發布新資料的便捷性以及全球大多數政府對高透明度的要求，大數據分析在現代研究中越來越突出。