|  |
| --- |
| 作品名稱：文字雲新聞與情分析 |
| 一、說明 |
| 本專題主要使用wordcloud透過Python、R程式及SQLServer軟體及搭配新聞文字進行字詞拆分繪製出文字雲圖，先在新聞網站上下載文字然後儲存於記事本上。使用Python執行文字雲前結果可以搭配png圖片產出文字雲效果、使用R執行文字雲前結果可以打上數字、英文或符號產出文字雲效果。 |
| 二、論文 |
| 根據標籤雲的作用(而非樣式)，在應用中可以將其分成三大類。其中，第一類用於描述網站中的每個獨立條目，而第二類則著力於從整體上刻畫網站所有條目的標籤情況：  第一類標籤云：每一個條目都有自己獨立的標籤雲，標籤字體越大，此條目中用戶使用過這個標籤的次數就越多，在頁面公開統計點擊且不要求精準數據的情況下十分適用。如Last.fm就是使用了第一類標籤雲。  第二類標籤云：網站一般會有一個超大型標籤雲，標籤字體越大，網站裡使用過這個標籤的條目數就越多。第二類標籤雲可以顯示出標籤的熱門程度，在實際應用中更為常見，如flickr。  第三類標籤云：在此類中，標籤作為一個數據項目的工具，用於表示在整個集合中里各個項目的數據量的大小。  從廣義來說，相似的可視化技術並不限於用於標籤雲，例如還可以用在文字雲或數據雲上。    大多數標籤雲通過內嵌HTML元素的方式實現。標籤可以按字母次序、重要次序、隨機次序等方式排列。部分網站還應用語義分組技術，讓內容相聯的標籤聚攏在一起。而啟發法更可以用於壓縮標籤雲的體積，幫助標籤分組。 |
| 三、實作 |
| 1. R語言繪製文字雲：   　　使用Rsutdio軟體的wordcloud及wordcloud2的套件繪製成文字雲圖，然後另存為圖片。ch\_news = c("請匯入新聞文字")需要再網站或記事本內找相關新聞文字複製並貼在"請匯入新聞文字"上面。  程式碼:   |  | | --- | | library(jiebaR)  library(wordcloud) # 第一代文字雲  library(wordcloud2) # 第二代文字雲  library(RColorBrewer)  ch\_news = c("請匯入新聞文字") # 匯入新聞  new\_terms = c("鄉鎮縣市", "民意代表", "店家", "買賣交屋")  writeLines(new\_terms, '/Users/hello/Documents/R/ch\_terms.txt') # 自訂字典,請輸入自己的路徑  stopwords = c( "在","的","上", "下","是", "個","來","為","亦","或", "之", "與", "於", "用", "都", "等", "日", "月", "年", "週", "嗎", "以", "就", "但", "及", "也", "了", "要", "不", "會", "和", "對", "著", "後", "她", "他")  writeLines(stopwords, '/Users/hello/Documents/R/ch\_stopwords.txt') # 自訂停用詞,請輸入自己的路徑  cutter = worker(user='/Users/hello/Documents/R/ch\_terms.txt', stop\_word = '/Users/hello/Documents/R/ch\_stopwords.txt') # 引用字典和停用詞,請輸入自己的路徑  ch\_news <- gsub("[0-9a-zA-Z]+?", "", ch\_news) # 刪除數字和字母  ch\_news <- cutter[ch\_news] # 斷詞  freq\_ch <- sort(table(ch\_news), T)  freq\_ch = as.data.frame(freq\_ch)  colnames(freq\_ch) <- c("Words", "Freq") # 字頻表  head(freq\_ch) # 查看前10筆資料  par(family=("NotoSansCJKtc-Medium")) # 設定字體 Mac  customed\_colors = c("#000080", "#ffff00", "#6495ed", "#00bfff", "#87cefa", "#db7093", "#ba55d3", "#b22222", "#008080", "#ff8c00", "#6b8e23") # 顏色  ch\_wordcloud = wordcloud(freq\_ch$Words, freq\_ch$Freq, min.freq = 2, random.order = F, ordered.colors = F, colors = customed\_colors); ch\_wordcloud # 第一代文字雲  ch\_wordcloud2 = wordcloud2(freq\_ch, size = 1.3, color = customed\_colors, backgroundColor="white"); ch\_wordcloud2  # 第二代文字雲  letterCloud(freq\_ch,'OK',size = 0.50)  letterCloud(freq\_ch,'R',size = 0.50) |   執行結果:  　　繪製好的文字雲圖將存成圖檔:請點選Export/Save as an image file/請選好路徑及命名/save即可。(執行文字符號時出現文字雲結果反應很慢甚至沒出現文字雲需要多式幾次就會成功)   |  | | --- | |  |  1. Python繪製文字雲：   使用Python軟體的wordcloud的套件繪製成文字雲圖，執行後會自動另存為圖片。  程式碼:   |  | | --- | | #程式  %matplotlib inline  import matplotlib.pyplot as plt  from wordcloud import WordCloud  from PIL import Image  import jieba  import numpy as np  # 如果檔案內有一些編碼錯誤，使用 errors='ignore' 來忽略錯誤  with open("財經\_房地產\_1.txt", encoding="Utf-8", errors='ignore') as f:  text = f.read()  # 設定使用 big5 斷詞  jieba.set\_dictionary('./dict.txt.big')  wordlist = jieba.cut(text)  words = " ".join(wordlist)  #文字雲造型圖片  mask = np.array(Image.open('pythonM 拷貝.jpg')) #文字雲形狀  # 從 Google 下載的中文字型  font = 'SourceHanSansTW-Regular.otf'  #背景顏色預設黑色，改為白色、使用指定圖形、使用指定字體  my\_wordcloud = WordCloud(background\_color='white',mask=mask,font\_path=font).generate(words)  plt.imshow(my\_wordcloud)  plt.axis("off")  plt.show()  #存檔  my\_wordcloud.to\_file('pythonword\_cloud.png') |   執行結果前:   |  | | --- | |  |   執行結果後:  　　背景一定要是純白色的才會配合黑字或黑圖演算法走，請不要用透明背景只有黑字或黑圖來執行出文字雲，這樣子會把整個文字圖給壟罩起來(天上少有，地下難尋)。   |  | | --- | |  |  1. SQLServer繪製文字雲：   程式碼:  　　使用SQL Server Management Studio(SSMS)內建Python套件(jieba、wordcloud套件)，使用R語言及Python繪製成文字雲圖，執行後會自動另存為圖片。   |  | | --- | | WITH Temp  AS  (  SELECT [NewsId] FROM [News] WHERE SUBSTRING([Label],1,1)='1'  ) #請確定好新聞ID  ,Temp2  AS  (  SELECT A.[NewsId],B.KeyWords,B.Cnts  FROM [Temp] AS A CROSS APPLY(SELECT \* FROM dbo.GetKeyWords(A.NewsId)) AS B  )  SELECT TOP(300) [KeyWords],SUM([Cnts]) AS Cnts  FROM Temp2  GROUP BY [KeyWords]  ORDER BY [Cnts] DESC  --------------------------------------------  EXEC sp\_execute\_external\_script  @language = N'Python'  , @script = N'  import matplotlib.pyplot as plt  from wordcloud import WordCloud  import pandas as pd  input\_data = InputDataSet  df=pd.DataFrame(input\_data)  nn = list(df.KeyWords)  vv = list(df.Cnts)  my\_dict=dict(zip(nn,vv))  #設定中文字體  font\_path = "C:\Windows\Fonts\kaiu.ttf"  wc = WordCloud(font\_path = font\_path,width = 800, height = 800,background\_color ="white",stopwords=None,min\_font\_size = 10)  wc.generate\_from\_frequencies(my\_dict)  wc.to\_file("C:\\DD\\p3.png")  '  , @input\_data\_1 = N'  WITH Temp  AS  (  SELECT [NewsId] FROM [News] WHERE SUBSTRING([Label],1,1)=''1''  ) #請確定好新聞ID  ,Temp2  AS  (  SELECT A.[NewsId],B.KeyWords,B.Cnts  FROM [Temp] AS A CROSS APPLY(SELECT \* FROM dbo.GetKeyWords(A.NewsId)) AS B  )  SELECT TOP(300) [KeyWords],SUM([Cnts]) AS Cnts  FROM Temp2  GROUP BY [KeyWords]  ORDER BY [Cnts] DESC  ' |   執行結果:  　　繪製好的文字雲圖目前只有單調的形狀，畢竟它也是結果也已經很不錯了。   |  | | --- | |  | |
| 四、結論 |
| 使用WordCloud繪製出文字雲前1.R語言要重頭開始執行到結尾將複製好的新聞文字貼在指定的位置然後就可以繪製成文字雲圖。它是用文字、數字、標點符號繪製成文字雲圖形狀。2.Python比較單純；將圖片、記事本等要放同一個資料夾及相同路徑就直接執行一整個程式碼就可以繪製成文字雲圖。用純白背景及黑字或黑圖繪製成文字雲圖形狀。3.SQLServer必須要跟Python、R關聯且下載執行套件，然後開啟SSMS軟體建立新聞資料庫，先上傳新聞資料再清洗新聞資料及檢查，最後繪製成文字雲圖。目前只有單調的形狀，畢竟它也是結果也已經很不錯了。  文字雲是可視化的文檔用詞頻率統計權重表，這項技術最近常用於具體化、形象化政治演講的話題和內容。 |
| 五、參考 |
| 好學校Hahow R語言系列課程。  資展國際R語言系列課程。  巨匠電腦Azure for ML及R語言資料分析。 |