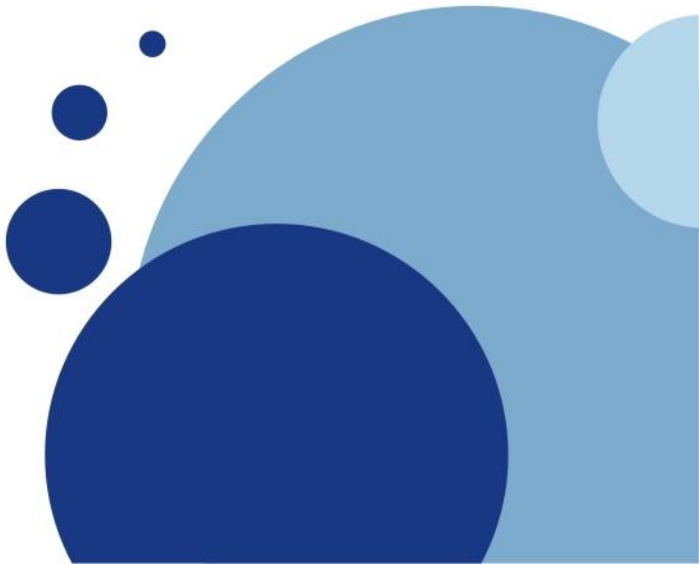
R
Ladies

運用 R Crawler 爬取 XML 資料

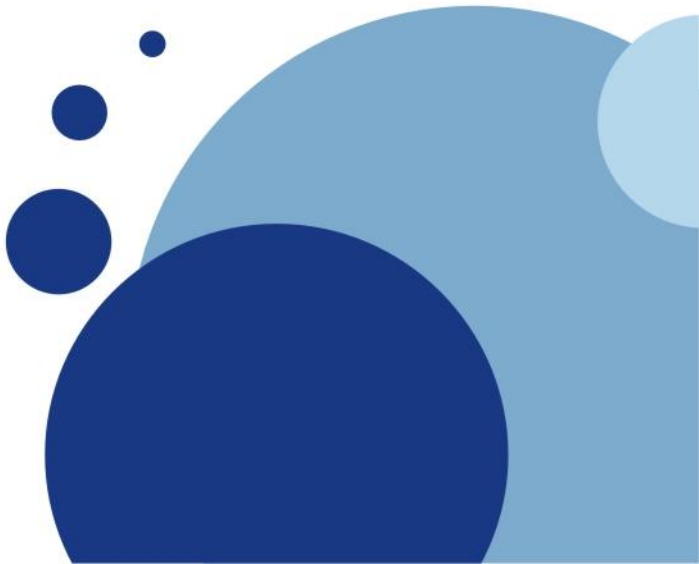
R Ladies 分享 by 陳露西



R
Ladies

About Me
陳露西

- HR發展 - 組織&人才
- 經營分析- 機會~在哪兒?
- 吃吃吃
- 旅遊
- 跳舞
- 閱讀 - 心
- 學習
- R
- 跑跑
- 散步

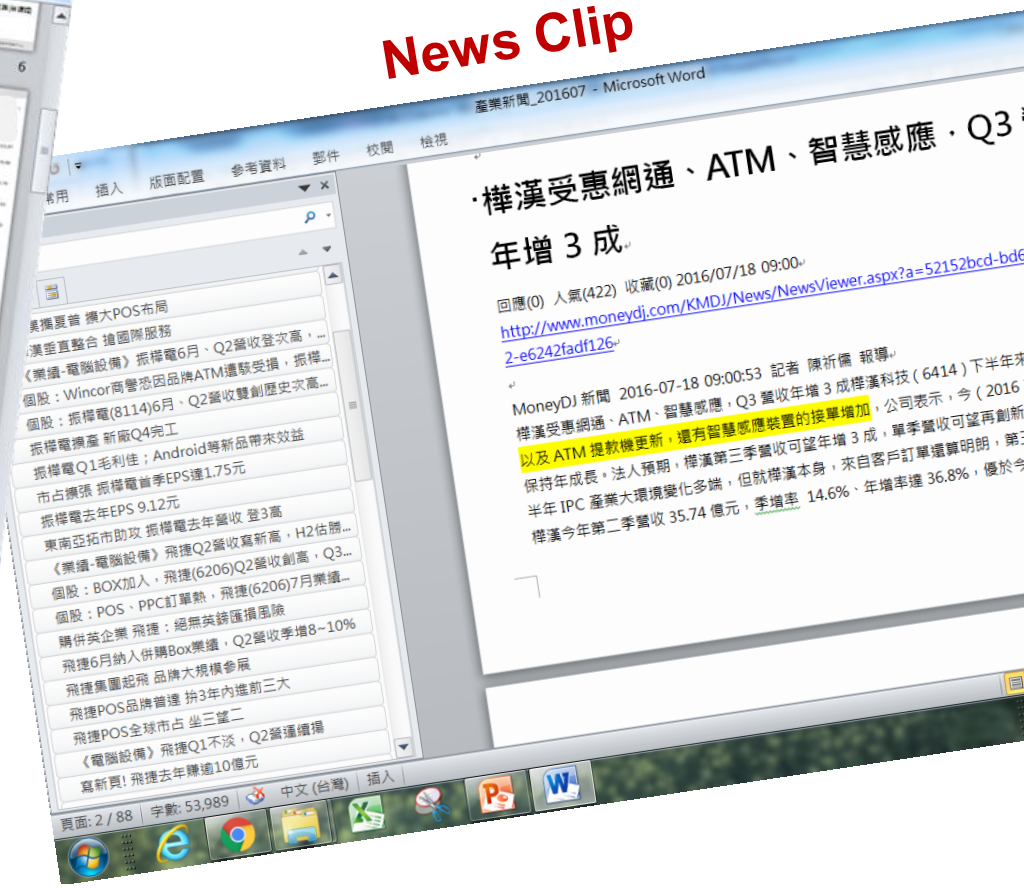


R Ladies

為什麼想要學爬文?

彙整報告

News Clip



R Ladies

運用 R Crawler 爬取 XML 資料

- XML & RCurl Package
- 根據網址爬取圖表
  - 固定網址
  - 網址組合
  - (範例) 爬取各種金屬價格趨勢圖
- 爬取Table資訊
  - 觀察網頁
  - 定義變數與函數
  - 爬取Table標題與內容
  - 排列與存取Table資訊
- 要再研究、集思廣益的部分

R Ladies

指令

- **download.file：下載檔案**

  download.file (url, destfile)

  destfile – Where the file should be saved (path with a file name).

- **getURL：下載URLs**

  下載一個或多個URI(URLs)：getURL(url)

  指定編碼與檔案類型：getURL(url,encoding="UTF-8", asText=TRUE)

- **htmlParse：爬HTML**

  parse HTML in a string into a representation of choice.

  htmlParse (URL, encoding="UTF-8")

  htmlParse(getURL(url))

- **xpathSApply：爬向量/矩陣資料**

  xpathSApply is a version of xpathApply which attempts to simplify the result if it can be converted to a vector or matrix rather than left as a list. In this way, it has the same relationship to xpathApply as sapply has to lapply.

  xpathSApply(html, "//text()[(ancestor::h3)]", xmlValue)

  xpathSApply(html, "//text()[(ancestor::table)]",

R Ladies

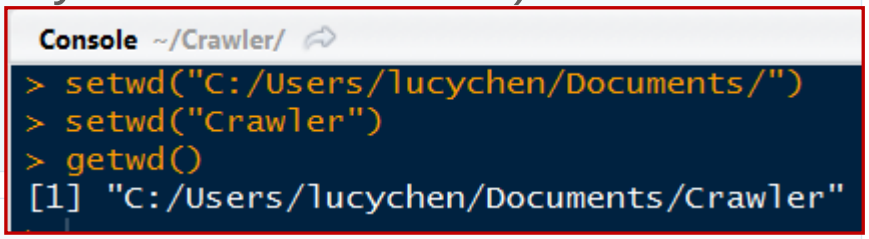
XML & RCurl Package

- 安裝(Instal)與載入(Library)XML及Rcurl套件

```
install.packages("XML")
install.packages("RCurl")
library(XML)
library(RCurl)
```

- 設定專案存放位址

```
setwd("C:/Users/lucychen/Documents/")
setwd("Crawler")
getwd()
```

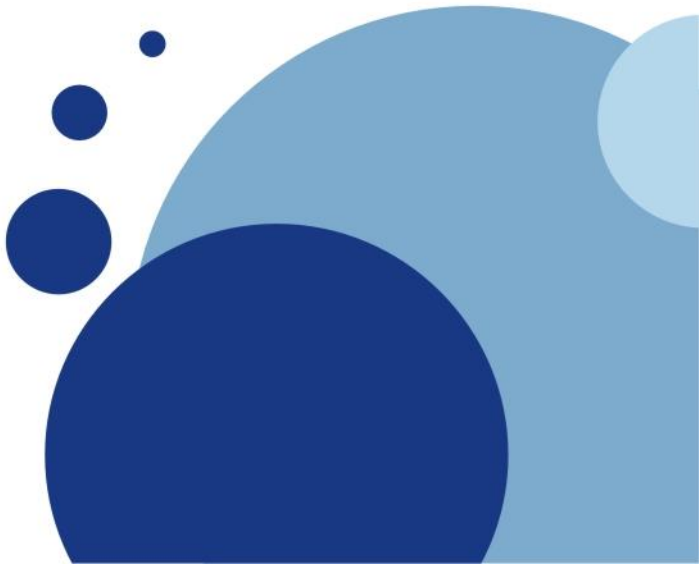



```
setwd
getwd()
```

若不特別指定位置，R會將專案位置->我的文件

R
Ladies

根據網址爬取圖表



根據網址爬取圖表

- 根據「固定網址」爬取圖片

```
URL<-"http://stockcharts.com/c-sc/sc?s=%24USD&p=D&yr=1&mn=0&dy=0&i=t47587542618&r=1469686084775"
download.file(URL,"stockcharts.png", mode = 'wb')
```

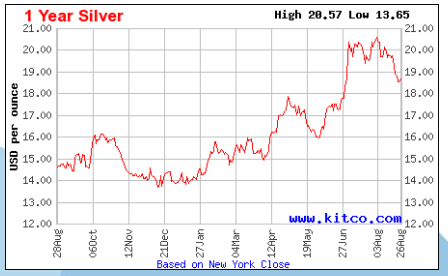
美元指數

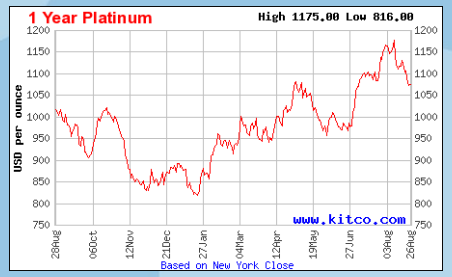
R Ladies

根據網址爬取圖表

```
URL3 <-
c("http://www.kitco.com/LFgif/au0365nyb.gif",
"http://www.kitco.com/LFgif/ag0365nyb.gif",
"http://www.kitco.com/LFgif/pt0365nyb.gif")
```

貴金屬現貨價格



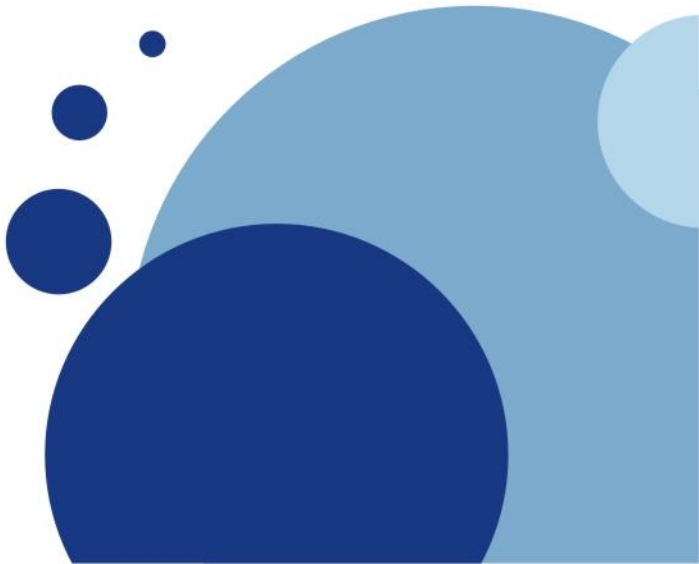


- 爬取同個網站「多張」圖片，找規律

```
count <-0
URL3 <-0
kitco ="http://www.kitco.com/LFgif/"
preciousmetal=c("au","ag","pt")
tail="0365nyb.gif"
for (k in(1:3))
    URL3[k] <- c(paste(kitco, preciousmetal[k], tail, sep=""))
for (i in (1:length(URL3))){
    download.file(URL3[i], paste('data',i,'.gif',sep=""), mode = 'wb')
    count=count+1
}
path<-getwd()
cat(paste("於",path,"共下載", count, "個gif文件\n", sep = ""))
```

R
Ladies

爬取Table資訊



R Ladies

目標網頁(1)...

R Ladies

目標網頁(2)...

R Ladies

預期結果

各國匯率

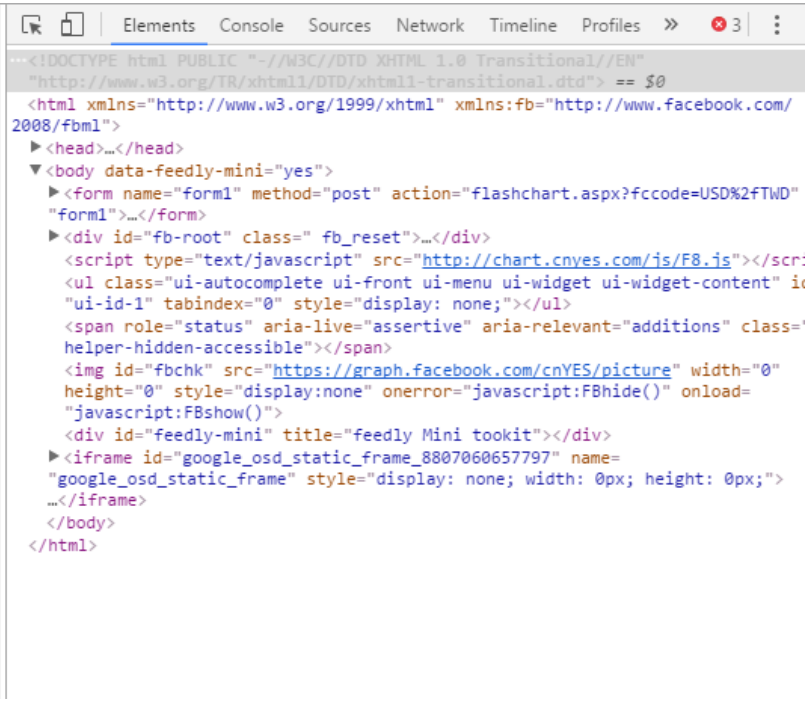
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		日期	收盤價	漲跌	漲%	開盤價	最高價	最低價	昨收	一週	一個月	三個月	六個月	一年	今年以來	52週最高	52週最低
2	美元/台幣	20160826	31.672	0.002	0.0063	31.651	31.7	31.624	31.67	0.17%	-1.23%	-2.63%	-5.43%	-3.35%	-4.22%	33.838	31.225
3	歐元/美元	20160826	1.1195	-0.0089	-0.7887	1.12833	1.13407	1.11807	1.1284	-1.15%	1.27%	0.78%	2.41%	-0.45%	3.10%	1.15268	1.05629
4	美元/人民幣	20160826	6.6688	0.0135	0.2028	6.6553	6.6696	6.6553	6.6553	0.26%	-0.02%	1.65%	2.02%	4.14%	2.76%	6.6984	6.318
5	美元/日元	20160826	101.81	1.276	1.2692	100.524	101.944	100.064	100.534	1.60%	-3.31%	-7.63%	-10.65%	-15.95%	-15.31%	123.614	99.921
6	美元/俄羅斯盧布	20160826	64.722	0.046	0.0711	64.8928	64.9904	63.8889	64.676	1.45%	-1.50%	-2.02%	-15.07%	-0.33%	-11.33%	81.9391	61.0899
7	美元/巴西里爾	20160826	3.2645	0.0341	1.0556	3.2314	3.2768	3.1905	3.2304	1.92%	0.01%	-9.67%	-18.42%	-8.18%	-17.59%	4.1695	3.1288
8	美元/馬來西亞林吉特	20160826	4.03	0.0205	0.5113	4.0095	4.03	4.0095	4.0095	0.49%	-1.06%	-1.10%	-4.50%	-4.46%	-6.14%	4.456	3.8375
9	美元/土耳其里拉	20160826	2.9508	0.0171	0.5829	2.9347	2.9587	2.9242	2.9337	0.60%	-2.18%	-0.17%	-1.22%	1.49%	1.16%	3.0884	2.7962
10	美元/印尼盾	20160826	13200	-38	-0.2871	13240	13251	13210	13238	0.34%	0.49%	-2.81%	-1.23%	-5.65%	-4.49%	14695	13010

金屬現貨價格

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		日期	收盤價	漲跌	漲%	開盤價	最高價	最低價	昨收	一週	一個月	三個月	六個月	一年	今年以來	52週最高	52週最低
2	錫-現貨(LME Tin Spot)	20160826	18,900.00	141	0.75	18,920.00	18,920.00	18,920.00	18,759.00	-0.47%	-0.80%	-2.63%	-5.43%	-3.36%	-4.22%	33.838	31.225
3	鎳-現貨(LME Nickel Spot)	20160826	9,757.25	-34.5	-0.35	9,770.00	9,770.00	9,770.00	9,791.75	-0.47%	-0.80%	-2.63%	-5.43%	-3.36%	-4.22%	33.838	31.225

R Ladies

觀察網頁



R
Ladies

定義變數與函數

- 定義變數
  - ✓ Exg：爬取的匯率名稱，也是URL的尾巴

    ```
    Exg <-
    c("USD/TWD","EUR/USD","USD/CNY","USD/JPY","USD/RUB"
    ,"USD/BRL","USD/MYR","USD/TRY","USD/IDR")
    ```

  - ✓ Metal：爬取的金屬名稱，也是URL2的尾巴

    ```
    Metal <- c("TNCS.html","NDCS.html")
    ```

- 定義函數

  1. 爬取匯率價格

     ```
     parsetablecontent <- function(Exg, URL)
     ```

  2. 爬取金屬價格

     ```
     parsetablecontent2 <- function(Metal, URL2)
     ```

  3. 表格清理

     ```
     cleartables <- function(Tables)
     ```

R Ladies

主程式

```
##3. crawler for table and figures
###3.1 URL for the exchange rate
source ("parsetable0829.R",encoding="UTF-8")
#函數parsetablecontent, parsetablecontent2 都寫在 "parsetable0829.R"檔案

Exg <-
c("USD/TWD","EUR/USD","USD/CNY","USD/JPY","USD/RUB","USD/BRL","USD/
MYR","USD/TRY","USD/IDR")
URL <- "http://www.cnyes.com/forex/flashchart.aspx?fccode="

###3.2 URL2 for the metal price
Metal <- c("TNCS.html","NDCS.html")
URL2 <- "http://www.cnyes.com/futures/javachart/"
```

```
###3.3 parse the exchagerate and metal price
temp1<-parsetablecontent(Exg,URL)
temp2<-parsetablecontent2(Metal,URL2)
```

```
###3.3 write to csv
write.csv(temp1, sprintf(paste("匯率",".csv"))) #檔名:匯率
write.csv(temp2, sprintf(paste("金屬",".csv"))) #檔名:金屬
```

函數1-1- parsetablecontent <- function(Exg, URL)

```
for (j in (1:8)){
    for (i in (1:length(Exg))){
      html <- htmlParse(getURL(paste(URL,Exg[i],sep="")), encoding="UTF-8", asText=TRUE)
      heads<-xpathSApply(html, "//text()[(ancestor::h3)]", xmlValue)    # h3會抓到標題列
      tables<-xpathSApply(html, "//text()[(ancestor::table)]", xmlValue) # table會抓數值列
      d<- as.data.frame(tables[(8+j)])      # tables的第 (8+j) 位是data   # as.data.frame資料格式
      colnames(d) = c(heads[9])             # heads的第9位是匯率幣別的標題
      if (i==1){
        temp <-d
      }else{
        temp <-cbind(temp, d)
      }
    }
    temp=t(temp) # transpose
    colnames(temp) = c(tables[j])
    if (j==1){
      Exgnames <-temp
    }else{
      Exgnames <- cbind(Exgnames, temp)
    }
  }
```

Tables

[1] "日期" [2] "收盤價" [3] "漲跌" [4] "漲%"
[5] "開盤價" [6] "最高價" [7] "最低價" [8] "昨收"
[9] "20160826" [10] "31.672" [11] "0.002" [12] "0.0063"
[13] "31.651" [14] "31.700" [15] "31.624" [16] "31.670"
[17] "一週" [18] "一個月" [19] "三個月" [20] "六個月"
[21] "一年" [22] "今年以來" [23] "52週最高" [24] "52週最低"
[25] "-0.47%" [26] "-0.80%" [27] "-2.63%" [28] "-5.43%"
[29] "-3.36%" [30] "-4.22%" [31] "33.838" [32] "31.225" ...

heads

[1] "新聞頭條" "新聞頭條" "新聞頭條"
[4] "新聞頭條" "鉅亨基金交易平台" "新聞頭條"
[7] "新聞頭條" "人氣排行榜" "美元/台幣"

美元/新台幣 讚 18 人說這讚。

日期	收盤價	漲跌	漲%	開盤價	最高價	最低價	昨收
20160826	31.672	0.002	0.0063	31.651	31.700	31.624	31.670

績效表現

一週	一個月	三個月	六個月	一年	今年以來	52週最高	52週最低
-0.47%	-0.80%	-2.63%	-5.43%	-3.36%	-4.22%	33.838	31.225

注意事項

此'美元/台幣'資料來源，與其他外幣資料來源相同，皆是來自國外的外匯經紀商報價，每個交易日18:00以後，我們再以台北外匯經紀公司收盤價更新當天價格。

R Ladies

函數1-1- parsetablecontent <- function(Exg, URL)

```
for (j in (1:8)){
    for (i in (1:length(Exg))){

      html <- htmlParse(getURL(paste(URL,Exg[i],sep="")), encoding="UTF-8", asText=TRUE)
      heads<-xpathSApply(html, "//text()[(ancestor::h3)]", xmlValue)    # h3會抓到標題列
      tables<-xpathSApply(html, "//text()[(ancestor::table)]", xmlValue) # table會抓數值列
      d<- as.data.frame(tables[(8+j)])      # tables的第 (8+j) 位是data   # as.data.frame資料格式
      colnames(d) = c(heads[9])             # heads的第9位是匯率幣別的標題

      if (i==1){
        temp <-d                    # i=1時
      }else{
        temp <-cbind(temp, d)  # i>1之後，將資料逐步併到右邊的column
      }
    }
    temp= t (temp)                      # transpose
    colnames(temp) = c(tables[j])    # transpose後，再加上第j個欄位名稱
    if (j==1){
      Exgnames <-temp                 # j=1時
    }else{
      Exgnames <- cbind(Exgnames, temp) # j>1之後，將資料逐步併到右邊的column
      }
    }
  }
```

R Ladies

函數1-2- parsetablecontent <- function(Exg, URL)

```
for (j in (1:8)){
    for (i in (1:length(Exg))){
        html <- htmlParse(getURL(paste(URL,Exg[i],sep="")), encoding="UTF-8", asText=TRUE)
        heads<-xpathSApply(html, "//text()[(ancestor::h3)]", xmlValue)
        tables<-xpathSApply(html, "//text()[(ancestor::table)]", xmlValue)
        d<- as.data.frame(tables[(24+j)])
        colnames(d) = c(heads[9])
        if (i==1){
            temp <-d
        }else{
            temp <- cbind(temp, d)
        }
    }
    temp <-t(temp)
    colnames(temp) <- c(tables[16+j])
    if (j==1){
        Exgnames2=temp
    }else{
        Exgnames2=cbind(Exgnames2, temp)
    }
  }
  TTLExgnames=cbind(Exgnames, Exgnames2) # 將二個table合併成一個總table
        }
  return(TTLExgnames)
}
```

Tables

[1] "日期" [2] "收盤價" [3] "漲跌" [4] "漲%"
[5] "開盤價" [6] "最高價" [7] "最低價" [8] "昨收"
[9] "20160826" [10] "31.672" [11] "0.002" [12] "0.0063"
[13] "31.651" [14] "31.700" [15] "31.624" [16] "31.670"
[17] "一週" [18] "一個月" [19] "三個月" [20] "六個月"
[21] "一年" [22] "今年以來" [23] "52週最高" [24] "52週最低"
[25] "-0.47%" [26] "-0.80%" [27] "-2.63%" [28] "-5.43%"
[29] "-3.36%" [30] "-4.22%" [31] "33.838" [32] "31.225" ...

heads

[1] "新聞頭條" "新聞頭條" "新聞頭條"
[4] "新聞頭條" "鉅亨基金交易平台" "新聞頭條"
[7] "新聞頭條" "人氣排行榜" "美元/台幣"

美元/新台幣 讚 18 人說這讚。

日期	收盤價	漲跌	漲%	開盤價	最高價	最低價	昨收
20160826	31.672	0.002	0.0063	31.651	31.700	31.624	31.670

績效表現

一週	一個月	三個月	六個月	一年	今年以來	52週最高	52週最低
-0.47%	-0.80%	-2.63%	-5.43%	-3.36%	-4.22%	33.838	31.225

注意事項

此'美元/台幣'資料來源，與其他外幣資料來源相同，皆是來自國外的外匯經紀商報價，每個交易日18:00以後，我們再以台北外匯經紀公司收盤價更新當天價格。

R Ladies

函數2-1- parsetablecontent2 <- function(Metal, URL2)

```
for (j in (1:8)){
    for (i in (1:length(Metal))){
        html <- htmlParse(getURL(paste(URL2,Metal[i],sep="")), encoding="UTF-8", asText=TRUE)
        heads<-xpathSApply(html, "//text()[(ancestor::h3)]", xmlValue)
        tables<-xpathSApply(html, "//text()[(ancestor::table)]", xmlValue)
        tables<-cleartables(tables)  # 清理表格(函數內容請見次二頁)
        d<- as.data.fra
        colnames(d) =
        if (i==1){
            temp =d
        }else{
            temp = cbind
        }
    }
    temp=t(temp)
    colnames(temp) =
    if (j==1){
        MetalName=te
    }else{
        MetalName = c
    }
}
```

```
> source("parsetable0829.R",encoding="UTF-8")
> Metal <- c("TNCS.html","NDCS.html")
> URL2 <- "http://www.cnyes.com/futures/javachart/"
> temp2<-parsetablecontent2(Metal,URL2)
> temp2
                                日期      \r\n      收盤價    \r\n    漲跌
錫-現貨(LME Tin Spot)\r\n     "開盤價" "最高價" "最低價" "\r\n" "昨收"
鎳-現貨(LME Nickel Spot)\r\n  "開盤價" "最高價" "最低價" "\r\n" "昨收"
                                \r\n              漲%          \r\n
錫-現貨(LME Tin Spot)\r\n     "\r\n    \r\n" "20160826" "18,900.000"
鎳-現貨(LME Nickel Spot)\r\n  "\r\n    \r\n" "20160826" "9,757.250"
                               -34.50    -0.35     9,770.000 9,770.000
錫-現貨(LME Tin Spot)\r\n     "一個月" "三個月" "六個月"   "今年以來"
鎳-現貨(LME Nickel Spot)\r\n  "一個月" "三個月" "六個月"   "今年以來"
                               9,770.000 9,791.750 績效表現 一週
錫-現貨(LME Tin Spot)\r\n     "一年"    "三年"    "五年"    "1.69"
鎳-現貨(LME Nickel Spot)\r\n  "一年"    "三年"    "五年"    "-5.33"
                               一個月
錫-現貨(LME Tin Spot)\r\n     "7.08"
鎳-現貨(LME Nickel Spot)\r\n  "-5.36"
```

R Ladies

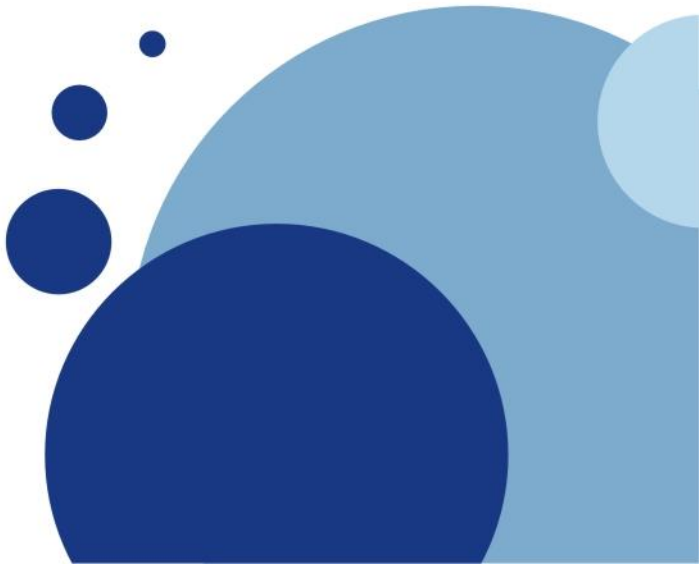
函數2-2- parsetablecontent2 <- function(Metal, URL2)

```
for (j in (1:9)){
    for (i in (1:length(Metal))){
        html <- htmlParse(getURL(paste(URL,Metal[i],sep="")), encoding="UTF-8", asText=TRUE)
        heads<-xpathSApply(html, "//text()[(ancestor::h3)]", xmlValue)
        tables<-xpathSApply(html, "//text()[(ancestor::table)]", xmlValue)
        tables<-cleartables(tables)  # 清理表格(函數內容請見次頁)
        d<- as.data.frame(tables[(24+j)])  # tables的第(24+j)位是第二列的數字
        colnames(d) = c(heads[9])  #heads的第9位是匯率幣別的標題
        if (i==1){
            temp <-d
        }else{
            temp <- cbind(temp, d)
        }
    }
    temp <-t(temp)
    colnames(temp) <- c(tables[16+j])
    if (j==1){
        Metalname2=temp
    }else{
        Metalname2=cbind(Metalname2, temp)
    }
}
TTLMetalnames=cbind(MetalName, Metalname2)
return(TTLMetalnames)
}
```

R Ladies

清理表格- cleartables <- function(Tables)

```
cleartables <- function(Tables){
    k=1
    Tables=trimws(Tables)
    #Remove leading and/or trailing whitespace "\r\n"
    while (k<=length(Tables)){
        if (Tables[k]==""){
            Tables=Tables[-k]
            #remove kth element
            k=1
        }
        k<-k+1
        if (k==length(Tables)){ #當全部檢查完後，跳出此while迴圈
            break
        }
    }
    return(Tables)
```

R
Ladies

Thanks for Joining