108-1
R語言與
資料科學導論
第一組

ComPyRison: R & Python PTT與Reddit交叉分析

一直error！怎麼問大神才會救我？

語言所 王凱弘 石晴方 張飛揚
工管系 張鈺琳
指導老師 謝舒凱 教授

專案動機與假說

學了一學期的R，在過程中遇到的挫折除了靠萬能的助教解惑外，我們更常在網路上問各路大神，因此好奇**怎麼樣的問題或發文**會在這些論壇上得到**較多的回覆與共鳴**。
我們也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**。此外，我們選擇台灣與美國的兩大論壇，比較**不同語言文化者**的程式學習或分享差異。

資料：R & Python版

	起始時間	觀察值*變量
Reddit	R: 2011-02-11	/ 5596*23
	P: 2008-03-19	/ 100158*23

📌 Author 📌 Self text 📌 Title 📌 Id

批踢踢		評論	貼文
	R: 2013-03-28	/ 17949*7	2884*10
	P: 2013-01-01	/ 34944*7	4690*10

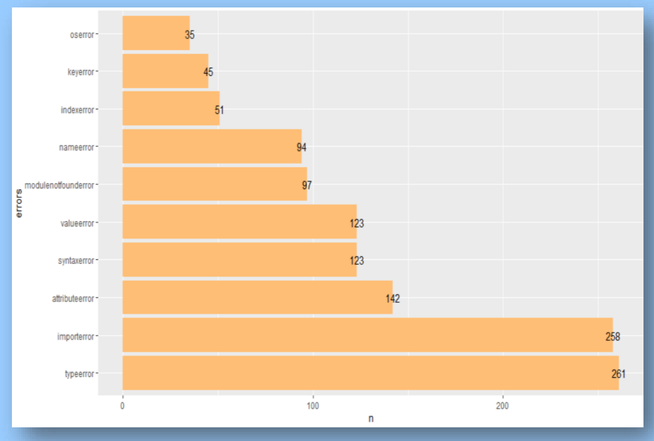
📌 Push id 📌 Push content 📌 Push time
📌 Post title 📌 Post content 📌 Post time

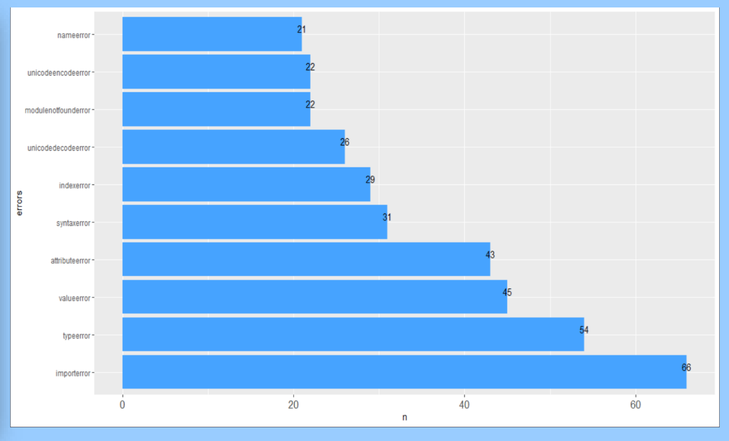
資料爬取與處理

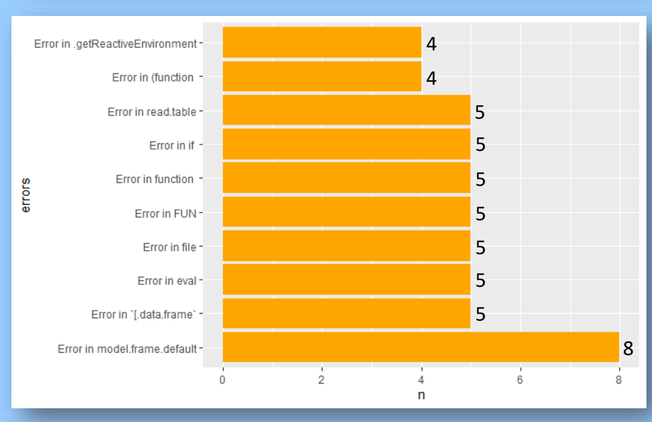
爬資料（套件：PTTmineR和rreddit）
→ 去除不需要的欄位
→ 添加常用套件的字典
→ 去除停用詞
→ 標題斷詞
→ 透過 string match檢驗標題
→ 找出高頻詞與常用套件
→ 產生新變量
→ 檢驗與回覆數的相關係數
→ 透過regex挑出錯誤訊息
→ 找出高頻錯誤

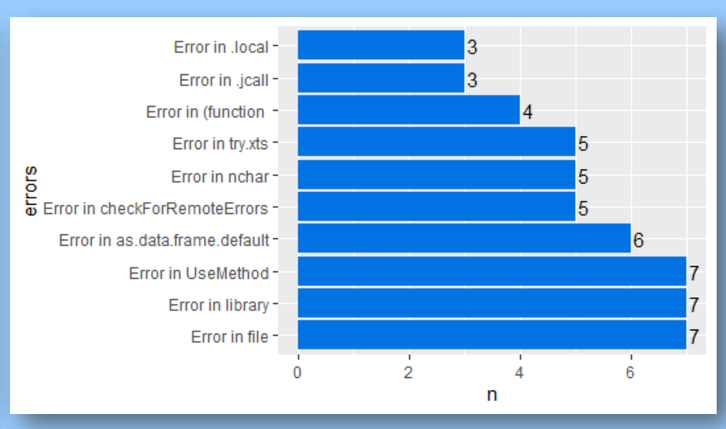
Error, Package & Word

各平台 R/Python版 前10大Error PTT

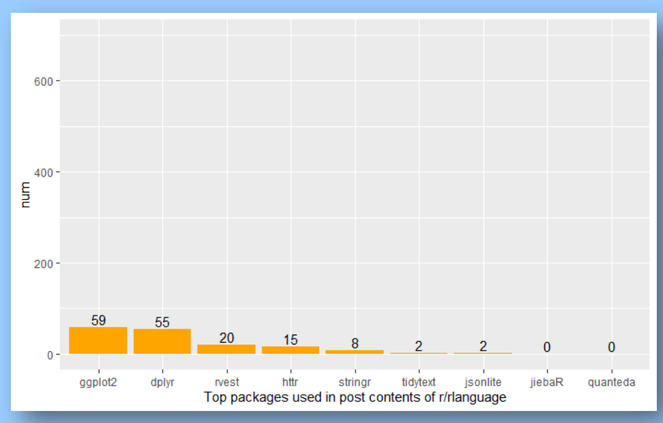


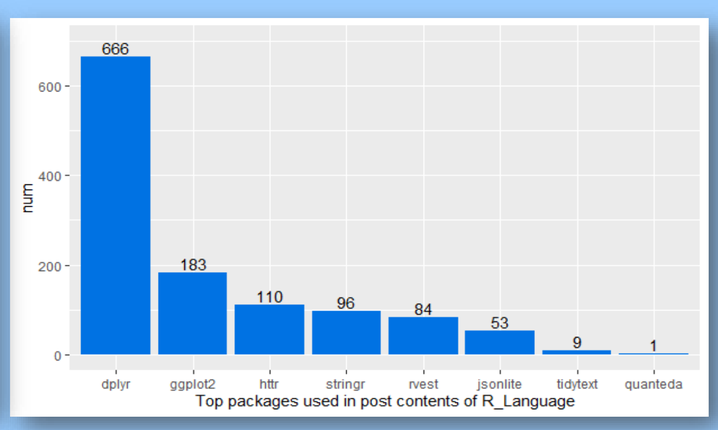






學過的R套件之版上排名





各版Title詞頻文字雲

各變量與回覆數的相關係數

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

文長、問號數量、禮貌詞數量與自謙詞數量對回覆數量的**不是無相關，就是相關性極小。**

（使用Pearson correlation test）