

字串搜尋和取代

- Textbook reading: Sections 10.3.9
- 字串搜尋指令 `grep` 語法：

```
grep(pattern, x, ignore.case=FALSE)
```

其中 `x` 為字串向量，`pattern` 為 regular expression 搜尋格式，`ignore.case=FALSE` 表示大小寫視為不同。

- 一些 regular expression pattern 格式

```
^      #比對開頭
$      #比對結束
.      #比對表示任何一個字元(換行符號除外)
*      #比對前一個字元零次或更多次
{n}    #比對前一個字元 n 次
{n,}   #比對前一個字元至少 n 次
[A-C]  #A或B或C
```

可用 `help(regex)` 查到更多

- Example 1. 執行以下指令生成字串向量 `x`

```
x <- c("abd123txt", "ab.txt", "bdk", "A35.png", "ahomework", "ak", "bak", "LALA")
```

寫下 R 指令完成下列工作

- 列出 `x` 中包含 “bd” 的字串
- 列出 Part (a) 之字串在向量 `x` 中的位置
- 列出 `x` 中以 “ab” 開頭的字串
- 列出 `x` 中以 “k” 結束的字串
- 列出 `x` 中以 a-k 中之字母結束的字串
- 列出 `x` 中以四個英文字母結束的字串。
- 列出 `x` 中以 “a” 開頭且以 “k” 結束的字串
- 列出 `x` 中以 “a” 開頭或以 “k” 結束的字串
- 列出 `x` 中以 “.txt” 結束的字串

Sol.

```
#a
x[grep("bd",x)]
```

```
#b
grep("bd",x)
```

```

#c
x[grepl("^ab",x)]

#d
x[grepl("k$",x)]

#e
x[grepl("[a-k]$",x)]

#f
x[grepl("[a-z, A-Z]{4}%",x)]

#or
x[grepl("[a-z]{4}%",x, ignore.case = TRUE)]
#ignore.case=TRUE 表示大小寫視為相同

#g
x[grepl("^a.*k%",x)]

#h
x[grepl("^a|k%",x)]

#i
x[grepl("[.]txt%",x)] #note that x[grepl(".txt%",x)] gives a different result

```

- 字串取代指令 gsub 語法：

```
gsub(pattern, newstring, x, ignore.case=FALSE)
```

其中 `x` 為字串向量，`pattern` 為搜尋格式，`newstring` 為取代字串，`ignore.case=FALSE` 表示大小寫視為不同。

Example 2. 執行以下指令生成字串向量 `x`

```
x <- c("abd123", "ab.txt", "bdk", "A35.png", "ahomework", "ak", "bak")
```

寫下 R 指令完成下列工作：

- 列出 `x` 中包含“t”的字串
- 將 Part (a) 每個字串中的“t”改成“T”並存成字串向量 `y`
- 列出 `x` 中以 a-z 中字母結束的字串
- 將 Part (c) 每個字串中的結束字母改成“A”並存成字串向量 `z`
- 將 `x` 字串中的“t”改成“T”並存成字串向量 `y`

Sol.

```

#a
x[grept("t",x)]

#b
y <- gsub("t", "T",x[grept("t",x)])

#c
x[grept("[a-z]$",x)]

#d
z <- gsub("[a-z]$", "A",x[grept("[a-z]$",x)])

#e
y <- gsub("t", "T",x)

```

- Regular expression pattern 也可配合list.files()使用。

Example 3. 執行以下指令生成檔案f1.txt, f2.txt:

```

cat("Start of f1", "\n", "End of file\n", sep="", file="f1.txt")
cat("One line\n", file="f2.txt")

```

寫下R指令完成下列工作：

- 列出工作目錄下檔名以.txt結束的檔案名稱。
- 將工作目錄下檔名以.txt結束的所有檔案內容依序合併到一個新檔案new.txt，合併要按照Part(a)列出的順序。
- 刪除工作目錄下檔名以.txt結束的檔案。

合併檔案的指令為file.append, 刪除檔案的指令為file.remove.

Sol.

```

#a
list.files(pattern=".txt$")

#b
flist <- list.files(pattern=".txt$")
n <- length(flist)
file.create("new.txt")
for (i in 1:n){
  file.append("new.txt", flist[i])
}

#c
file.remove(list.files(pattern=".txt$"))

```

- Note that

```
flist <- list.files(pattern=".txt$")
```

is equivalent to

```
flist <- list.files()  
flist <- flist[ grep(pattern=".txt$", flist) ]
```

- 使用Regular expression pattern 配合cat指令可修改檔案內容

Example 4. 執行以下指令生成檔案f1.txt:

```
cat("Start of f1", "\n", "End of file\n", sep="", file="f1.txt")
```

寫下R指令將檔案f1.txt內容中所有的“f”換成“g”。

```
x <- scan(file="f1.txt", what="char", sep="\n")  
x <- gsub("f", "g",x)
```

```
cat(x,file="f1.txt",sep="\n", append = FALSE)  
#append預設為 FALSE，若省略不設效果也相同
```