ComPyRison

R & Python in PTT & Reddit 交叉分析

2019-01-12

簡介

學了一學期的 R，在過程中遇到的挫折除了靠萬能的老師和助教解惑外，我們更常在網路上請教各路大神，因此好奇怎麼樣的問題或發文能在論壇上得到較多的回覆與共鳴。 我們透過台灣與美國的兩大論壇 PTT 與 Reddit，試著從發文內容中比較 R 與 Python 使用者的使用情境或關注點是否有差異，並比較不同語言文化者的程式學習或分享差異。

方法

本專案使用的套件有: dplyr, stringr, jiebaR, tidytext, ggplot2, PTTminerR 以及 rreddit。 爬取兩論壇的資料後，由於兩套件所能爬出的資料眾多，故先將多餘的欄位剔除以加快讀取和運行速度，也使文件易於閱讀。首先使用 dplyr 整理表格，僅留下發文者、po 文標題、po 文內容等資訊。並且搭配使用 regex 與 stringr ，產生四個新欄位: 問號數、字數，禮貌詞多寡，以及自謙詞多寡，以探討此四變量是否會影響回覆數，各版常見的 Error 也是我們欲分析的目標。我們使用 jiebaR 對文本進行斷詞，為分析本課程所學過的 R 套件在兩論壇上的討論度，我們加入了自訂的套件名稱字典，避免套件名稱在斷詞時被斷開。此外，停用詞字典也加入以去除停用詞。

資料取得

使用兩套件 PTTminerR 以及 rreddit 爬取兩論壇上 python 版以及 r 版的貼文以及留言資料。禮貌詞與自謙詞由組員觀察數篇貼文後，自行記錄頻繁出現的禮貌與自謙詞種類。reddit 版上並未出現自謙詞(例如:小弟、在下…等)，因此此變項不納入 reddit 之討論。以下是資料整理步驟：

1. 計算內文長度、問號數量、禮貌詞數量、自謙詞數量

我們直接在 dplyr 下 mutate 出這三個數值，其中內文長度使用 strsplit 簡單斷詞，問號和禮貌詞使用 stringr::str_match_all 匹配正則表達式後，直接計算數組長度。

2. 計算標題和內文的詞頻，並通過圖像展示

在這裏，主要使用 tidytext 進行斷詞，處理停用詞，計算詞頻等操作，並使用 ggplot2 和 wordcloud2 繪製頻率圖表和雲圖

3. 關於 error 的分析

使用 stringr::str_match 匹配 error 的形式，並進行統計和繪圖

4. 計算各變量與回覆數的關係

我們在這裏使用 pearson correlation test 檢驗內文長度，禮貌詞長度和問號數與回覆數的相關性。

結果

1. 關於 error：

從圖表中可以看出：在 python 中，出現較多的是 import error 和 type error；而在 R 中，error 則相對不集中。

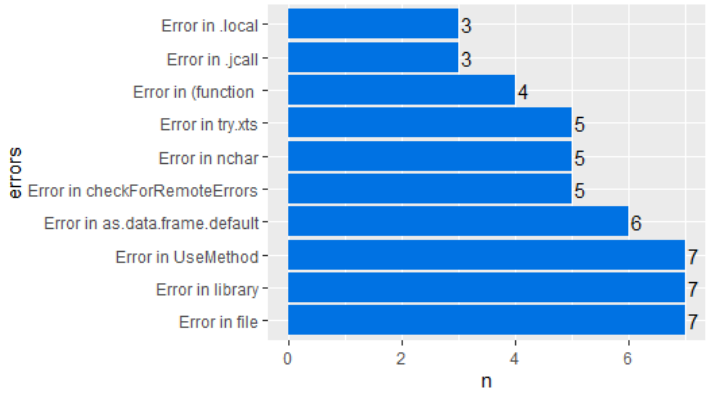

圖 1 PTT 中 R 版的 error 頻數表

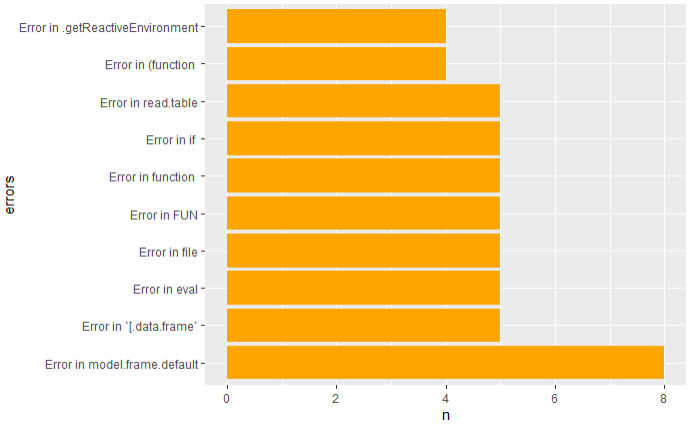

圖 2 reddit 中 R 版的 error 頻數表

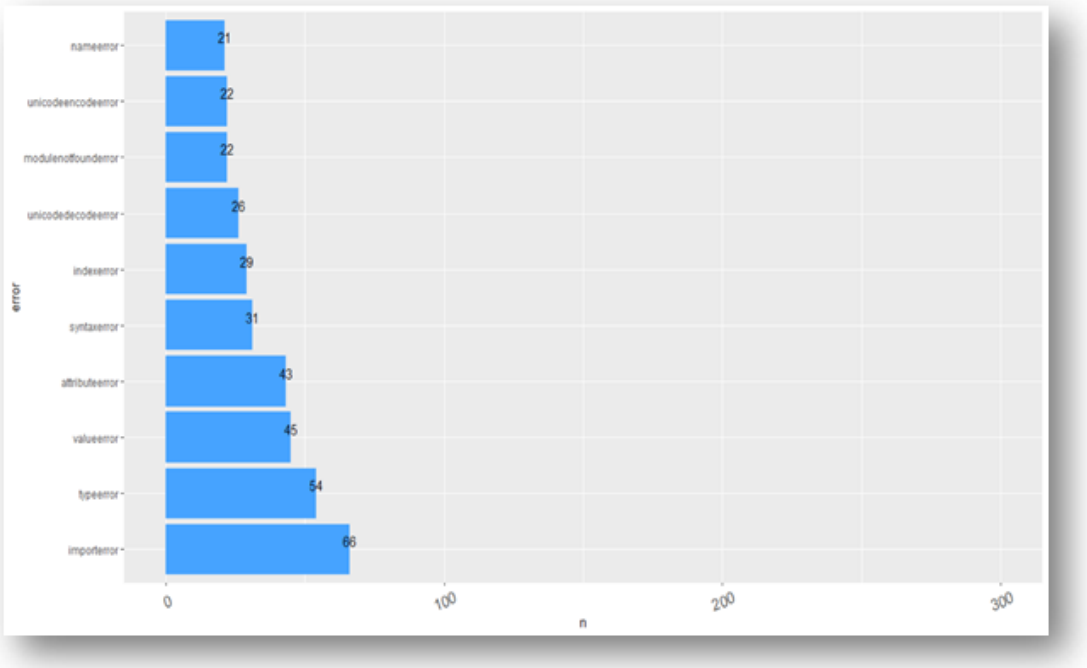

圖 3 PTT 中 Python 版的 error 頻數表

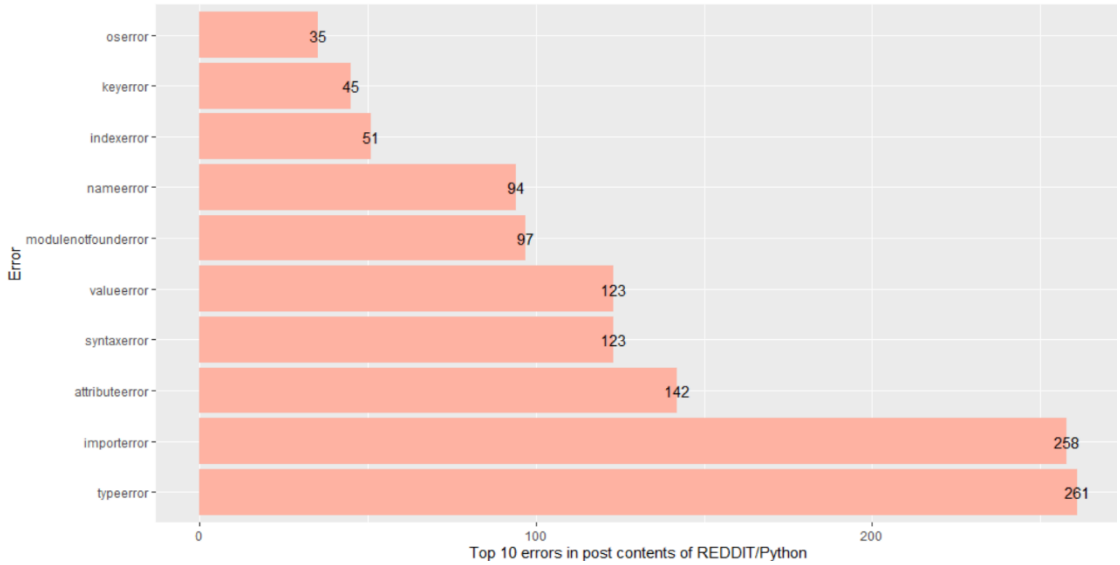

圖 4 reddit 中 Python 版的 error 頻數表

2. 關於學過的 R 套件的熱門程度

從運行結果可以看出，dplyr 和 ggplot2 是最熱門的兩個套件，而 reddit 上的數值又遠小於 PTT 上的。

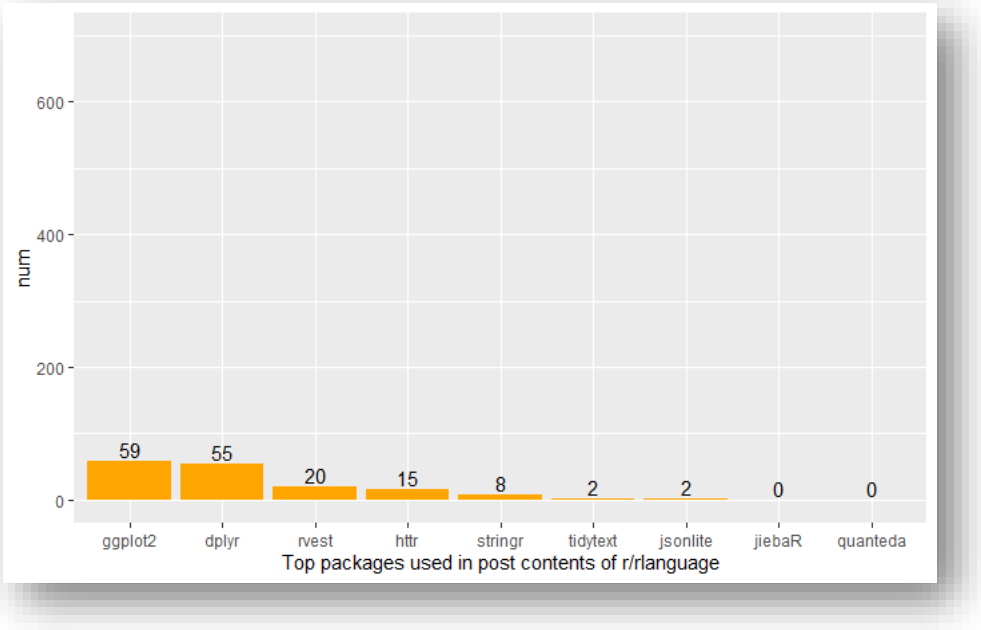

圖 5 reddit 中 R 版的套件頻數表

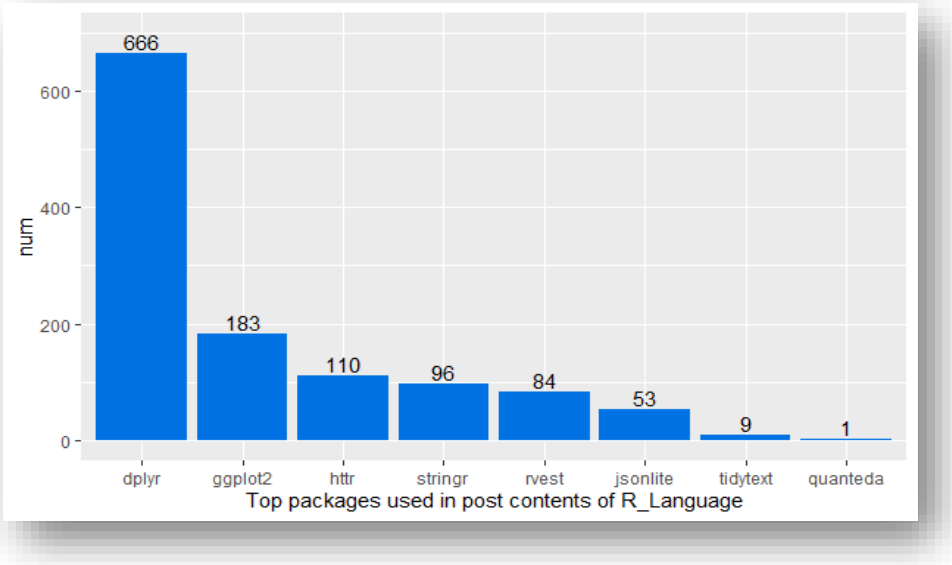

圖 6 PTT 中 R 版的套件頻數表

3. 關於標題詞的雲圖

從運行結果可以看出，兩個網站的 python 版最高頻的詞都是 python，而 R 版的則是資料(data)。具體的雲圖如下所示：

圖 7 PTT 中 R 版的標題詞雲圖

圖 8 reddit 中 R 版的標題詞雲圖

圖 9 PTT 中 Python 版的標題詞雲圖

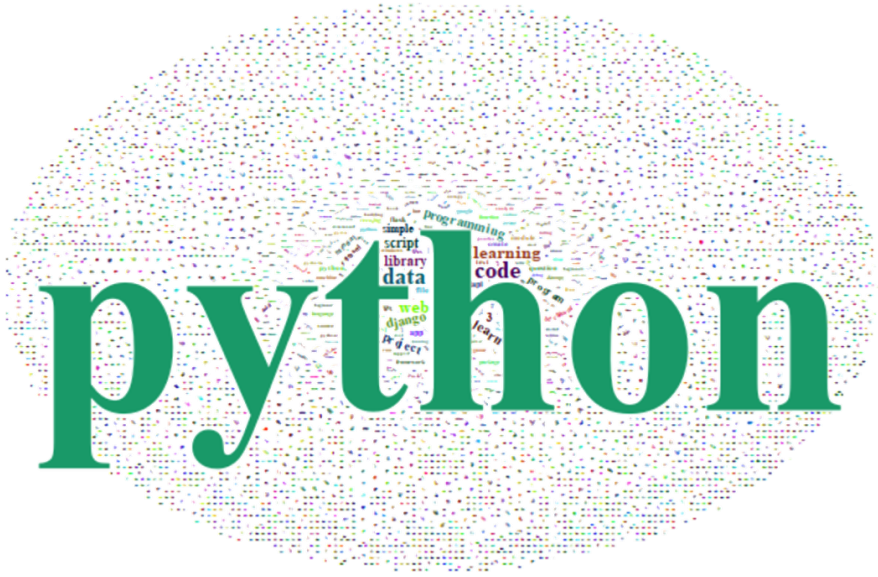

圖 10 reddit 中 Python 版的標題詞雲圖

4. 各變量與回覆數的相關性

通過 pearson correlation test，我們計算出了以下關係

論壇	看板	文長	問號數量	禮貌詞數量	自謙詞數量
Reddit	R	0.0473	0.0981	0.0529	（無自謙詞）
	Python	0.0224	0.0314	不相關	
PTT	R	不相關	0.0438	0.0503	0.0369
	Python	不相關	不相關	不相關	0.0319

表 1 各變量與回覆數的 pearson correlation test 表

可以看出，文長、問號數量、禮貌詞數量與對回覆數量的相關性不是無相關，就是相關性極小。

討論與貢獻

Python 版在兩論壇的資料數均多出很多，可能原因是開版時間早，使用人數多。

1. 最常出現的 Error

   從圖表中可以看出：在 python 中，出現較多的是 import error 和 type error；而在 R 中 error 則相對不集中。這一現象可能是由於 R 的 error 形式并不不統一而造成的，不同的 code 會出現不同的 error，同一個 code 在不同的套件下也有可能出現不同的 error。

2. 學過的 R 套件

   在兩論壇上均是 dplyr 和 ggplot2 佔前兩名，可以看出 dplyr 和 ggplot2 作爲功能比較基礎的套件，確實在資料處理與展示上是很常用的。本學期學過的套件在 reddit 上討論度不如 PTT 高，推測是台灣的 R 使用者較常使用這些套件，國外使用者也許風行的是其他套件。

3. Title 詞頻表

   可以發現都有 “資料” 或 “data”，表示 python 與 R 的使用者多使用程式處理資料。

4. 各變量與回覆數的相關係數

   是本專案最主要也是最重要的部分。多打字，多打問號，更加禮貌或自謙地提問，到底能不能有效提高回覆數呢？實驗結果給了這個問題一個接近否定的回答：可以說幾乎沒有關係，或者就算有關係，也是非常微小的關係。換句話

說，無論怎麼樣提問，可能對問題的解決都無法提供太大幫助。 造成這個局面的原因可能是多方面的：一方面，可能如何提問，確實對回覆數沒有太大的影響；另一方面，可能對回覆數能造成很大影響的要素不在這幾個變量之內。例如說，如果有人提了個很難的，大家都不會的問題，那麼就算他多麼有禮貌多麼誠懇地提問，回覆數也不會太多；相反地，如果有人提了個大家可能都會（至少是自認爲會）的問題，那麼討論的熱度也會高一些。如果事實如同後者所說的那樣，那麼也可以說明，這些版上的用戶，相比外在的形式（例如提問的方法）而言，更加注重內容（具體的問題）。

附錄

組員分工

王凱弘: PTT python 版資料爬取、整理與繪圖，海報設計

石晴方: PTT R_Language 版資料爬取、整理與繪圖，書面報告撰寫

張飛揚: Reddit r/rlanguage 版資料爬取、整理與繪圖，書面報告撰寫

張鈺琳: Reddit r/python 版資料爬取、整理與繪圖，海報設計