國防大學政治作戰學院
「文字探勘與社會科學應用工作坊」
前置作業須知

主　講 人：倪世傑
課程時間：2024/06-07
主辦單位：政治作戰學院大數據研究中心
地　　點：復興崗

課程前置作業

我們在課程進行前，需要安裝工作環境，同時，測試這個工作環境能否正常運作。
我們的工作環境需要以下幾項軟體：

下載位置
#R
## R for Mac https://cran.r-project.org/bin/macosx/ (注意：區分處理器M版本與老的Intel版本)
## R for Windows https://cran.r-project.org/bin/windows/base/

#Rstudio
##Rstudio download for Mac　　https://posit.co/download/rstudio-desktop/
##Rstudio download for Windows https://posit.co/download/rstudio-desktop/

#sublime text 4
#https://www.sublimetext.com/download
##需要注意的是，Mac、Windows、Linux系統要下載不同的版本。
##現在比較流行的是開源代碼編輯器是VS Code
VS Code(Visual Studio Code)https://code.visualstudio.com/.

請各位參與者務必下載並運行之，確保每一項都能夠運作。
我簡單地安排了兩項工作，請各位務必要進行，當作對系統安裝的測試。
在設定的部分需要各位參與者進行的原因在於如果設定出了狀況，我們可以通過Line群組，在課程進行前先解決，節省下寶貴的課程函授時間。
RStudio是我們未來最主要的工作區域，RStudio也有線上版本RStudio Cloud，只要使用Google , Apple的帳號都可以使用免費提供的空間，可資利用。(☠機敏性質的文件請勿上傳網路空間以免觸法並危及國安☠)

1.設定：設定您未來的工作路徑(working directory)。
## 1_1 設定路徑：你的工作區在哪裡？ setting yr working directory.
需要事先在Rstudio中設定工作路徑。
但在設定之前，我們先在「文件」（Documents）中設立一個資料夾，我是預設為一個叫做AIHS的資料夾，
我們之後所有需要用的檔案，都會在這個資料夾儲存與輸出。
資料夾絕對不要設定在「桌面」或「下載項目」，請設定在「文件」。設定工作區域的資料夾：

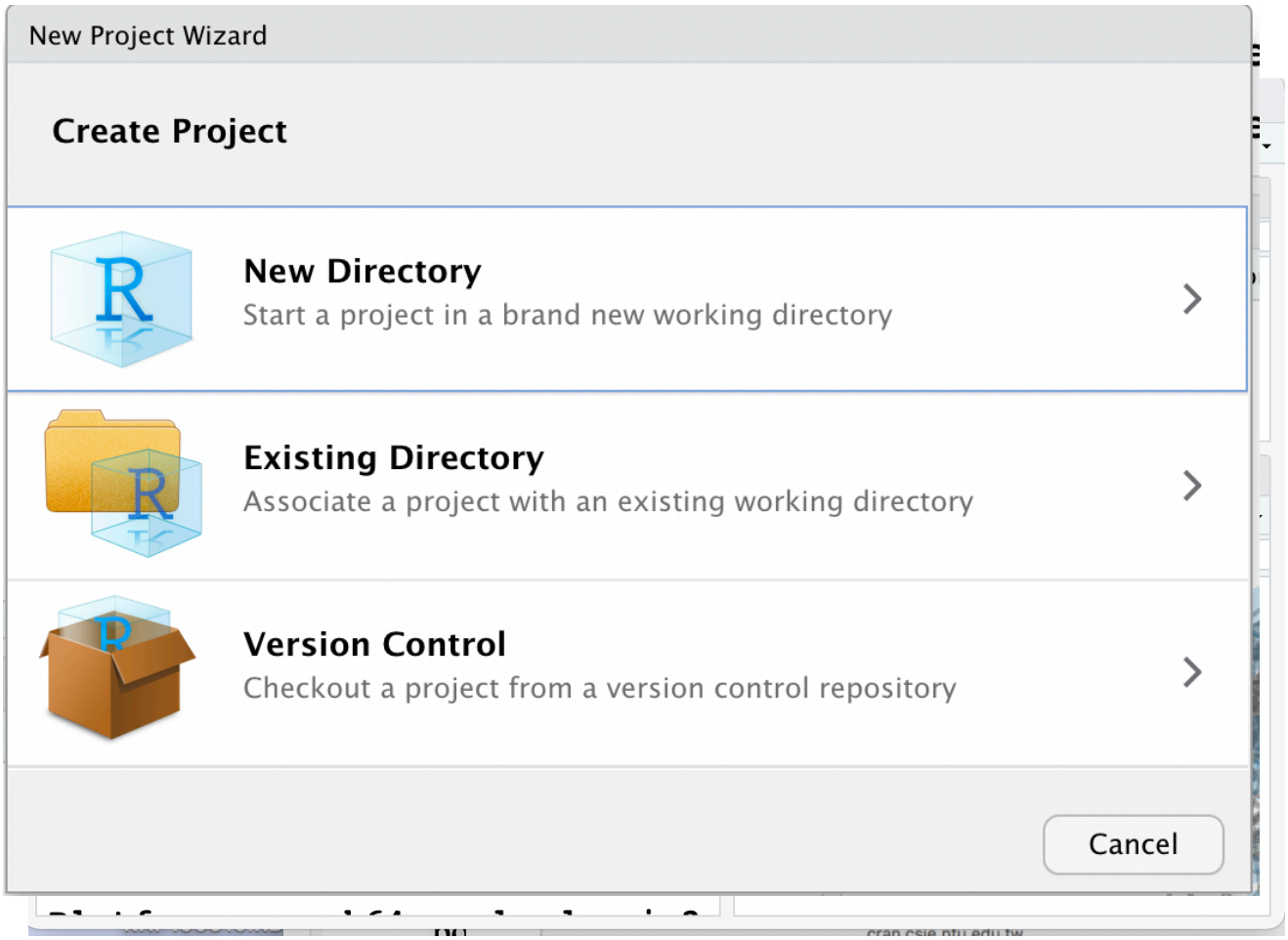
1.1.打開你的RStudio

Rstudio操作說明：請參考這一篇「(Day2) RStudio安裝與介紹」，主要是「RStudio環境設定(Options)」這個部分。
https://ithelp.ithome.com.tw/m/articles/10190868

我們要將Rstudio的版面設定為上圖的狀態。

2.設定資料夾

2.1 上圖右上角在 R FXG的這個地方進行設定。你的螢幕現在的狀況應該是 R Project (None)。



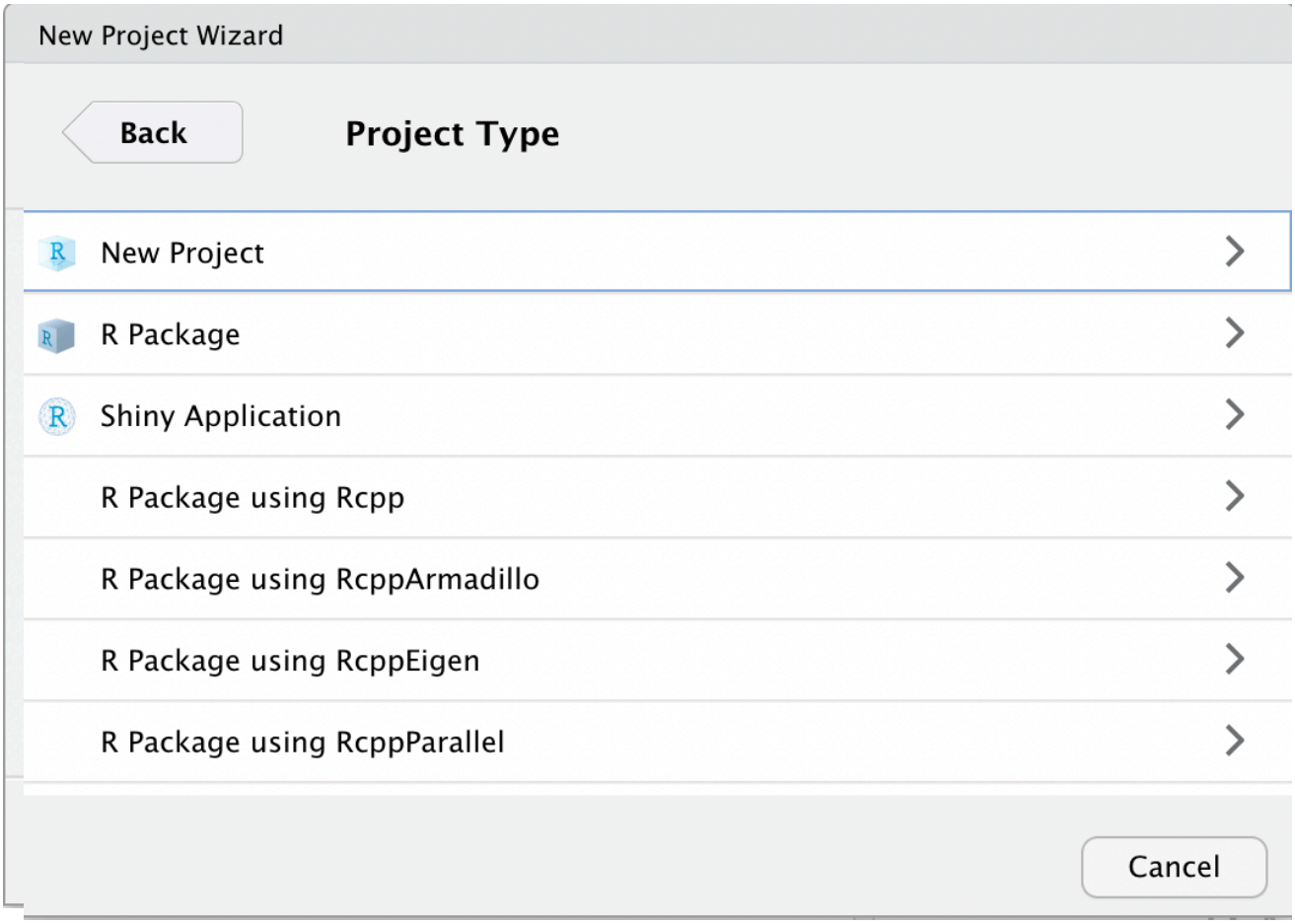

2.2 按下 R Project (None) 會出現下圖，請選R New Directory。

2.3 之後請點選 R New Project

2.4 設定資料夾。

2.4.1 先設定工作路徑。在“Create project as subdirectory of:”，這裏先設定為直接在「文件」（Documents）這裏，子檔案夾會直接設定在「文件」下面。

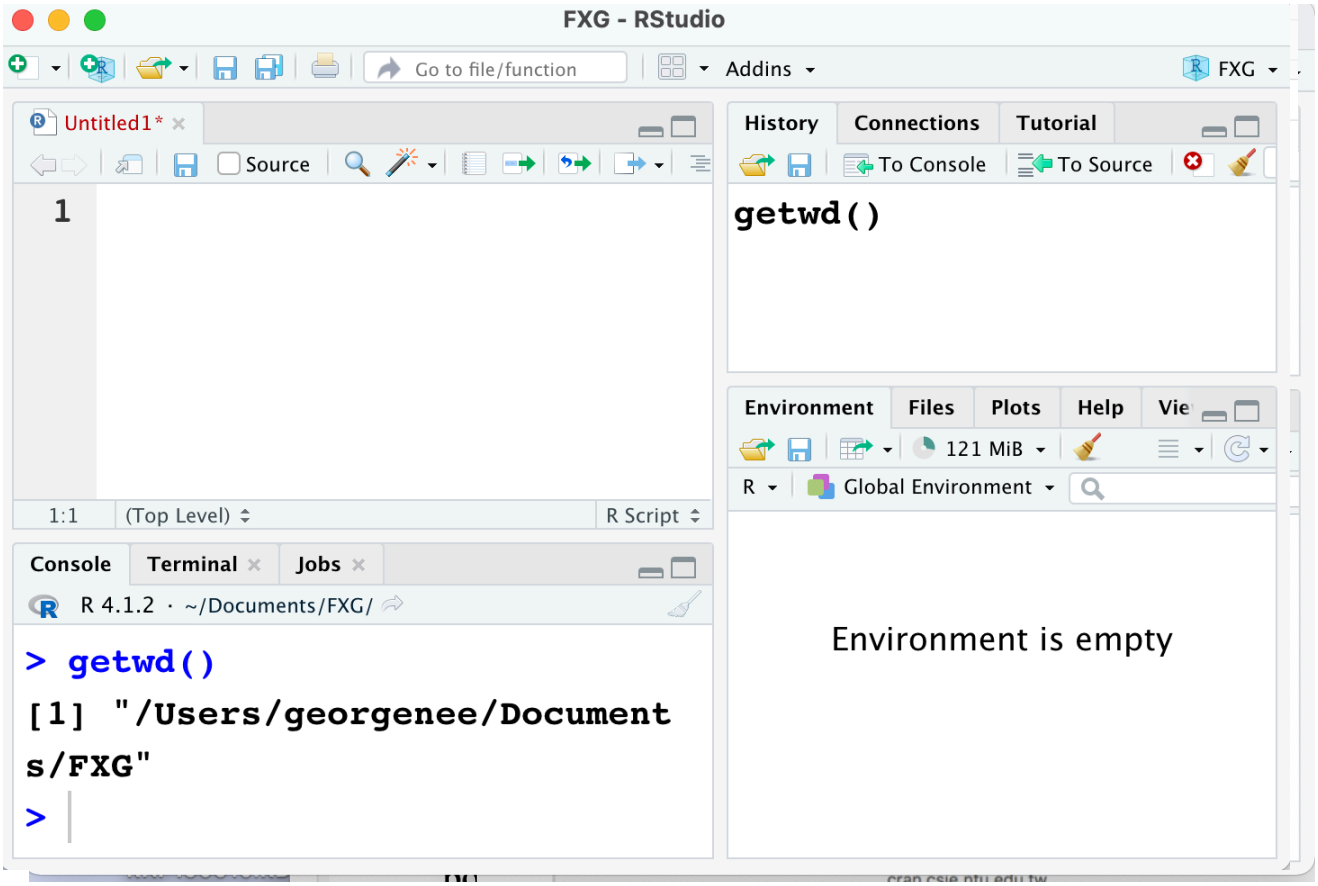
2.4.2 之後，我們在Directory name 這裡設定子資料夾名稱。請各位設定為FXG。（畫面為Fuxinggang, 請改輸入為FXG）。填入後按下右下方Create Project。




2.4.3 之後，各位參與者會發現，工作畫面就出來了。注意，右上角已經從R Project (None)轉為FXG了。（請各位參與者下載最新的R版本。）

2.4.4 回到文件(Documents)，各位參與者會發現你已經有了一個新的資料夾 FXG。

2.4.5 回到RStudio，在左下角的Console(中文：控制台)輸入“getwd()”意即取得當前的工作路徑(getting yr working directory），就得以知道我們現在在哪個路徑下工作。georgenee是講者自己的路徑，在各位參與者自己的電腦上，出現的會是各位自己當初設定電腦「使用者」的名稱。




3. 下載並運行應用軟體。

因應不同的需求，在R的世界中存在許多開發者撰寫各種「包」(package)以支應不同的需求。像是ggplot2, igraph等包用在繪圖，sjPlot則是應用在社會科學統計等等不一而足。

我們先試著下載幾個包，並且運作一下，看看是否完成下載。

指令install.packages()為下載功能，因為在使用R的過程中，必須時常下載（或是更新）這些「包」，因此，在應用時最好在網路長通的情況下使用，這一點必須時時注意各單位相關的資安規定，以免誤觸法網。當然，如果已經事先已經下載好自己所需要的「包」，要使用的時候拿出來library()一下就可以了，不連網也是可以的。

在Console輸入：

install.packages("quanteda") #表示下載quanteda這個包。

library(quanteda) #表示開始執行這個包。

install.packages("ggplot2")
library(ggplot2)

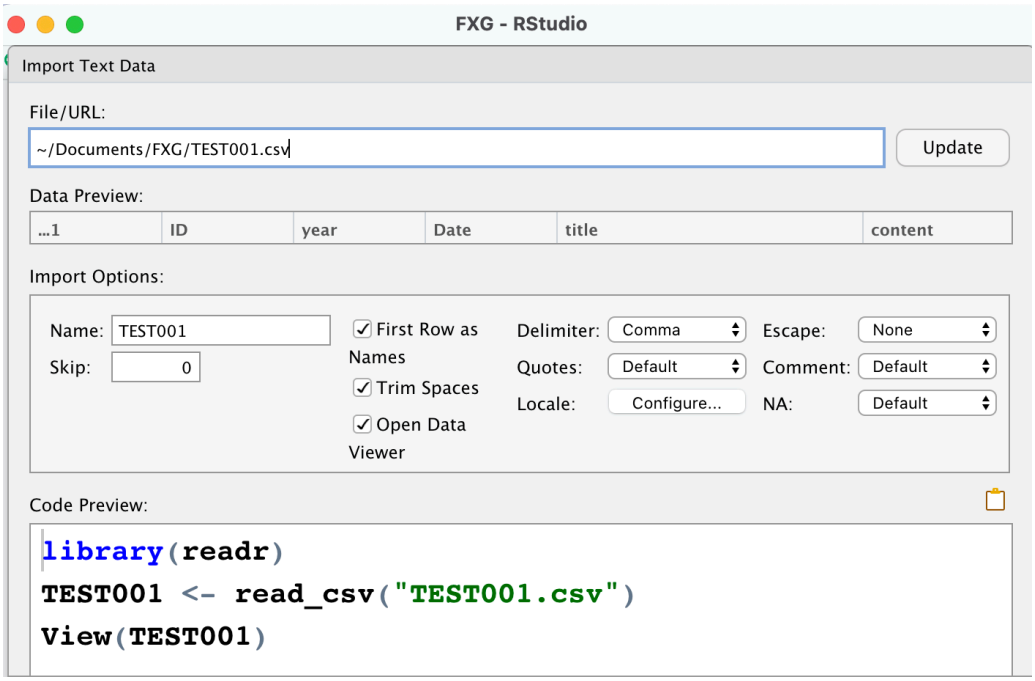
請各位參與者下載並執行以上兩個包。

必須注意的是，在R語言中，引號必須是"ggplot2", 而不能是“”。但是我們在Microsoft Word以及MacOS Pages都很容易在輸入時變成後者。因此，我們通常不在以上兩者最普遍使用的文字編輯作業軟體中進行程式語言編寫，這也是為什麼需要下載程式語言編寫相關程式的原因。
在本次課程中，我們採用Sublime Text3進行程式語言的編寫。
相關的套件非常多，像是更廣泛應用的VSCode（https://code.visualstudio.com/)。但未課程進行順暢之故，我們在本次課程中統一以Sublime Text3進行編輯作業。

4. 開啟檔案。
我會先給各位參與者兩個檔案進行測試，測試的目的是確定作業系統能夠讀進檔案。這兩個檔案叫做Test001.csv 與Test002.csv。前者是繁體中文，後者是簡體中文。我們看看是否能夠讀進作業系統。

為什麼需要讀進作業系統？
在R裏面，如同我們在其他作業系統如Microsoft Excel, MacOS Numbers 一樣，我們必須系統性地處理資料，把資料的各項特徵（即變項）以dataframe的方式呈現。我們把資料讀進R裡面才能夠以R語言進行資料編輯與修改。
.csv 的副檔名，與.xlsx, .doc, .pdf一樣，就是一個副檔名。但是儲存以.csv為副檔名的檔案往往是以dataframe的格式儲存。我們在將檔案讀入 R 的過程中需要講檔案儲存為.csv的格式。當然，如果各位參與者原先工作的檔案是以EXCEL儲存，我們也是可以讀進去 R 裡面的，這部分我們在函授的時候會教各位參與者怎麼處理。

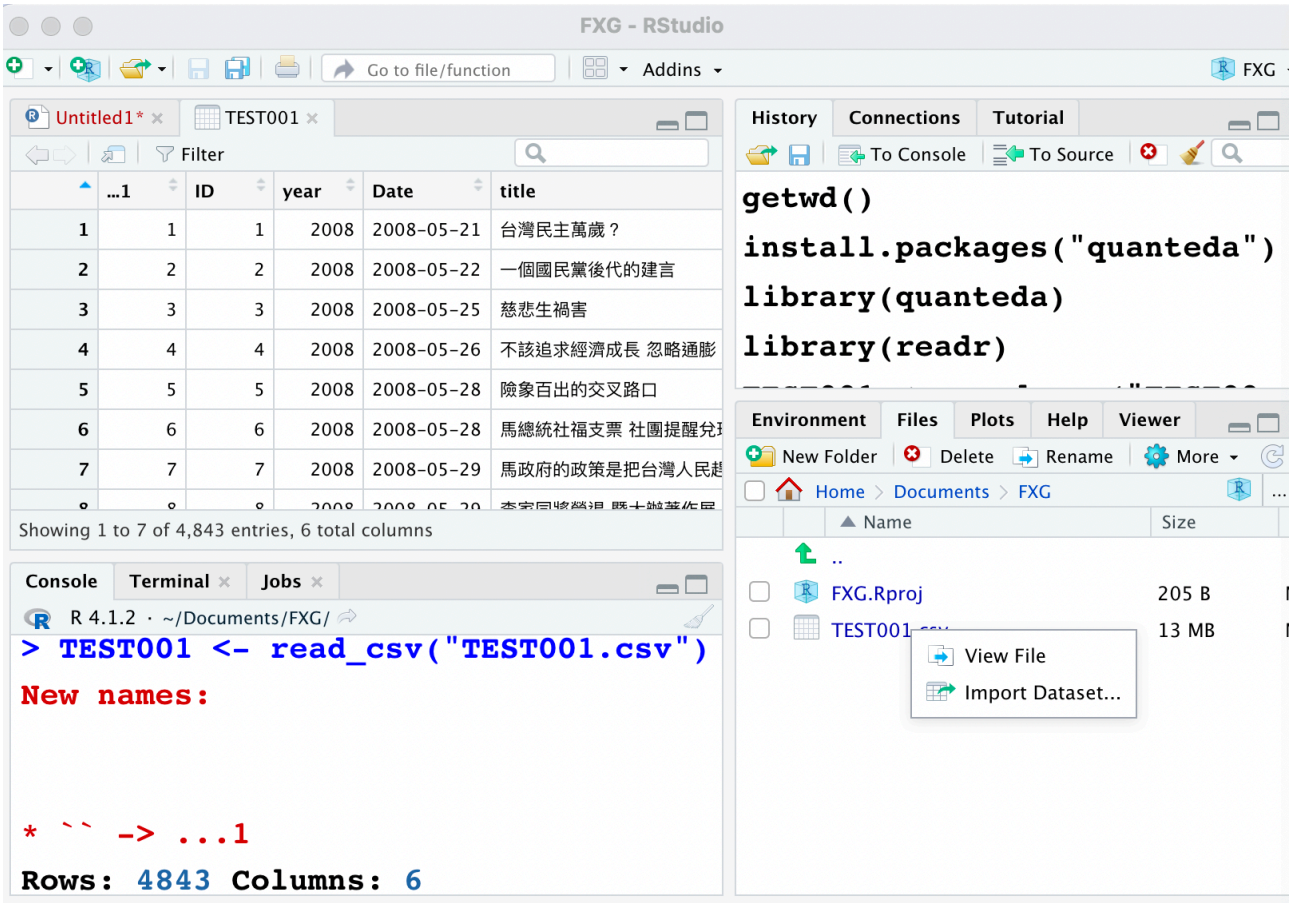

4.1 開啟檔案 Test001.csv。請各位參與者將LIne 通訊群組中Test001.csv檔下載後移到FXG這一個資料夾。之後您就會在RStudio工作頁面上的右下角，也就是Environment（工作環境）這一個區域看到，Test001.csv，TEST002.csv這兩個檔案已經進來了？




4.1.1用滑鼠左鍵點一下檔名的部分，會出現View File與 Import  Dataset兩個選項，請點擊Import Dataset。

4.1.2 之後會出現這一個頁面框，頁面框下會有一個Import選項（圖片中截圖沒截好未顯示，請各位參與者自行找一下。）

4.1.3 點選Import之後檔案就進來了，可查看RStudio左上角的作業框。




4.2 我們檢查一下，讀進去的兩個檔案，是否都分別以中文繁體字以及中文簡體字顯示？

4.3 最後，我們進行一些簡單的設定。

4.3.1 編碼。 我們在進行」中文文字分析，以utf-8編碼。

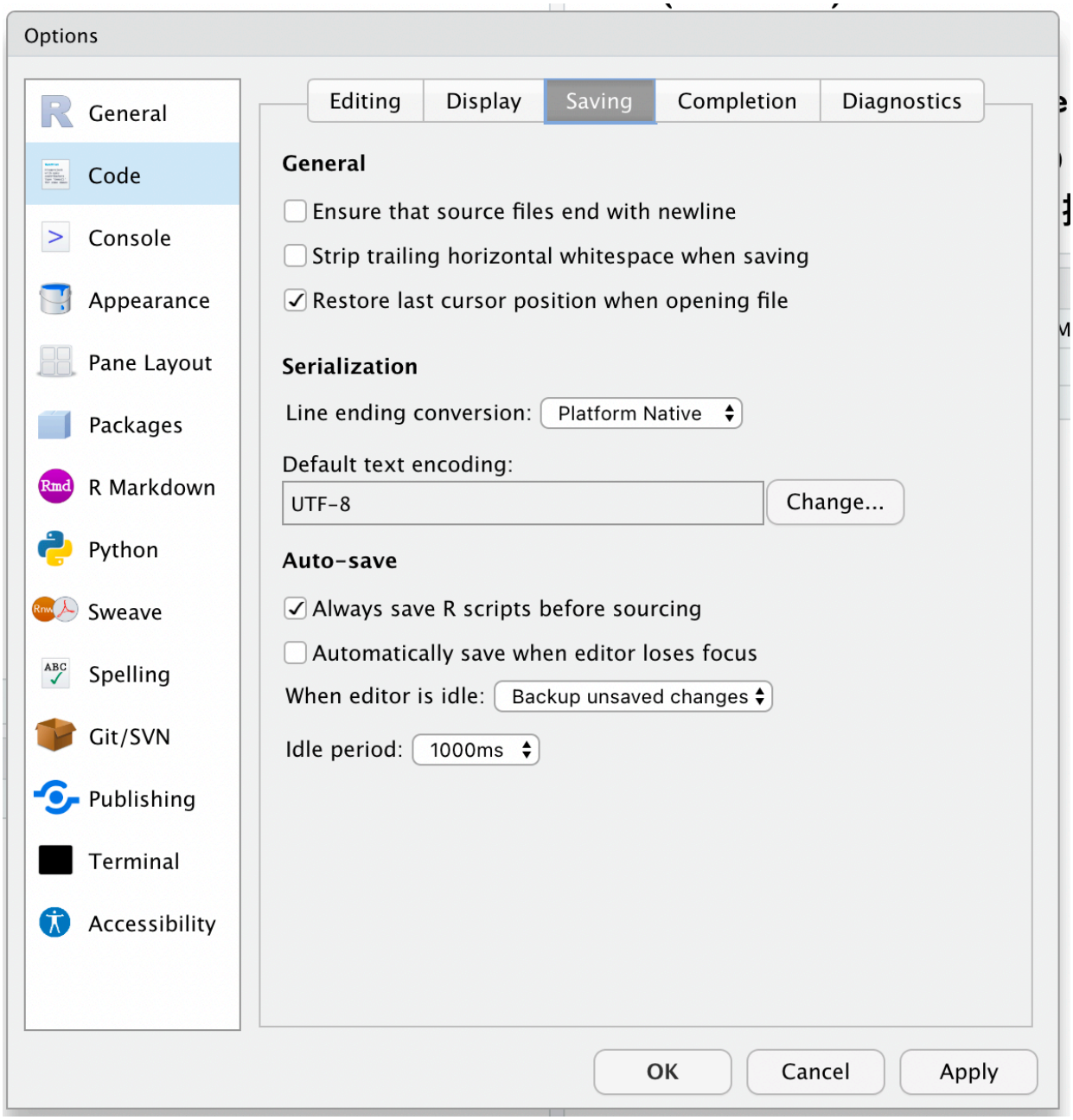
路徑：RStudio-Tools-Global Option- Code-Saving

各位參與者可以看到Default text encoding的欄目中可以進

行編碼(encoding)的格式，我們選擇UTF-8。

點選Apply之後點選OK。

雖說UTF-8是RStudio預設的編碼，我們仍需要檢視是否確實。

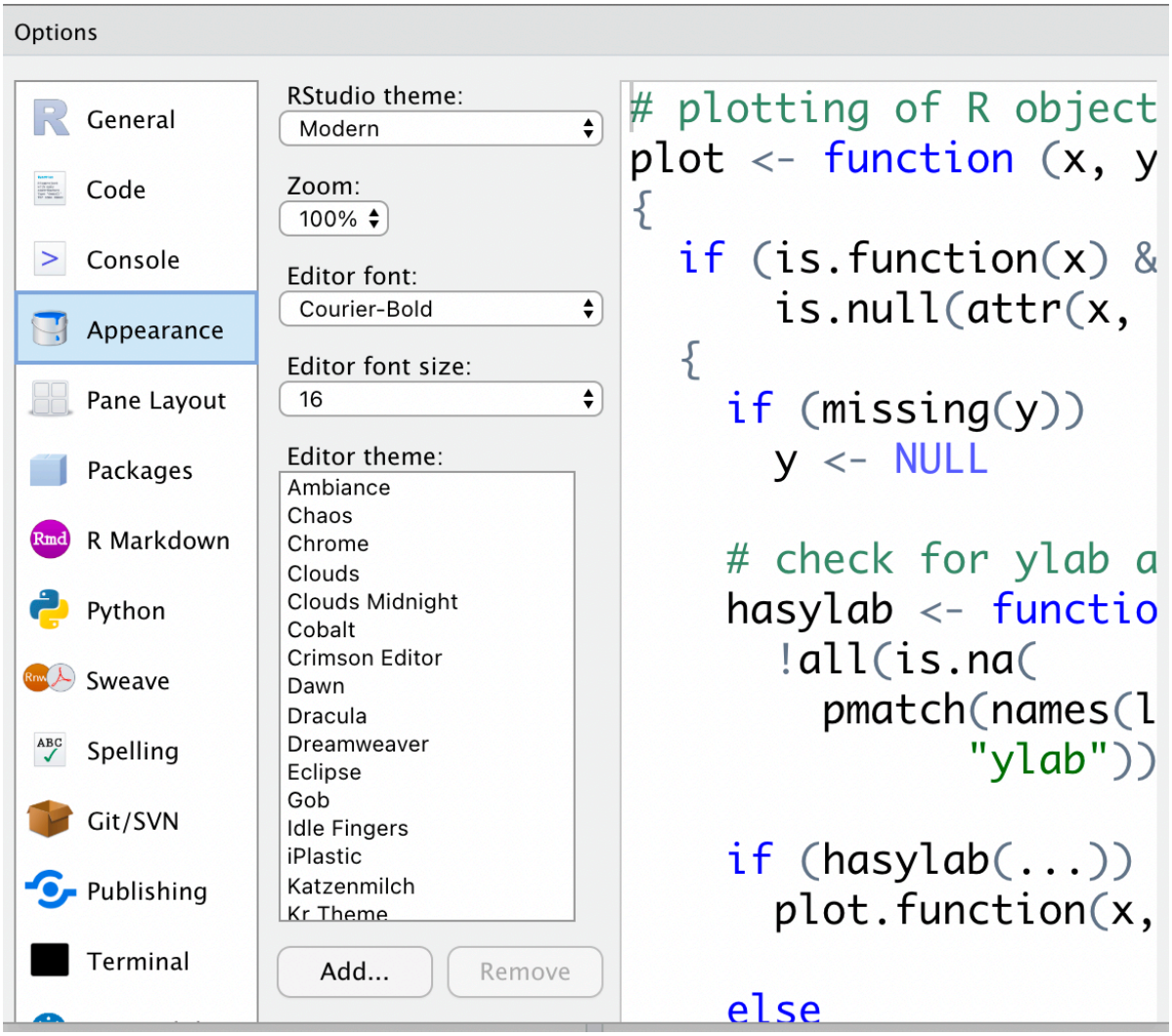


### 4.3.2 調整字體大小、字型與頁面風格

路徑：RStudio–Tools–Global Option– Code– Appearance

請自行調整，尤其是字體大小，調到16級字（以上）。

選擇結束後，請點選add, OK。




最後：請主辦單位、最有力、最強大的教育推手孫懋嘉博士準備多條延長線供各位參與者插電。我們的設定到這邊暫時告一個段落！

文本探勘論文寫作參考用論文：
情緒分析：https://www.nature.com/articles/s41599-023-01925-2。
主題建模：https://www.researchgate.net/publication/342423719_Mining_PIGS_A_structural_topic_model_analysis_of_Southern_Europe_based_on_the_German_newspaper_Die_Zeit_1946-2009。