資料科學 Data Science 系列

R 語言資料分析實務 (2)

文字探勘 - 文字雲製作

姓名：羅左欣

日期：2016/10/17(一)

部落格：http://shouzo.github.io/

Agenda

(一) Prepare：預備工作

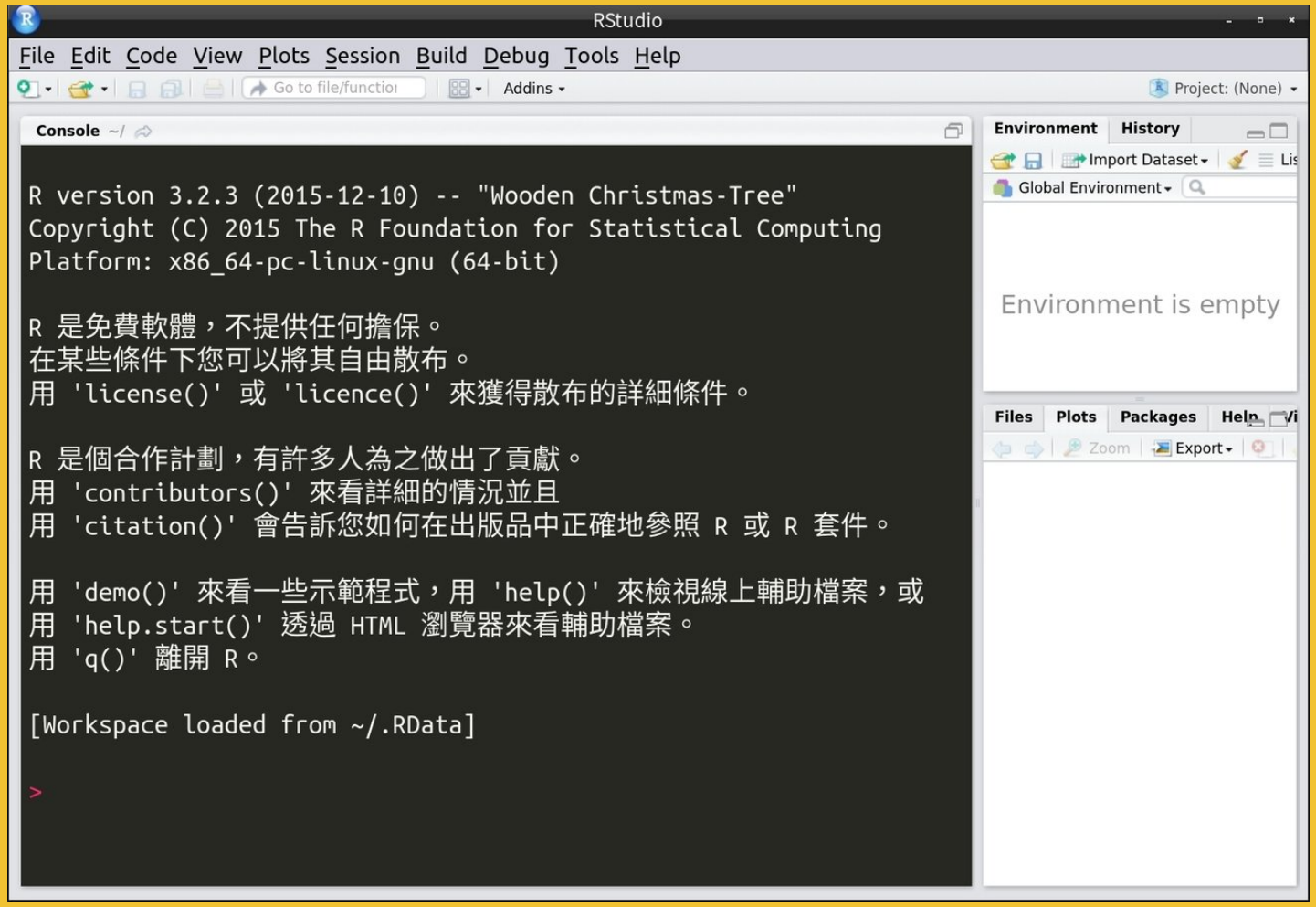
(二) Basic：基本介紹與操作

(三) Theme：文字雲製作

(四) Reference：學習資源

(一) Prepare：預備工作

(一) Prepare：預備工作



在這個系列的簡報中，主要以 "RStudio" 做為主要軟體。

(二) Basic：基本介紹與操作

1. 網頁分析

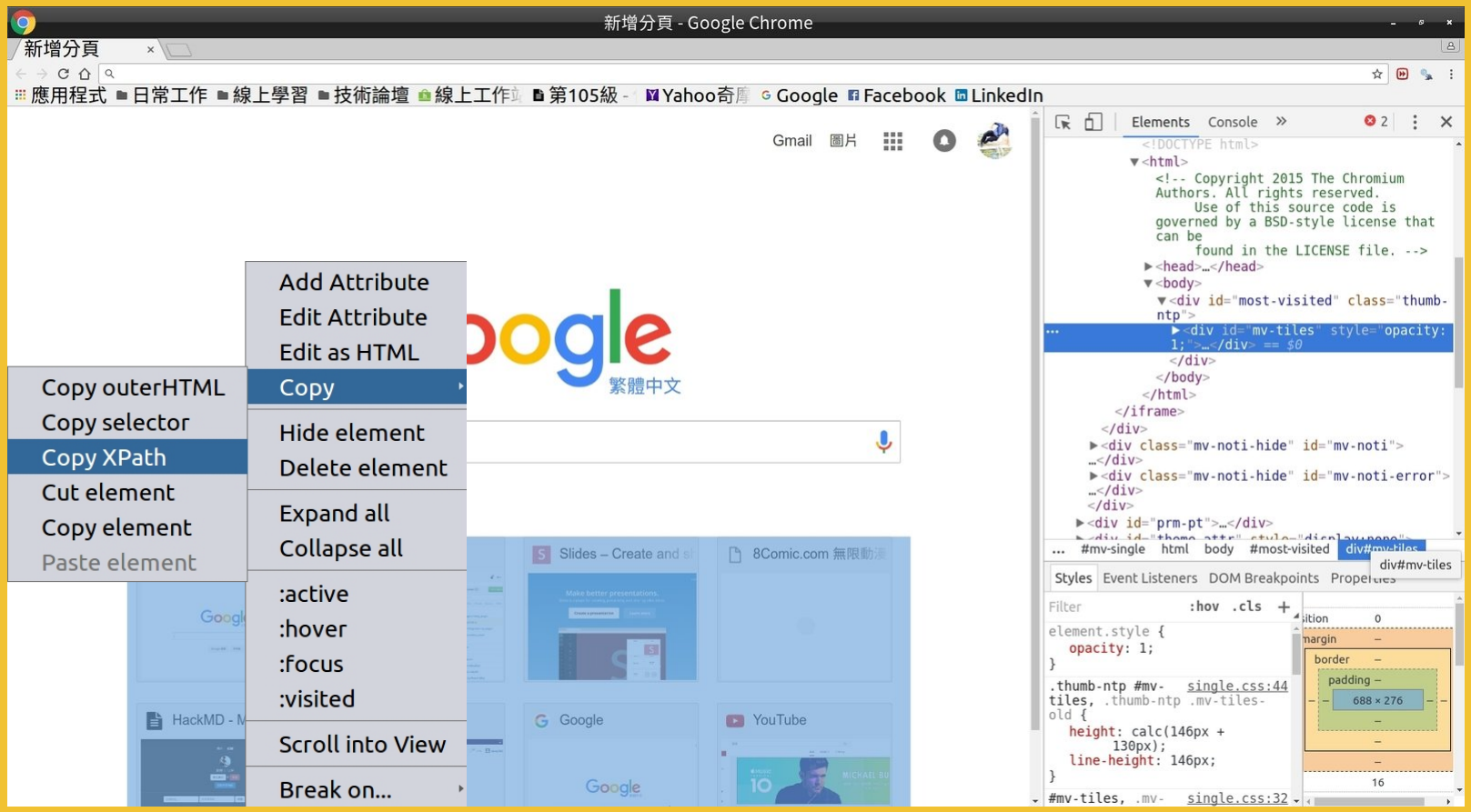
參考教材：使用 R 與 rvest 套件擷取網頁資料

➡ 教材網址： https://blog.gtwang.org/r/rvest-web-scraping-with-r/

尋找 Xpath

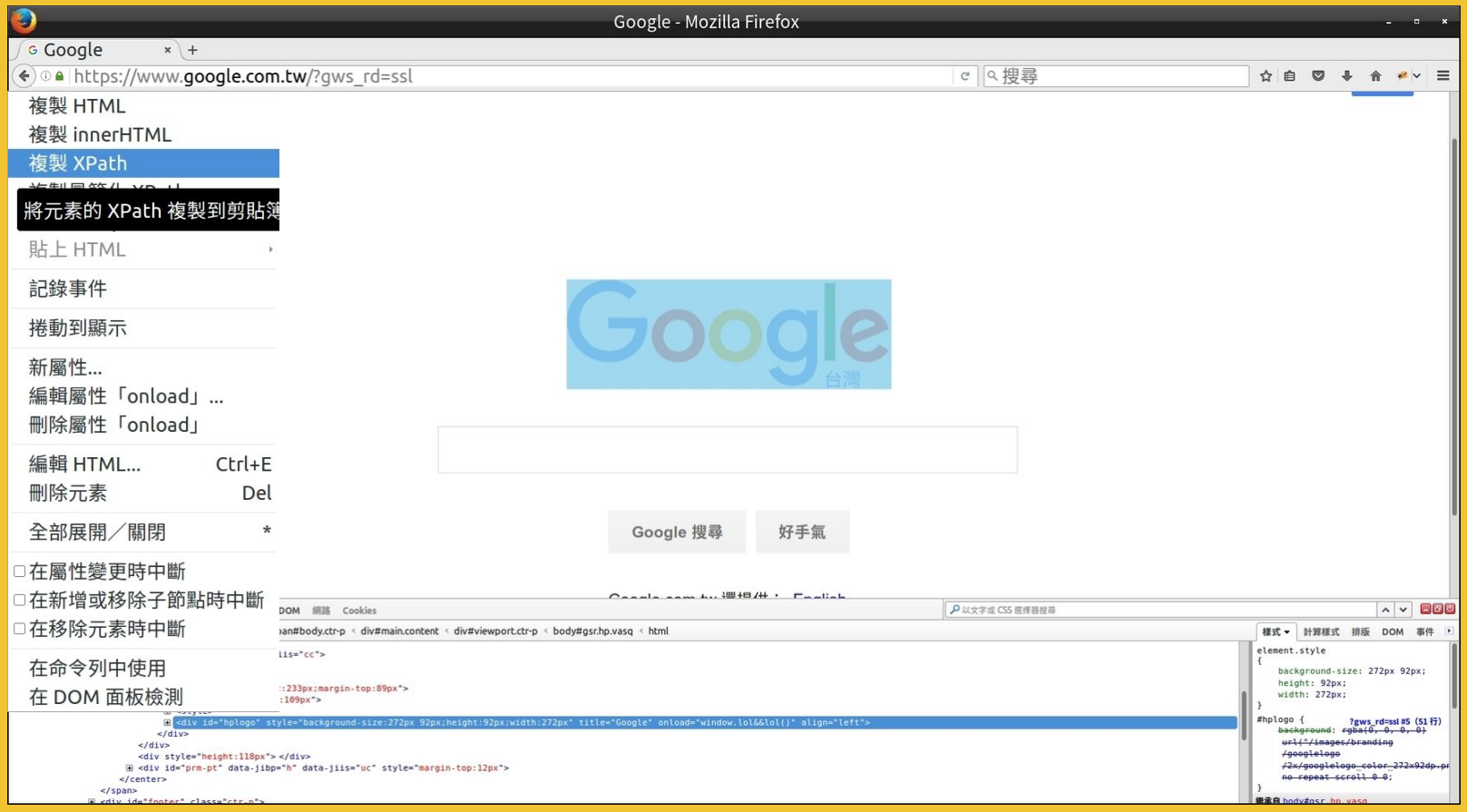
(二) Basic：基本介紹與操作　　1. 網頁分析

Google 瀏覽器 (Chrome) - 開發人員工具



(二) Basic：基本介紹與操作　　1. 網頁分析

Mozilla Firefox 瀏覽器 - **擴充套件 FireBug**

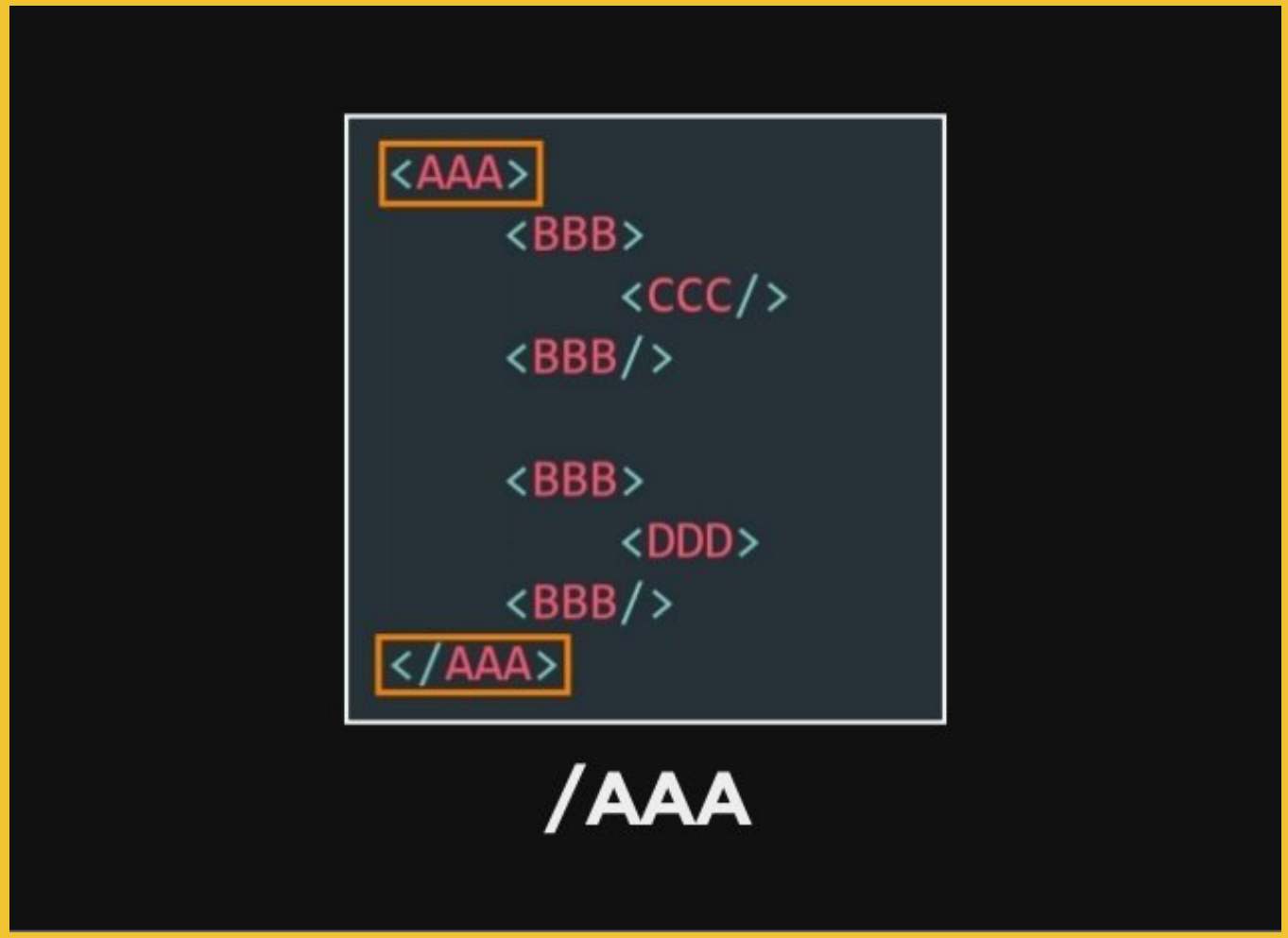


Xpath 概觀

註：本篇內容引用自 *aweimeow* ➡ https://github.com/aweimeow

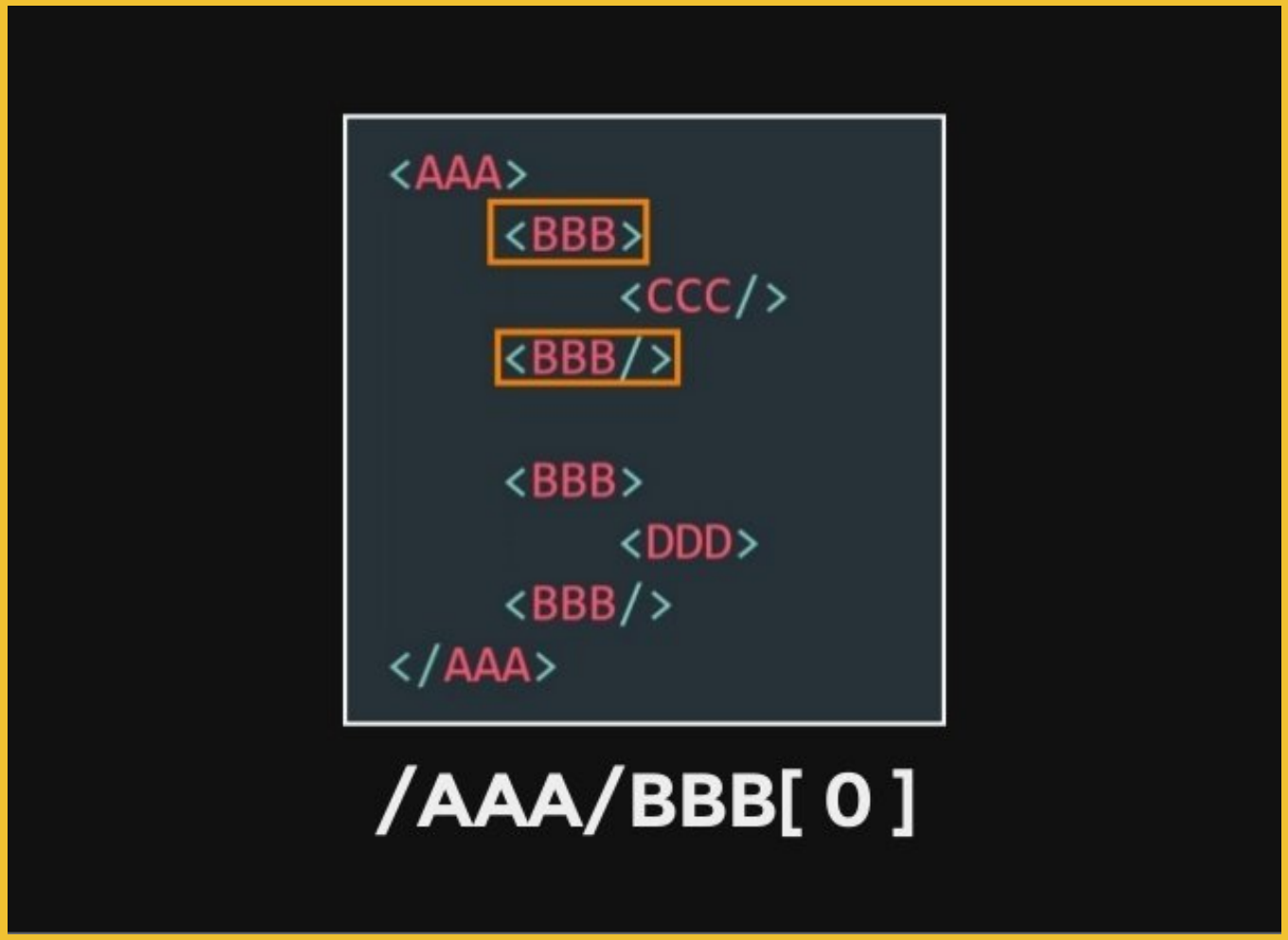
(二) Basic：基本介紹與操作

1. 網頁分析



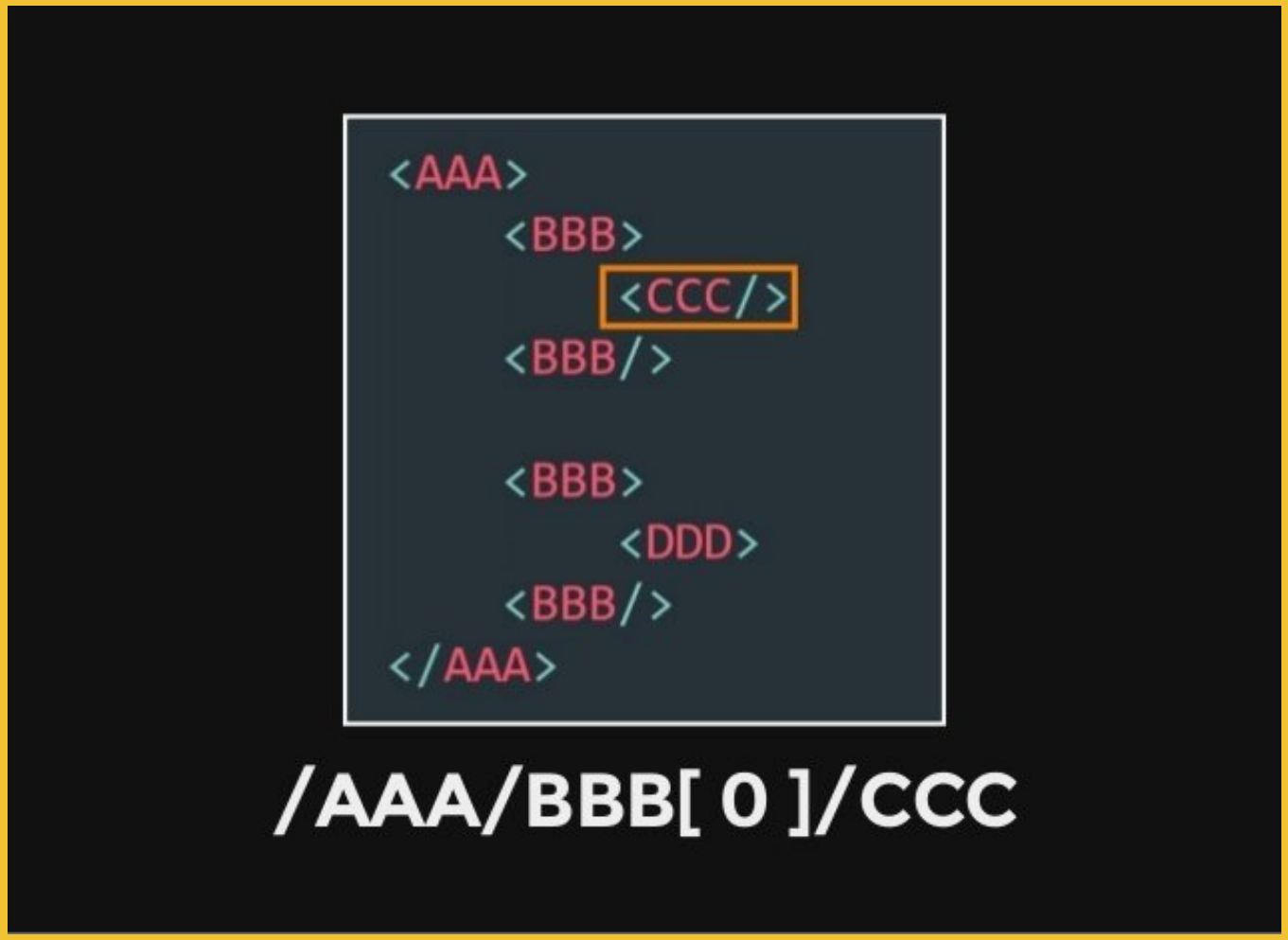
(二) Basic：基本介紹與操作

1. 網頁分析



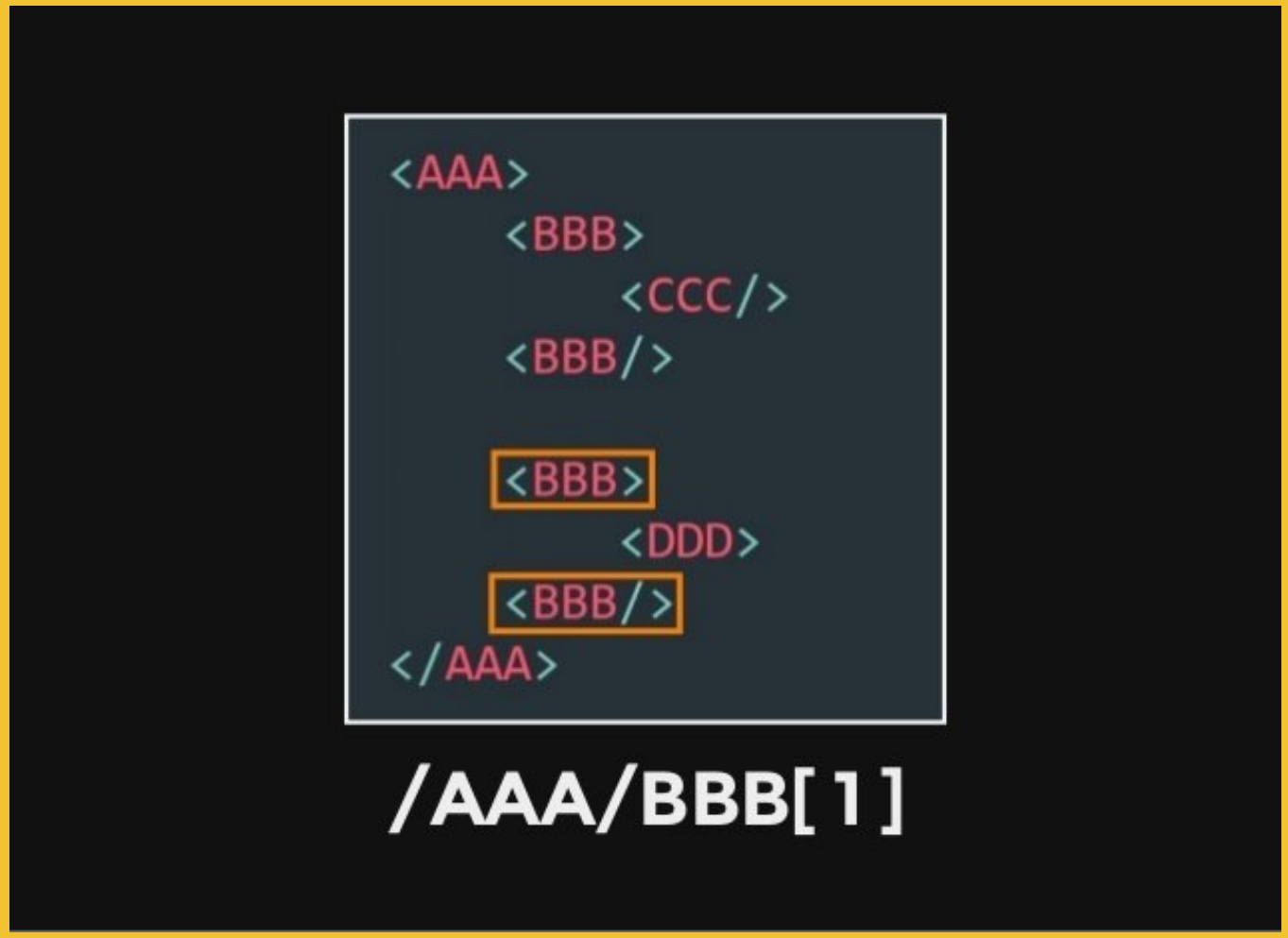
(二) Basic：基本介紹與操作

1. 網頁分析



(二) Basic：基本介紹與操作

1. 網頁分析



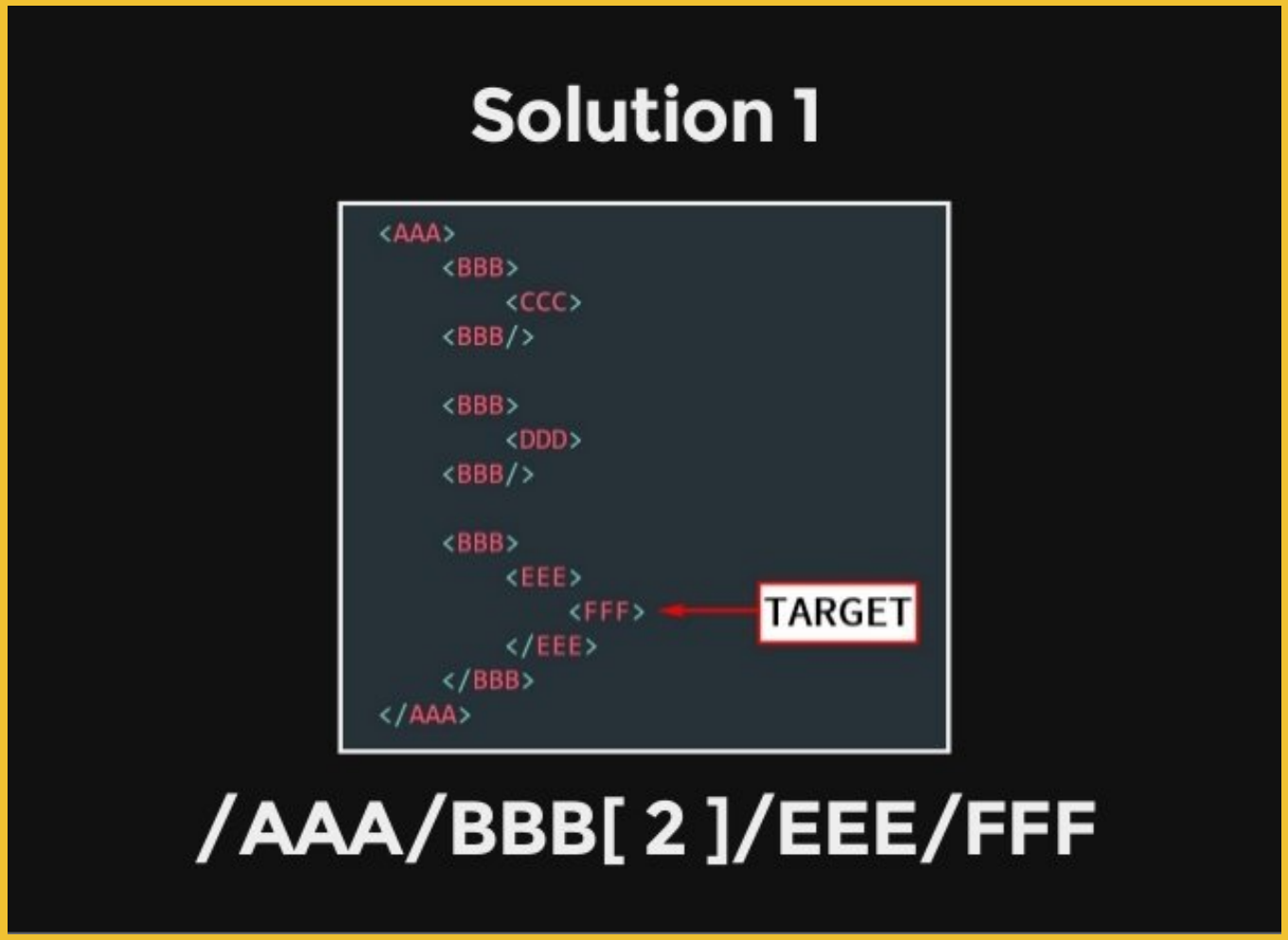
尋找特定物件的 Xpath

(二) Basic：基本介紹與操作

1. 網頁分析

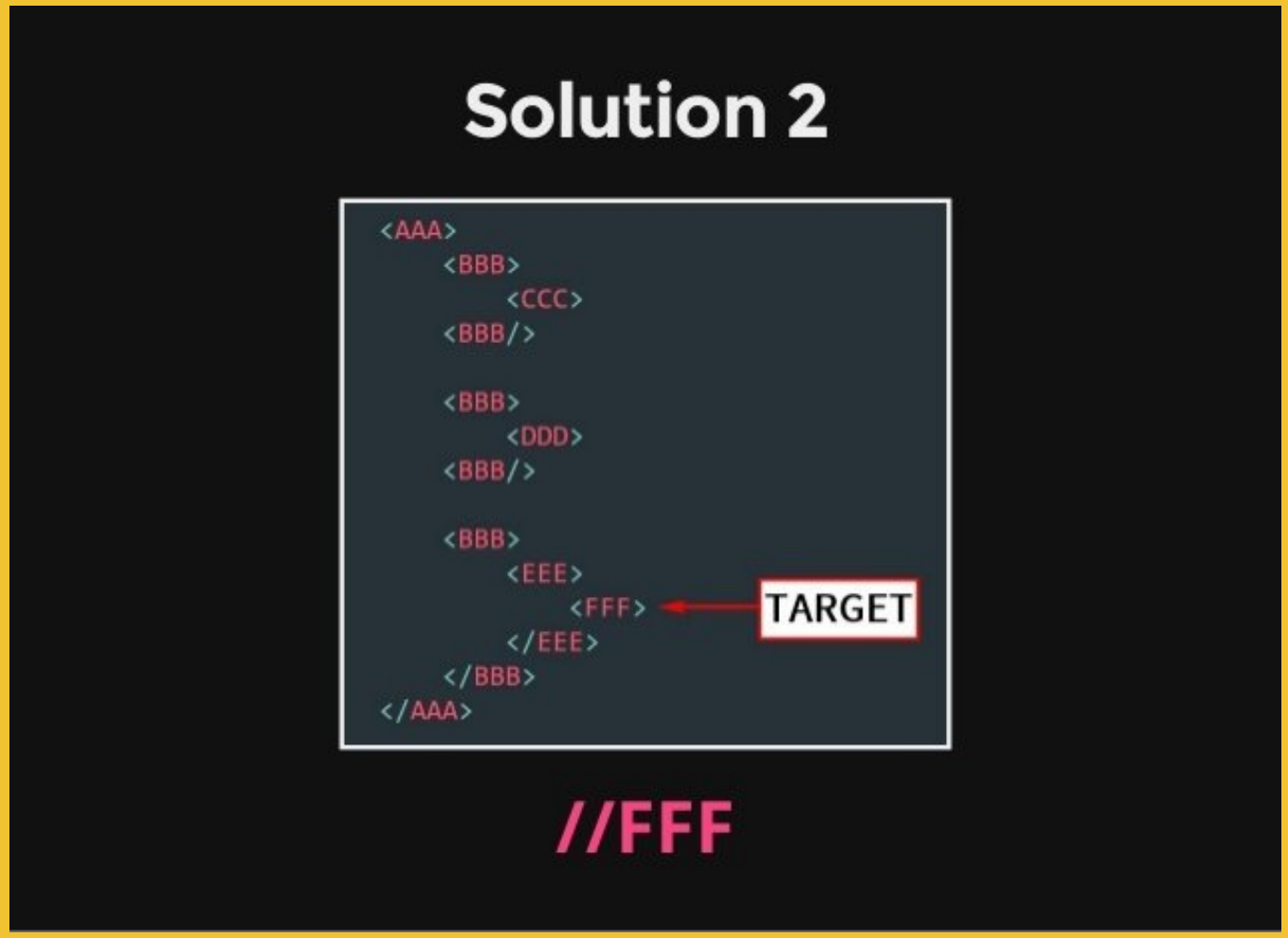
(二) Basic：基本介紹與操作

1. 網頁分析



(二) Basic：基本介紹與操作

1. 網頁分析



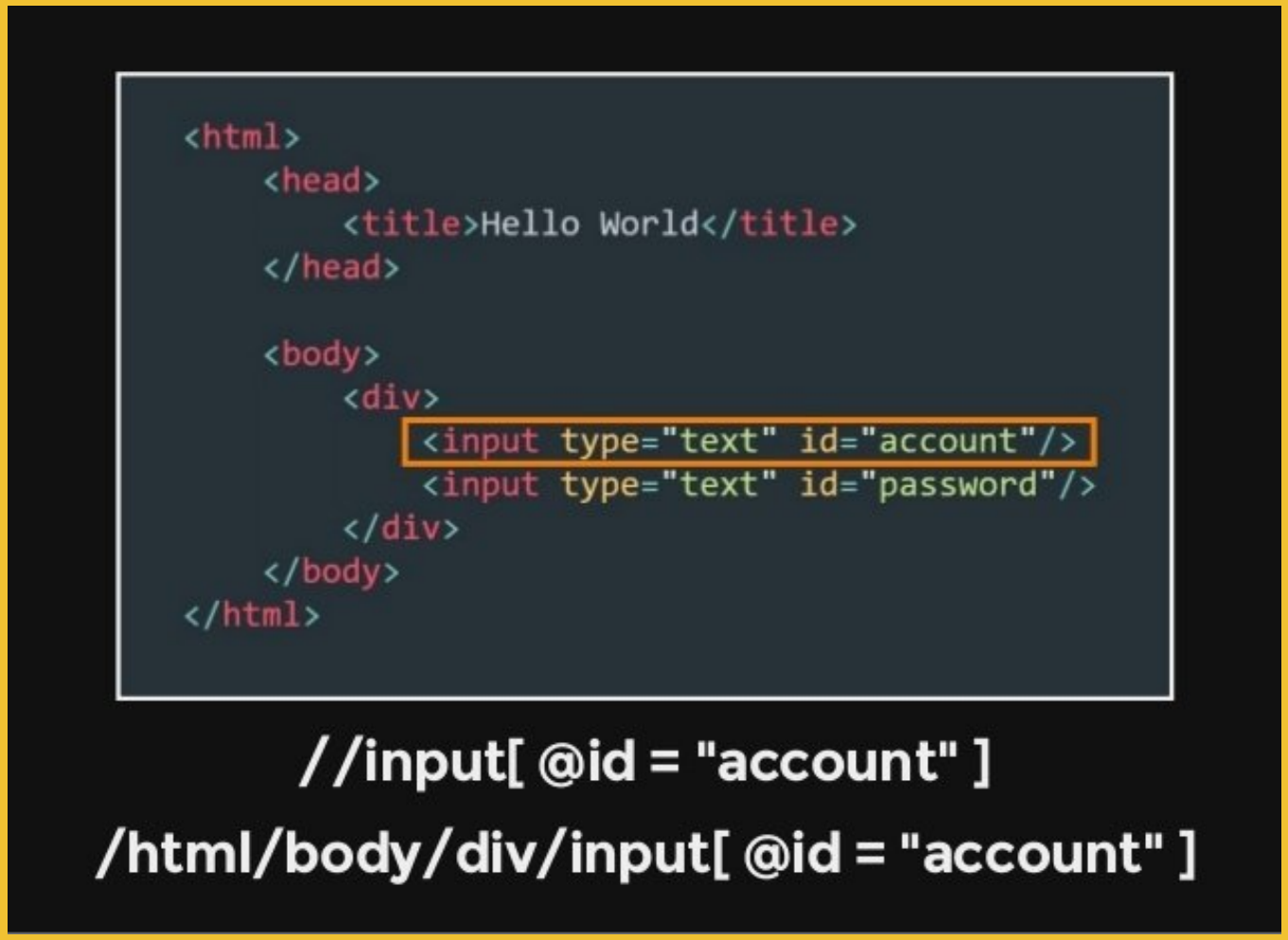
透過 TAG 屬性尋找特定物件

(二) Basic：基本介紹與操作

1. 網頁分析

(二) Basic：基本介紹與操作

1. 網頁分析



2. 資料的讀取

常見的資料格式：

(1) CSV

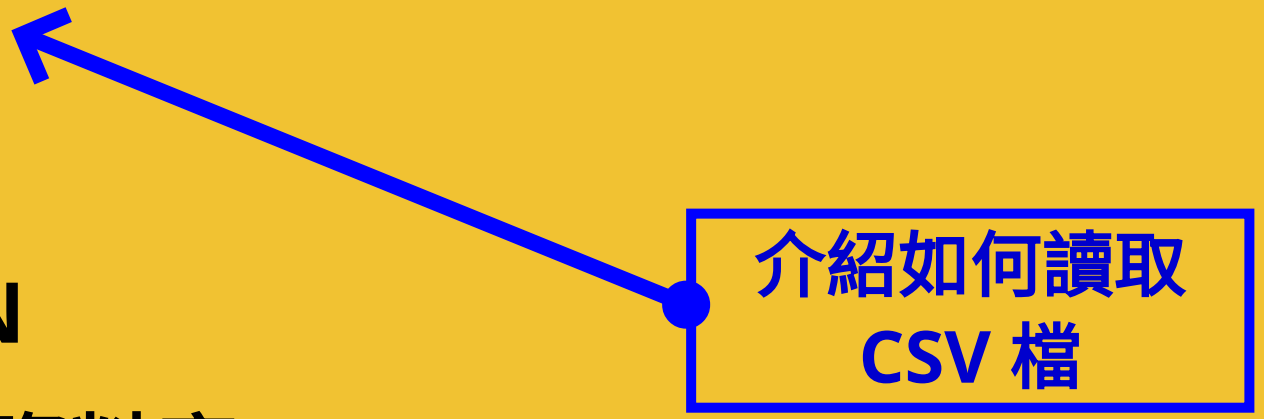
(2) XML

(3) JSON

(4) DB (資料庫)

(5) RData

(6) SPSS、Stata、SAS、Octave ...



(二) Basic：基本介紹與操作　　2. 資料的讀取

讀取CSV：　STEP1：使用read.table()

[用法]

read.table (file = *檔案路徑* **, header =** *TRUE or FALSE*, ***sep*** **= "***分隔符號***")**

[參數設定]

file　*設定檔案的完整路徑*

header　*設定是否將資料的第一橫列設為直行名稱*

sep　*設定用來分隔資料的分隔符號*

```
# 讀取檔案的完整路徑(在此為網路位址)
theUrl <- "http://www.jaredlander.com/data/Tomato%20First.csv"


# 將檔案載入R，在這裡設定
tomato <- read.table ( file = theUrl, header = TRUE, sep = ",")
```

若發現 CSV 檔(或 tab 分隔值檔)內容有缺漏，例如分隔資料格的分隔符號出現在儲存格內。

在這個情況下應該改用 read.csv2() 或 read.delim2() 讀取資料。

讀取CSV：

STEP2：使用head()

[用法]

head(*資料表名稱*)

STEP3：使用data.frame()

[用法]

data.frame (*變數1 = 名稱1*, *變數2 = 名稱2*, *變數3 = 名稱3*,

..... , stringsAsFactors = *TRUE or False*)

[引數設定]

stringsAsFactors *防止含 character (字元)的直行被轉為 factor，保持 character 直行為原有的資料型態*

STEP2、STEP3：執行結果

```
> head(tomato)     # 查看資料表的第一部分
  Round             Tomato Price      Source Sweet Acid Color Texture Overall
1     1         Simpson SM  3.99 Whole Foods   2.8  2.8   3.7     3.4     3.4
2     1  Tuttorosso (blue)  2.99     Pioneer   3.3  2.8   3.4     3.0     2.9
3     1 Tuttorosso (green)  0.99     Pioneer   2.8  2.6   3.3     2.8     2.9
4     1     La Fede SM DOP  3.99   Shop Rite   2.6  2.8   3.0     2.3     2.8
5     2       Cento SM DOP  5.49  D Agostino   3.3  3.1   2.9     2.8     3.1
6     2      Cento Organic  4.99  D Agostino   3.2  2.9   2.9     3.1     2.9
  Avg.of.Totals Total.of.Avg
1          16.1          16.1
2          15.3          15.3
3          14.3          14.3
4          13.4          13.4
5          14.4          15.2
6          15.5          15.1
>
> x <- 10:1
> y <- -4:5
> # "q"是一個 character 型態的向量
> q <- c("Hockey", "Football", "Baseball", "Curling", "Rugby", "Lacrosse", "Basketball",
>
> theDF <- data.frame( First = x, Second = y, Sport = q, stringsAsFactors = FALSE)
> theDF$Sport
 [1] "Hockey"     "Football"   "Baseball"   "Curling"    "Rugby"      "Lacrosse"
 [7] "Basketball" "Tennis"     "Cricket"    "Soccer"
```

(三) Theme：文字雲製作

(三) Theme：文字雲製作

(1) 是"文字探勘"上常用的呈現手法之一

(2) 出現頻率越高的字詞，會加以突顯出來

(3) 比起表格類型的結果，文字雲更美觀

1. 處理英文資料

處理步驟：

STEP 1：準備要分析的資料
STEP 2：安裝和載入所需的套件
STEP 3：進行"文字探勘"
STEP 4：製作"字詞矩陣"
STEP 5：產生"文字雲"

參考教材：Text mining and word cloud fundamentals in R：5 simple steps you should know

➡ 教材網址： https://goo.gl/snM2nZ

(三) Theme：文字雲製作

1. 處理英文資料

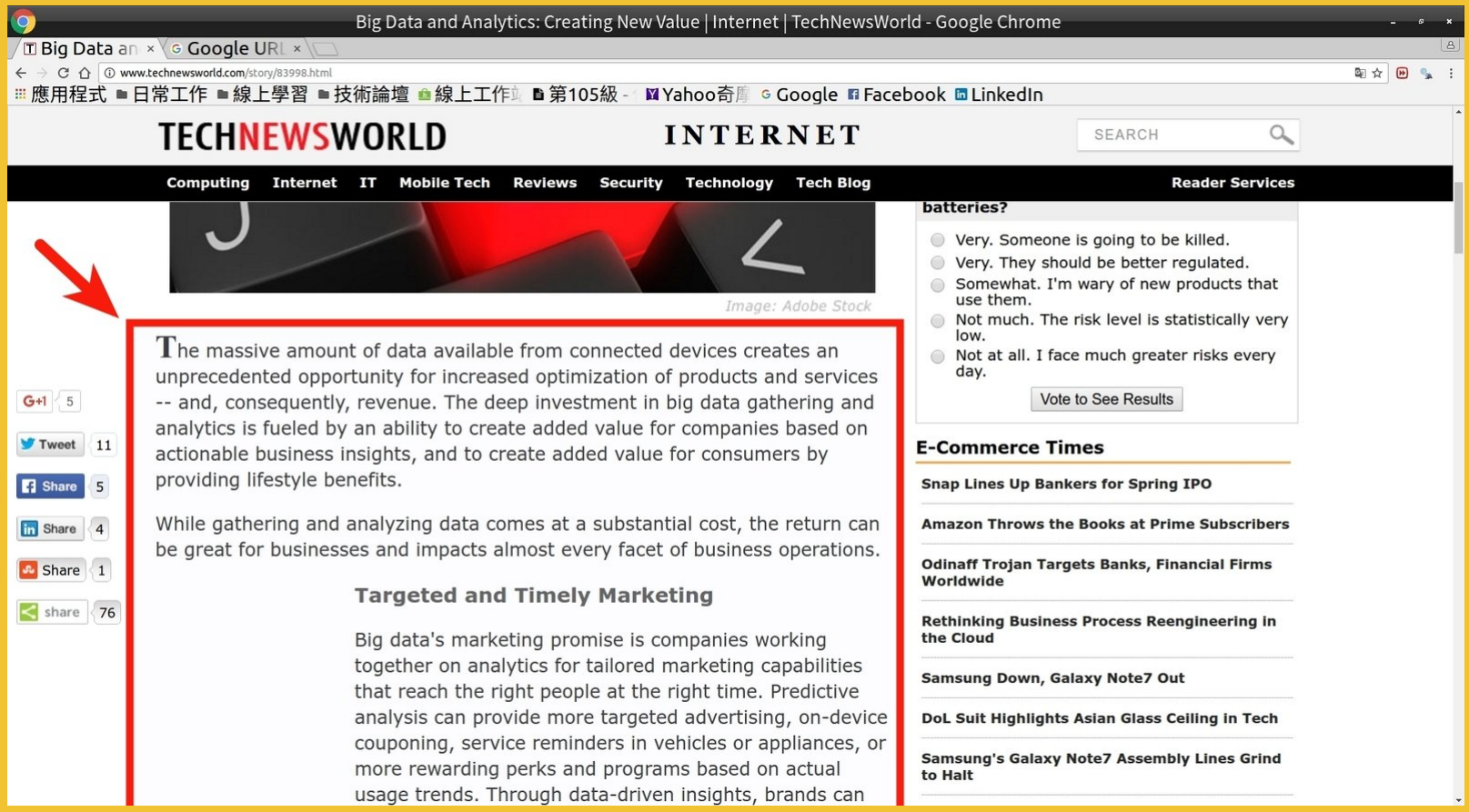
STEP 1：準備要分析的資料

➡ 網址：http://www.technewsworld.com/story/83998.html

(三) Theme：文字雲製作

1. 處理英文資料

STEP 1：準備要分析的資料



紅線圈選區域為本次欲分析之內容

(三) Theme：文字雲製作

1. 處理英文資料

STEP 2：安裝和載入所需的套件

開啟 RStudio，在命令列中輸入以下指令：

```
# 安裝套件
install.packages("rvest")     # "網頁分析"用
install.packages("tm")  # "文字探勘"用
install.packages("SnowballC") # Text stemming
install.packages("wordcloud") # 產生"文字雲"用
install.packages("RColorBrewer") # Color palettes


# 載入套件
library("rvest")
library("tm")
library("SnowballC")
library("wordcloud")
library("RColorBrewer")
```

(三) Theme：文字雲製作

STEP 3：進行"文字探勘"

1. 處理英文資料

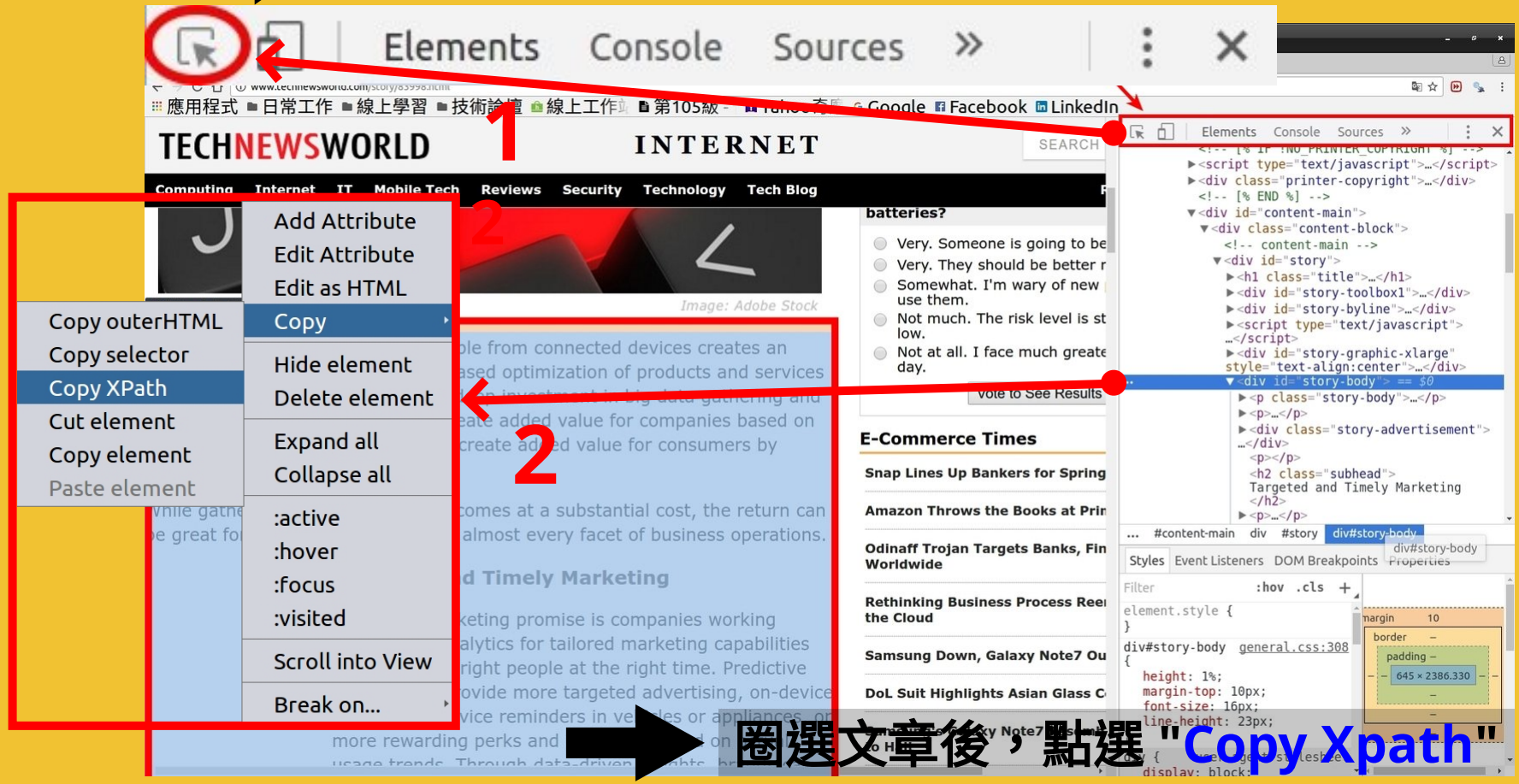
在 Chrome 點選"**開發人員工具**" (亦可按下 "**F12**"鍵)

(三) Theme：文字雲製作

STEP 3：進行"文字探勘"

1. 處理英文資料

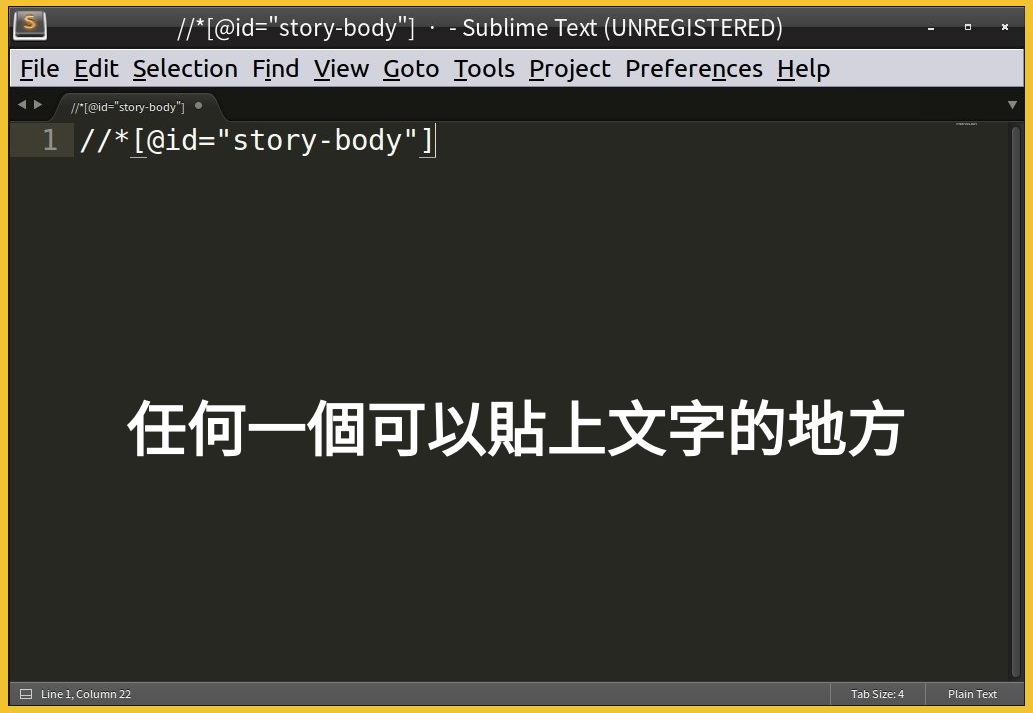
➡ 利用選取工具找到段落後，在對應節點按右鍵



➡ 圈選文章後，點選"Copy Xpath"

STEP 3：進行"文字探勘"

➡ 將取得的 Xpath 貼在記事本上 (稍後會用到)



文章的 Xpath： `//*[@id="story-body"]`

(三) Theme：文字雲製作

1. 處理英文資料

STEP 3：進行"文字探勘"

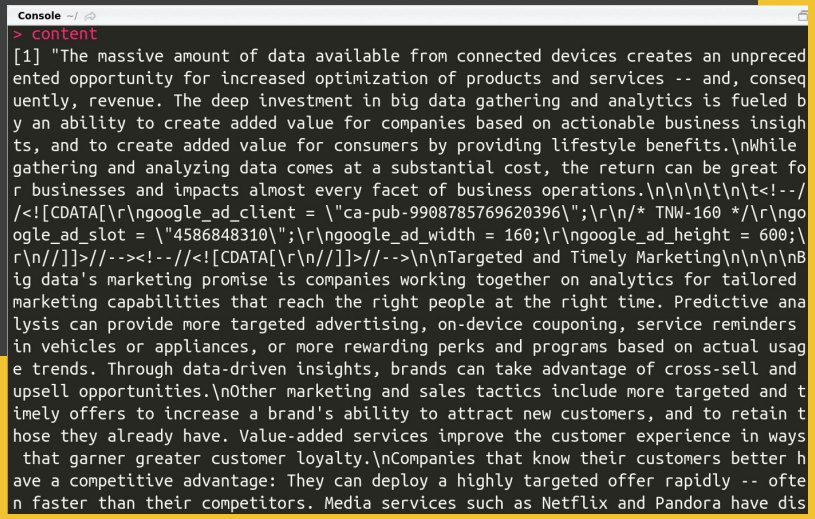
➡ 在命令列中輸入以下指令：

```
# 擷取網頁內容，將網頁下載後存入"source.page"物件
source.page <- read_html("http://www.technewsworld.com/story/83998.html")

# 利用 Xpath 取得文章內容
source.content <- html_nodes(source.page, xpath =  '//*[@id="story-body"]')

# 取得 HTML 中的文字資料
content <- html_text(source.content)

# 顯示資料 (此時文章仍包含多餘字元)
content
```



(三) Theme：文字雲製作

1. 處理英文資料

STEP 3：進行"文字探勘"

➡ 在命令列中輸入以下指令：

```
# 將內容以"語料庫"的形式儲存
docs <- Corpus(VectorSource(content))


# 檢查內容
inspect(docs)
```

```
Console ~/
> inspect(docs)
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 1

[[1]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 5860
```

(三) Theme：文字雲製作

1. 處理英文資料

STEP 3：進行"文字探勘"

過濾特殊字元：在命令列中輸入以下指令，將特殊字元以"空白"取代

```
# 將特殊的字元以"空白"取代
toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))

docs <- tm_map(docs, toSpace, "/")     # 將"/"以"空白"取代
docs <- tm_map(docs, toSpace, "@")     # 將"@"以"空白"取代
docs <- tm_map(docs, toSpace, "\\|")   # 將"\\|"以"空白"取代
```

過濾贅詞、符號：在命令列中輸入以下指令，移除贅詞和多餘的符號

```
# Convert the text to lower case
docs <- tm_map(docs, content_transformer(tolower))
docs <- tm_map(docs, removeNumbers)     # 移除數字

# 移除常見的"轉折詞彙"
docs <- tm_map(docs, removeWords, stopwords("english"))
docs <- tm_map(docs, removePunctuation)  # 移除標點符號
docs <- tm_map(docs, stripWhitespace)    # 移除額外的"空白"
```

(三) Theme：文字雲製作

1. 處理英文資料

STEP 4：製作"字詞矩陣"

➡ 在命令列中輸入以下指令：

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


# 顯示前10個出現頻率最高的字詞
head(d, 10)
```

Console ~/

```
> head(d, 10)
                word freq
data            data   22
service      service   11
can              can    9
companies companies    9
customer   customer    9
business   business    8
product     product    8
analytics analytics    6
marketing marketing    6
models       models    5
```

(三) Theme：文字雲製作

1. 處理英文資料

STEP 5：產生"文字雲"

➡ 在命令列中輸入以下指令：

```
# 設定可重複的亂數序列
set.seed(1000)


# 製作文字雲
wordcloud(words = d$word, freq = d$freq, min.freq =  2,
          max.words = 30, random.order = FALSE, rot.per = 0.35,
          colors = brewer.pal( 8, "Dark2"))
```

2. 處理中文資料

處理步驟：

STEP 1：準備要分析的資料
STEP 2：安裝和載入所需的套件
STEP 3：進行"文字探勘"
STEP 4：製作"字詞矩陣"
STEP 5：產生"文字雲"

參考教材：文字資料探勘實作

➡ 教材網址： http://andrew.ga/works/TextMining/

(三) Theme：文字雲製作　　2. 處理中文資料

STEP 1：準備要分析的資料

➡ 網址：http://www.appledaily.com.tw/realtimenews/article/life/20161016/968938/

(三) Theme：文字雲製作

2. 處理中文資料

STEP 1：準備要分析的資料

紅線圈選區域為本次欲分析之內容

STEP 2：安裝和載入所需的套件

開啟 RStudio，在命令列中輸入以下指令：

```
# 安裝套件
install.packages("rvest")       # "網頁分析"用
install.packages("jiebaR")      # "中文斷詞"用
install.packages("tm")          # "文字探勘"用
install.packages("wordcloud2")  # 產生"文字雲"用


# 載入套件
library("rvest")
library("jiebaR")
library("tm")
library("wordcloud2")
```

(三) Theme：文字雲製作

2. 處理中文資料

STEP 3：進行"文字探勘"

➡ 在 Chrome 點選"**開發人員工具**" (亦可按下 "**F12**"鍵)

(三) Theme：文字雲製作

2. 處理中文資料

STEP 3：進行"文字探勘"

➡ 利用選取工具找到段落後，在對應節點按右鍵

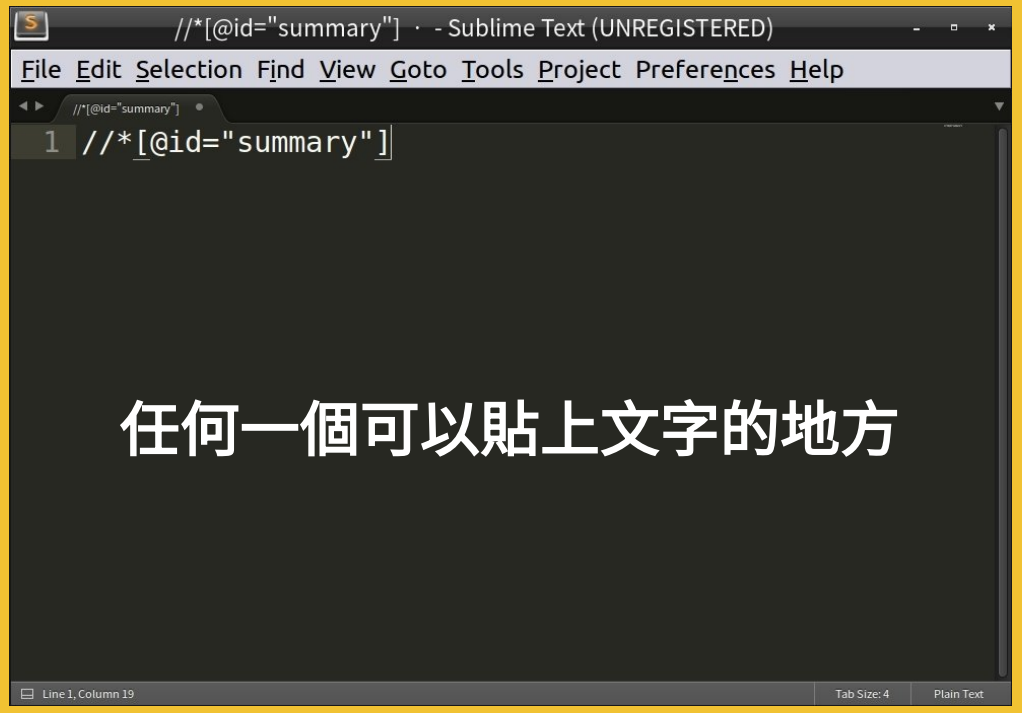
➡ 圈選文章後，點選"Copy Xpath"

(三) Theme：文字雲製作

2. 處理中文資料

STEP 3：進行"文字探勘"

➡ 將取得的 Xpath 貼在記事本上 (稍後會用到)



文章的 Xpath： `//*[@id="summary"]`

(三) Theme：文字雲製作

2. 處理中文資料

STEP 3：進行"文字探勘"

➡ 在命令列中輸入以下指令：

```
# 擷取網頁內容，將網頁下載後存入"source.page"物件
source.page <- read_html( "http://www.appledaily.com.tw/realtimenews/article/life/20161016/968938/"  )

# 利用 Xpath 取得文章內容
source.content <- html_nodes(source.page, xpath =  '//*[@id="summary"]' )

# 取得 HTML 中的文字資料
content <- html_text(source.content)

# 顯示資料 (此時文章仍包含多餘字元)
content

# 啟用 jiebaR 套件裡的斷詞引擎
mixseg = worker()
content.vec <- segment(code = content, jiebar = mixseg)
```

```
Console ~/
> content
[1] "(新增配音影片)雙颱接力擾台，未來一周有雨。21號颱風莎莉佳昨增強為中度颱風，預估今天穿越菲律賓呂宋島後將進入南海。中央氣象局預報員吳依帆指出，明起受莎莉佳外圍雲系影響，東半部地區有局部大雨，北部、南部也有陣雨。莎莉佳將影響至周二，周三雨勢趨緩，周四起受22號颱風海馬外圍雲系影響，又會開始下雨。海馬今晨距離台灣2000多公里，朝西北方向前進，其路徑較莎莉佳偏北，氣象局研判可能通過呂宋島北方或巴士海峽一帶，不排除更靠近台灣，影響程度有待觀察。據中央氣象局下午14時預報指出，22號颱風海馬增強為中度颱風，目前位於北緯 10.4 度，東經 139.2 度，以每小時17轉22公里速度，向西北轉西北西進行。颱風中心氣壓 972 百帕，近中心最大風速每秒 33 公尺，瞬間之最大陣風每秒 43 公尺，七級風半徑 150 公里，十級風半徑 50 公里，預估預估它的外圍雲系將在周四、五影響台灣天氣，屆時各地降雨機率逐漸增加，東半部、南部地區及北部山區有短暫陣雨，其他地方也會有局部短暫陣雨。但實際影響台灣的時間及程度，仍需視其路徑及發展，不排除未來有機會進一步增強為強烈颱風的可能性。（王嘉慶、即時新聞中心／台北報導）出版時間 07：45更新時間 19：00"
```

(三) Theme：文字雲製作

2. 處理中文資料

STEP 3：進行"文字探勘"

➡ 在命令列中輸入以下指令：

```
space_tokenizer = function(x){
    unlist(strsplit(as.character(x[[ 1]]), '[[:space:]]+'))
}

jieba_tokenizer = function(d){
    unlist(segment(d[[ 1]], mixseg))
}

# 撰寫 CNCorpus 副程式
#### CNCorpus Function Start ####
CNCorpus = function(d.vec){

    doc <- VCorpus(VectorSource(d.vec))
    doc <- unlist(tm_map(doc ,jieba_tokenizer), recursive =  F)
    doc <- lapply(doc , function(d)paste(d, collapse = ' '))
    Corpus(VectorSource(doc))
}
#### CNCorpus Function END ####
```

CNCorpus 副程式：將內容以"語料庫"的形式儲存

(三) Theme：文字雲製作

2. 處理中文資料

STEP 4：製作"字詞矩陣"

➡ 在命令列中輸入以下指令：

```
content.corpus = CNCorpus(list(content.vec))     # 執行 CNCorpus 副程式
content.corpus <- tm_map(content.corpus, removeNumbers)     # 移除數字

control.list = list(wordLengths = c(2, Inf),tokenize = space_tokenizer)
content.dtm <- DocumentTermMatrix(content.corpus, control = control.list)


inspect(content.dtm)     # 檢查內容
```

```
Console ~/
> inspect(content.dtm)    # 檢查內容
<<DocumentTermMatrix (documents: 1, terms: 116)>>
Non-/sparse entries: 116/0
Sparsity           : 0%
Maximal term length: 5
Weighting          : term frequency (tf)

    Terms
Docs 一周 一帶 七級風 下午 下雨 中央氣象局 中度颱風 中心 今天 今晨 位於 偏北
   1    1    1      1    1    1          2        2    3    1    1    1    1
    Terms
Docs 公尺 公里 其他 出版 前進 北方 北緯 北部 十級 半徑 南海 南部 即時新聞
   1    2    3    1    1    1    1    1    2    1    2    1    2        1
    Terms
Docs 受莎莉佳 可能 可能性 台北 台灣 各地 吳依帆 呂宋島 周三 周二 周四 地區 地方
   1        1    1      1    1    4    1      1      2    1    1    2    2    1
    Terms
Docs 報導 增加 增強 外圍 多公里 大雨 天氣 實際 將在 小時 局部 屆時 山區 巴士海峽
   1    1    1    3    3      1    1    1    1    1    1    2    1    1        1
    Terms
Docs 強烈颱風 影片 影響 後將 指出 排除 接力 擾台 新增 方向 明起 時間 更新 最大
   1        1    1    6    1    2    2    1    1    1    1    1    3    1    2
```

(三) Theme：文字雲製作

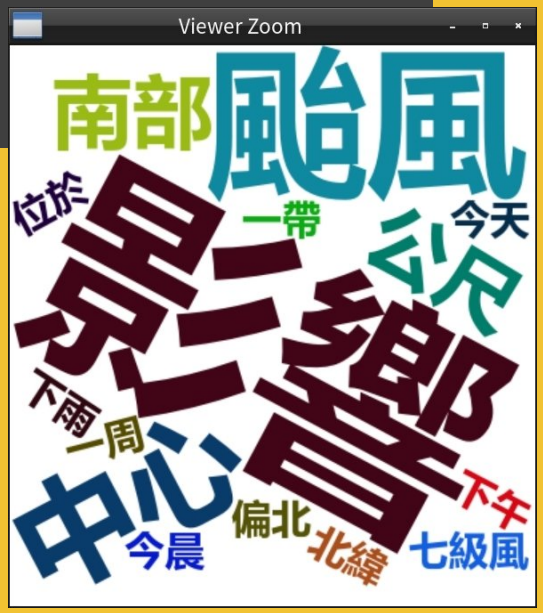
2. 處理中文資料

STEP 5：產生"文字雲"

➡ 在命令列中輸入以下指令：

```
frequency <- colSums(as.matrix(content.dtm))
frequency <- sort(frequency, decreasing = TRUE)[1:100]

wordcloud2(as.table(frequency), fontFamily = '微软雅黑', shape = 'star')
```



(四) Reference：學習資源

(四) Reference：學習資源

線上教材

一、中文教材

1. **R 語言翻轉教室** - Wush Wu、Chih Cheng Liang、Johnson Hsieh
   http://datascienceandr.org/

2. **手把手教你 R 語言資料分析實務** - 張毓倫&陳柏亨
   http://goo.gl/18mwug

3. **R 軟體與資料探勘之開發與應用** - 陳志華
   https://goo.gl/NPdzzP

二、英文教材

1. **DataCamp**
   https://www.datacamp.com/

2. **R for Data Science**
   http://r4ds.had.co.nz/

(四) Reference：學習資源

推薦書籍

R 軟體資料分析基礎與應用

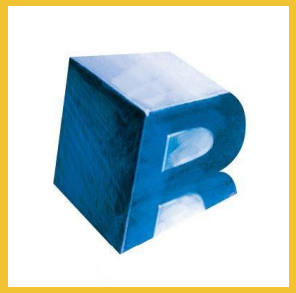
作者： Jared P. Lander
譯者：鍾振蔚
出版社：旗標

(四) Reference：學習資源

相關社群

台灣資料科學年會

https://www.facebook.com/twdsconf/



Taiwan R User Group

https://www.facebook.com/Tw.R.User/

資料視覺化 / Data Visualization

https://www.facebook.com/data.visualize/

Q & A