Text mining

一開始的構想為利用"東京著衣"fb 粉絲專頁，去進行 Text mining 分析技術找出最新流行的衣服穿搭，但在詞雲分析的部分，所得出的資訊並不如我預期的多，有用的訊息非常少，因此，我轉向利用"天蠍座(Scorpio) - 10/24~11/22"fb 粉絲專頁，進行 Text mining 分析，希望透過此技術能更快搜尋到天蠍座的特性。

第一部分

首先先去搜尋此粉絲專頁的前 100 篇文章，以及 70 篇文章中粉絲留言的訊息，分別存成 postScorpio.csv 以及 commentScorpio.csv

第二部分

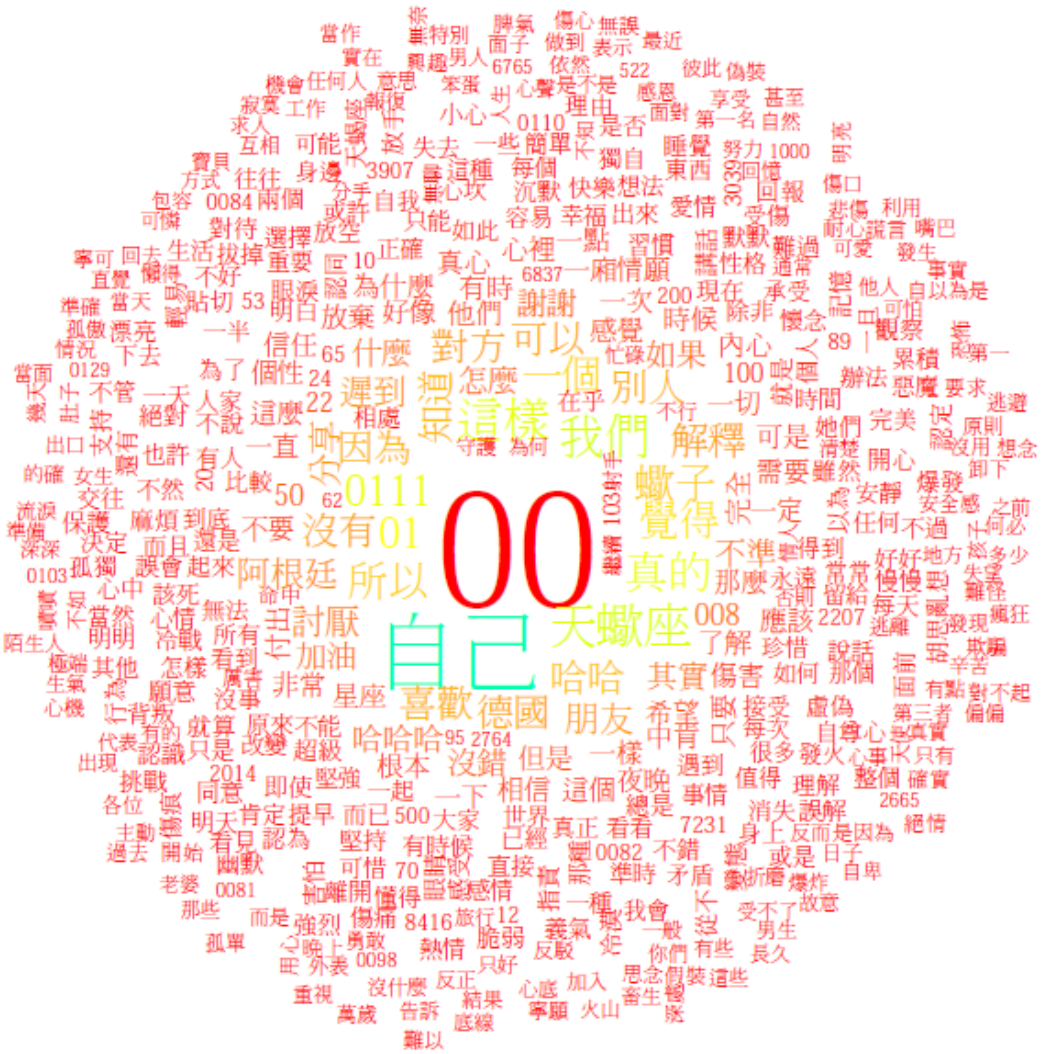
接著製作詞雲，針對文章我選擇文字最小頻率大於 5 才會輸出在 pdf 中，針對粉絲留言的訊息也是用文字最小頻率大於 5 才會輸出在 pdf 中。

文章詞雲

從文章所建構的詞雲發現天蠍座次數是出現最多的，但其於出現次數較多的字例如:25、自己、分析、他們、全文、所以……，皆與想挖掘出的性格特徵沒有相關，在第三部分會將這些不重要的詞刪除，再建構一次詞雲，這樣就可以更清

楚的看到，哪一些字詞出現的次數最多。

粉絲留言詞雲



從文章所建構的詞雲發現 00 這個符號是出現最多的，但其於出現次數較多的字例如:自己、天蠍座、我們、蠍子、德國、哈哈、朋友、阿根廷、喜歡……，也與想挖掘出的性格特徵沒有相關，其實我們可以發現頻率更低的詞，例如:幽默、包容、自以為是、獨自、信任……，已經透露出一些天蠍座的個性，在第三部分會將一些不重要的詞刪除，再建構一次詞雲，這樣可以更清楚的看到，大多數的粉絲留言包含了哪些天蠍座特值的訊息。

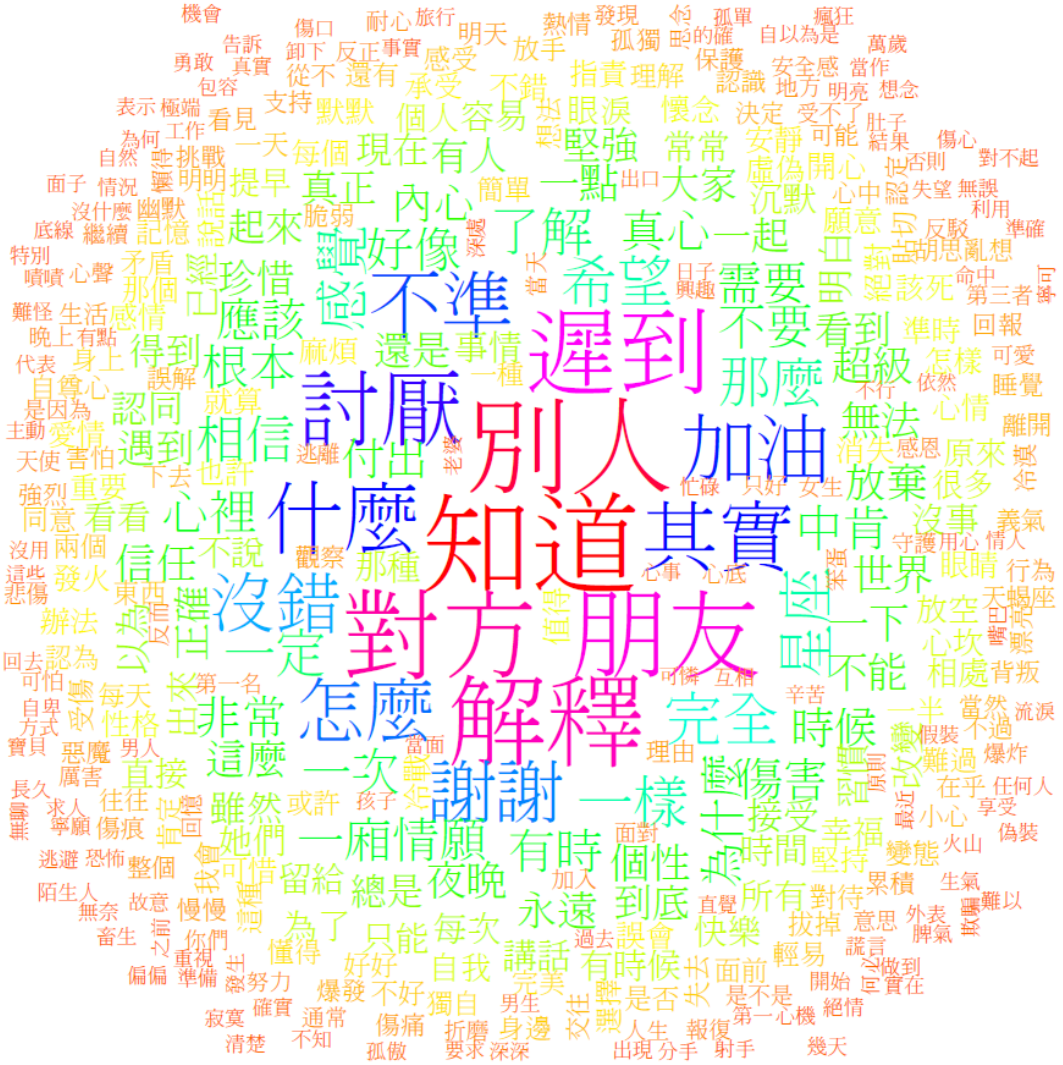
第三部分

利用粉絲留言資料去刪除字不重要的字("真的","哈哈哈","分享","因為", "比較", "這樣", "目前", "所以", "覺得","變成","這個","沒有","一個","天蠍座","蠍子","全文","可以","他們","25","分析","自己","喜歡","我們","德國","阿根廷")，去建構粉絲留言詞雲，輸出的 pdf 包括，還未設定文字出現頻率以及設定文字出現最小

頻率大於 5。

刪減字後粉絲留言詞雲(還未設定文字出現)

刪減字後粉絲留言詞雲(設定文字出現最小頻率大於 5)



觀察詞雲發現"遲到"這個詞出現的次數蠻多的，當然"不準"也是出現次數蠻高的訊息之一。

第四部分

利用 100 則發文內容，先進行刪減文字，但這次先不把"天蠍座"去除，接著抓出發文內容中文字頻率超過一次的，以及發文內容中文字頻率超過兩次的，接著找出與"天蠍座"關聯程度為 0.1 以上的詞。

發文內容中文字頻率超過一次

```
 [1] "一針見血" "一連串"   "一廂情願" "一輩子"   "一點點"   "二話不說" "千回百轉" "小心翼翼" "小孩子"
[10] "不可思議" "不在乎"   "不得不"   "不論是"   "不懷好意" "天蝎座"   "天翻地覆" "天蠍座"   "心上人"
[19] "心甘情願" "只不過"   "正前方"   "交朋友"   "任何人"   "全世界"   "全神貫注" "因人而異" "曲高和寡"
[28] "有時候"   "百分之百" "老朋友"   "冷嘲熱諷" "局外人"   "忍不住"   "沒良心"   "狂風暴雨" "事實上"
[37] "忽冷忽熱" "忽閃忽閃" "信以為真" "怎麼樣"   "是不是"   "是是非非" "為什麼"   "耍脾氣"   "胡思亂想"
[46] "陌生人"   "面面俱到" "捉摸不定" "格格不入" "桃花運"   "神秘感"   "骨子裡"   "排行榜"   "理所當然"
[55] "第一名"   "第一眼"   "第二名"   "第三名"   "第三者"   "細水長流" "莫名其妙" "善解人意" "無所事事"
[64] "無論如何" "虛情假意" "想像力"   "愛出風頭" "感同身受" "運用自如" "遍體鱗傷" "實際上"   "寧缺毋濫"
[73] "對不起"   "監聽器"   "與眾不同" "說了算"   "說三道四" "靜若處子" "瞧不起"   "舉手投足" "雙子座"
```

可以利用超過一次頻率的的字，約略勾勒出天蠍座的性格，"一針見血"、"小孩子"、"曲高和寡"、"與眾不同"、"忽冷忽熱"、"虛情假意"、"神秘感"……的性格。

發文內容中文字頻率超過兩次

```
 [1] "不得不"   "不懷好意" "天蝎座"   "天蠍座"   "有時候"   "怎麼樣"
"為什麼"   "胡思亂想" "捉摸不定"
[10] "神秘感"   "骨子裡"   "第一名"   "第三者"   "想像力"
```

在 100 文章中出現頻率超過兩次的非常少，較難看出確切的性格特性，但還是有幾個詞值得注意，"不還好意"、"捉摸不定"、"神秘感"、"想像力"，這幾個詞是有關性格的描寫。

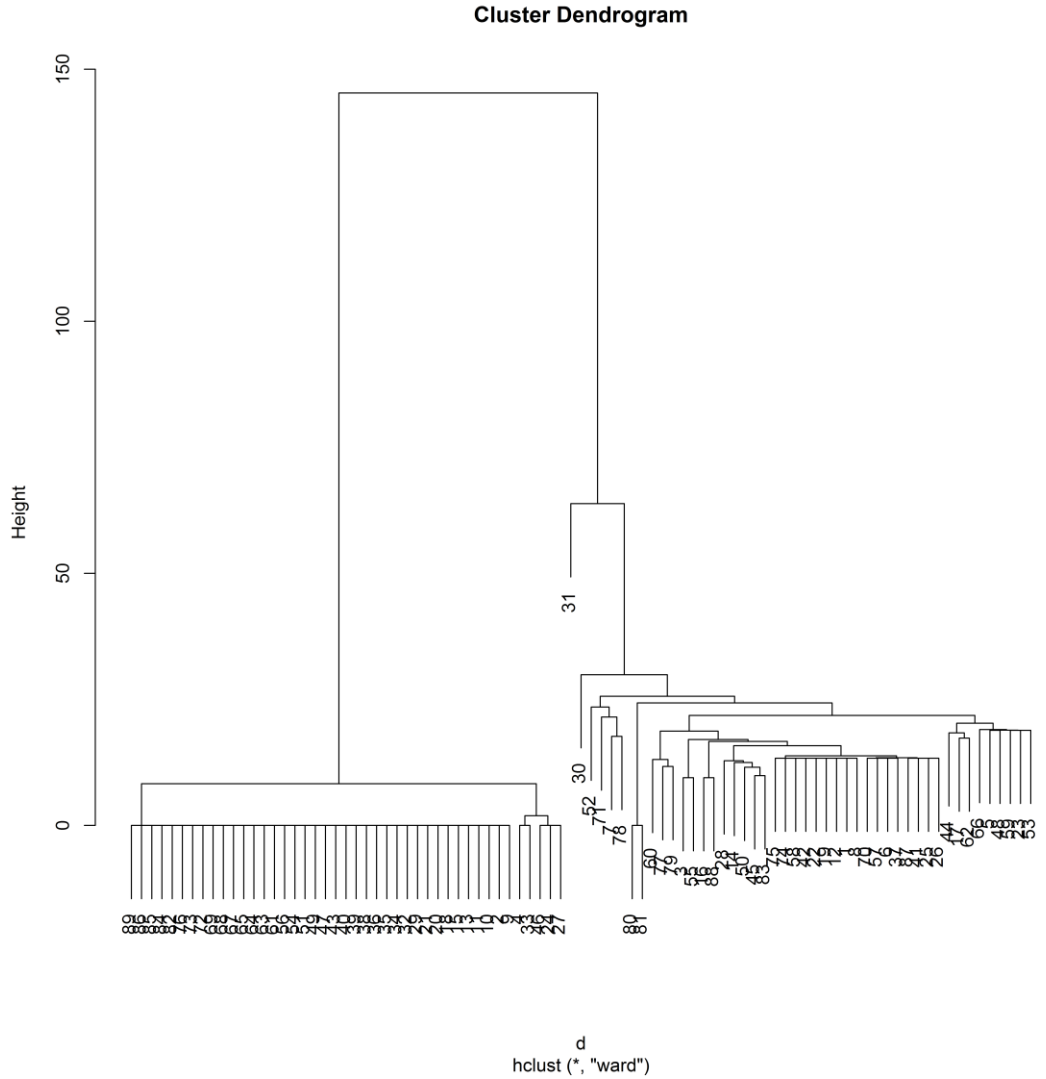
與"天蠍座"關聯程度為 0.1 以上的詞

一針見血	0.73	無論如何	0.73	全神貫注	0.12	神秘感	0.46
只不過	0.73	想像力	0.73	曲高和寡	0.12	瞧不起	0.29
任何人	0.73	愛出風頭	0.73	局外人	0.12	捉摸不定	0.23
因人而異	0.73	運用自如	0.73	胡思亂想	0.12	骨子裡	0.21
冷嘲熱諷	0.73	監聽器	0.73	面面俱到	0.12	細水長流	0.21
忍不住	0.73	靜若處子	0.73	格格不入	0.12	莫名其妙	0.21
事實上	0.73	舉手投足	0.73			一廂情願	0.12
桃花運	0.73					一輩子	0.12
善解人意	0.73					二話不說	0.12

從關聯性可以更明確的看出哪一些詞是跟天蠍座高度相關的，特別是第一、二欄相關程度大於 0.7 以上。

第五部分

沿用第四部分 100 則發文內容，進行集群分析，利用歐氏距離做群集分析並利用 ward 法。

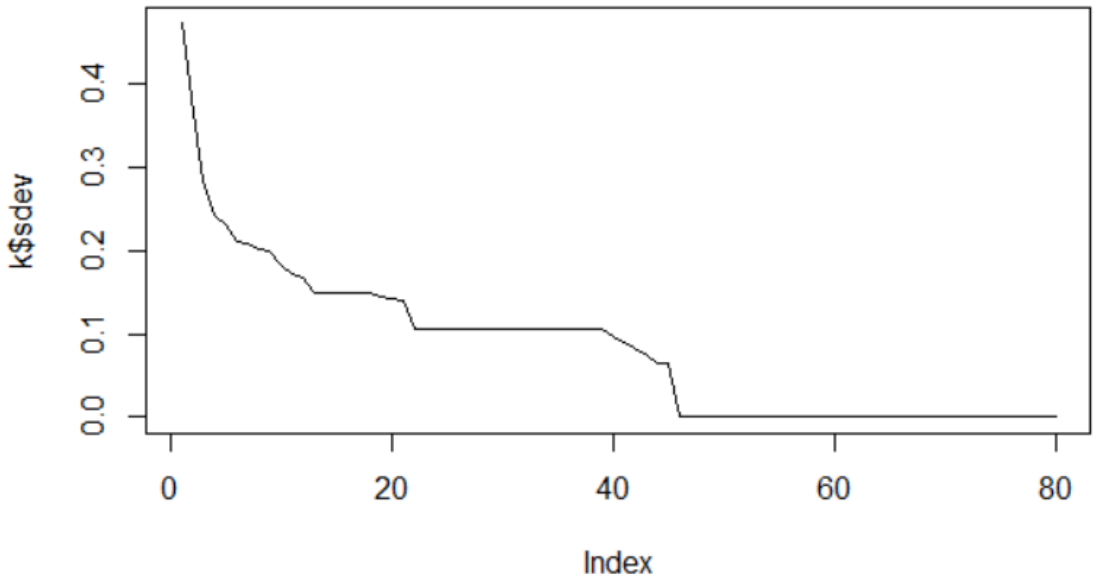


很清楚的可以看到大致上是分成兩大群，而特別的是 31 是自己單獨為一個小群，集群分析的結果大致上還蠻清楚可以去歸納。

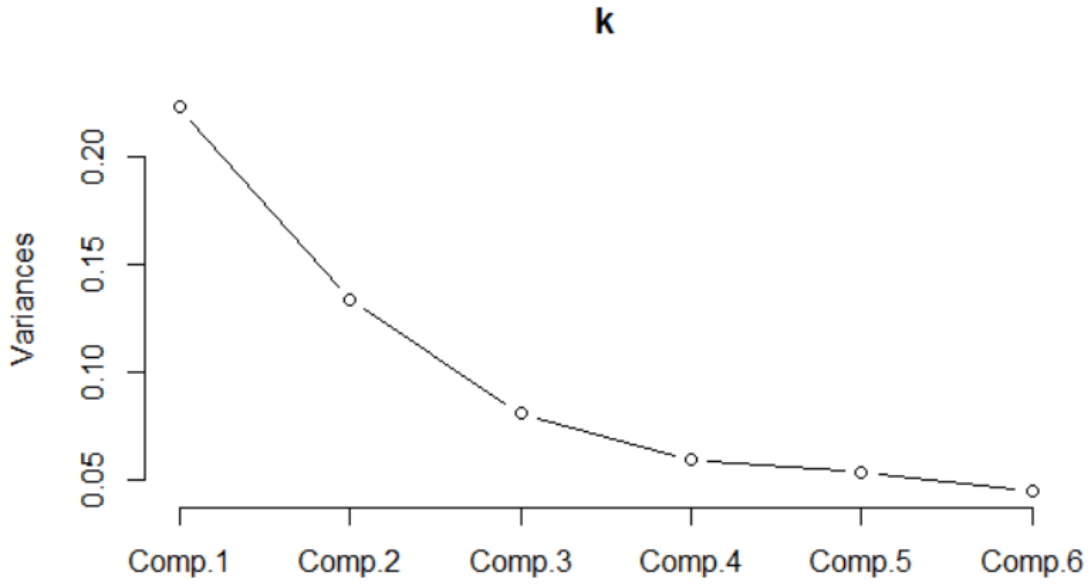
第六部分主成分分析

沿用第四部分 100 則發文內容，進行主成分分析。

這是還未設定主成分個數



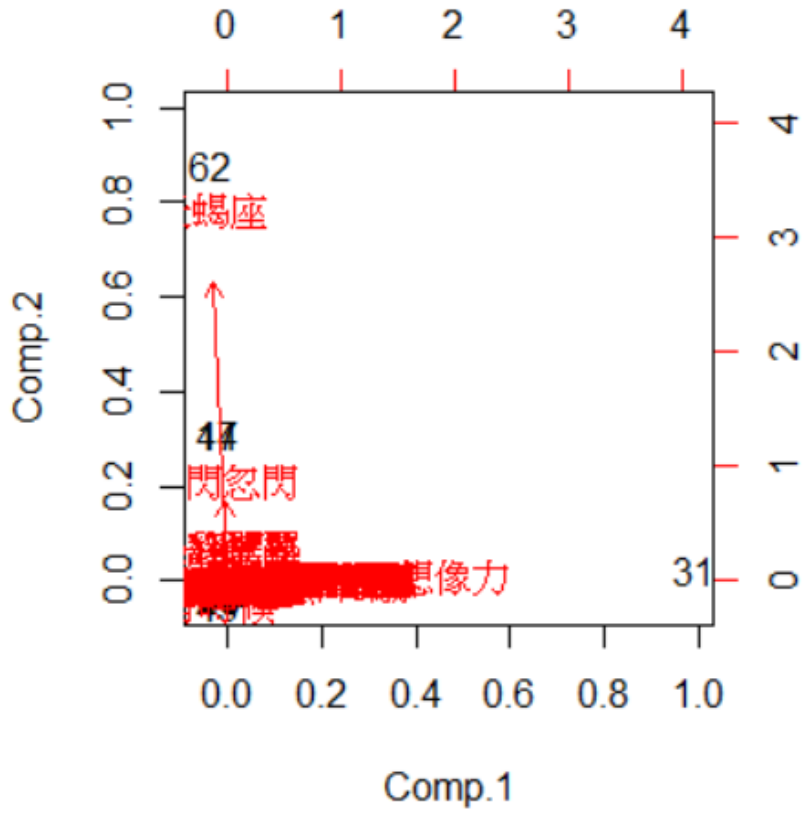
這是設定主成分個數為 6



雖然從圖型上有點難判斷，但如果以 6 個主成分來選擇，可能是選擇組成分 3 或是組成分 4，來看一下 Cumulative Proportion 以及 Proportion of Variance，發現主成分 3 與 4 的累積變異跟各別變異相差不多，因此選擇組成分 4。

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Standard deviation	0.4725354	0.3659012	0.28458388	0.24396882	0.23146226	0.21137717
Proportion of Variance	0.1789073	0.1072722	0.06489033	0.04769008	0.04292594	0.03579939
Cumulative Proportion	0.1789073	0.2861795	0.35106984	0.39875992	0.44168585	0.47748524

接著利用主成分 1 與主成分 2 畫出雙標圖



發現想像力與忽閃忽閃式次數最多的。