Dataholic 資料癮

《紅樓夢》的R語言文字探勘探索

前陣子一位網友-黎辰在他的微信公眾號上分享用機器學習判定紅樓夢後40回是不是曹雪芹所寫的文章造成滿大的轟動，雖然我知道文字探勘相關技術在文學與社會科學已經越來越多相關的應用，但對於這樣有趣的方式分析歷史懸案仍然讓我眼睛為之一亮。所以就決定用同樣的題材來試著自己玩玩看。

過去嘗試過的文字探勘都是以英語為主，這次算是我第一次實作中文的文字探勘內容，所以除了用呼叫函式庫之外也試著理解背後的演算機制。這次用的中文分詞是R語言的"結巴分詞"（jiebaR），基本上就是Python的jieba中文分詞移植過來的，結巴分詞的演算法支持最大概率法（Maximum Probability）、隱式馬可夫模型（Hidden Markov Model）、索引模型（QuerySegment）和混合模型（MixSegment）四種分詞模式，另外也有詞性標註、關鍵詞提取等等的功能，本篇目前僅用到分詞的功能。

分詞是什麼

分詞就是將我們說的語言切割成一個一個有意義的最小單位，在電腦做文本分析或自然語言處理的時候，分詞通常都需要將一句話甚至一篇文章切割成一個一個詞。例如：紅樓夢是中國四大小說之一。就會希望被切割成：紅樓夢/是/中國/四大/小說/之一。這樣一個個的詞，而要如何分的精準且符合原意，就被背後的演算機制影響著。至於結巴分詞背後的演算法理解，因為

牽涉到綠性代數、自動機理論和一點點的機率論，所以在本篇中就不再多做說明。有興趣的話我有找到一篇解釋結巴分詞背後隱式馬可夫模型演算法、字典樹等等的投影片供參考。理解完背後的演算法之後，我試著濃縮成精簡的說明：每個被分析的文本，都會被丟進事先定義好的分詞字典中，找出相對應的分詞，而當某些新詞沒有出現在字典中時，就會將連續的詞語序列的出現方式和結構，猜出一個機率最大的組合方式，例如“即”這個字後面可能會出現“將”、“時”、“刻”等等每種排列組合都會被計算一組機率，並找出最大機率的組合，就是最後的結果。

把手弄髒吧！

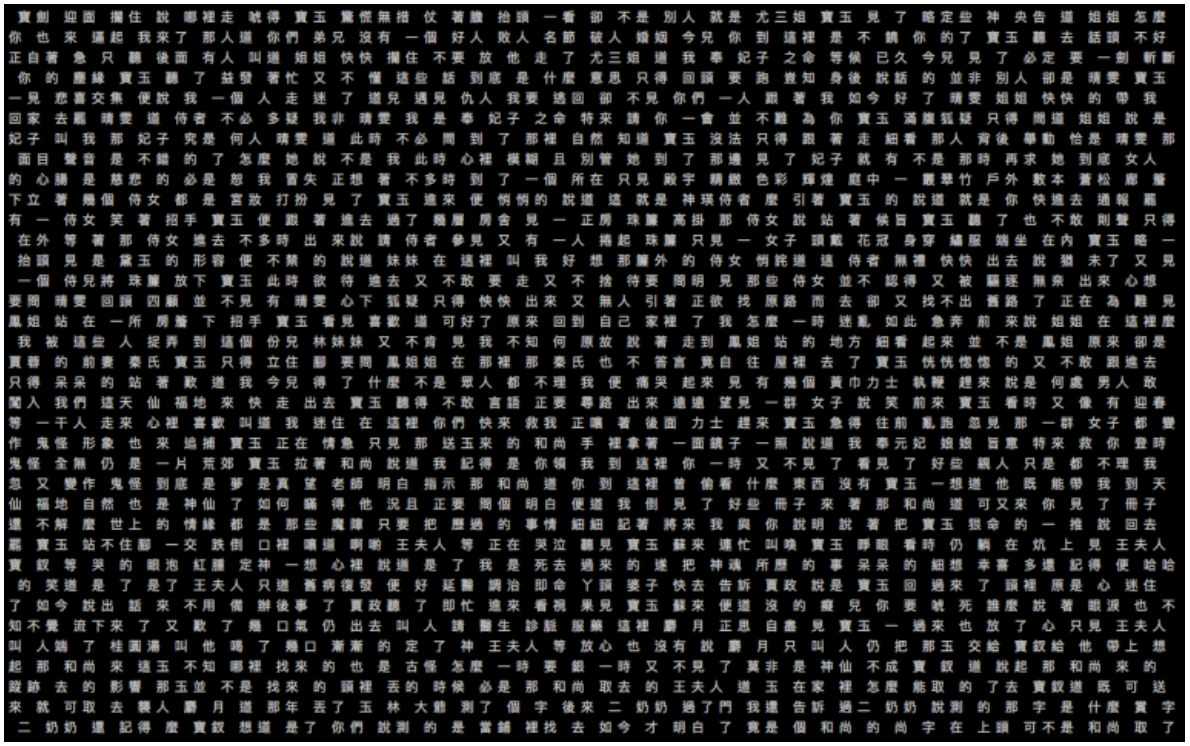
這次簡單的實作用到的package有jiebaR, tm, wordcloud

首先設定默認參數和分詞引擎

```
#Accept default parameter and bulit cutter engine
wk = worker()

wk <= './RedDream.txt'
```

分詞完後的結果會類似這樣的檔名"RedDream.segment.2016-11-27_17_49_03”出現在資料夾中:



接著我用tm函式庫將分詞後的結果建立成文本：

```
library(tm)
mycorpus <- Corpus(DirSource("RedDream.segment.2016-11-27_17_49_03"),
tdm <- TermDocumentMatrix(mycorpus, control = list(wordLengths = c(1,
```

然後手動分了前八十回和最後四十回，並且分別統計了個分詞出現的頻率，產出這樣的結果：

了	的	我	你	也	是	又	說	寶玉	她	去	來
12684	8196	4837	4136	3722	3377	3376	3302	2539	2298	2291	2215
他	著	道	有	不	便	都	就	在	笑道	人	這
2040	1936	1858	1776	1745	1729	1712	1705	1651	1592	1470	1389
聽	什麼	等	一個	那	好	叫	呢	只	和	要	得
1259	1095	1064	1012	1002	990	939	919	896	847	847	846
上	我們	見	忙	與	們	吃	倒	寶母	才	你們	一面
842	832	791	763	733	716	702	701	689	688	681	670
如今	再	出來	個	姑娘	兩個	鳳姐	把	因	知道	說道	王夫人
666	615	613	596	590	587	586	579	578	575	571	571
看	起來	拿	奶奶	老太太	這個	只見	自己	到	為	笑	太太
559	558	555	532	523	519	516	515	510	510	502	493

最後試著用文字雲的方式來呈現前八十回和後四十回的文字出現頻率：

```
#Setting Chinese font on word cloud
par(family=("Heiti TC Light"))

wordcloud(cloud1$word, cloud1$freq, min.freq = 100, random.order = F,
colors = rainbow(length(row.names(mtr1))))

wordcloud(cloud2$word, cloud2$freq, min.freq = 100, random.order = F,
colors = rainbow(length(row.names(mtr2))))
```

第一頁是前八十回的結果，第二頁是後四十回的結果。排除出現頻率最高的"你、我、他、的、了"等等這些主受詞或語助詞，可以看出來在橘色部份的連接詞或介系詞，這種牽涉到個人寫作習慣的詞在分佈上有些許的不同，因此其實不難看出兩者的差異。

因為此篇內容主要是透過這樣的練習來瞭解中文分詞系統演算機制運作的關係，所以還沒有牽涉到後續的機器學習部份，也許改天有機會可以再接下去試試看用不同的模型來分析！

後記

其實我想過找後四十回最有可能的編纂者高鄂和程偉元的其他文獻來驗證是不是他們寫的，但找了相關的文獻後發現高鄂工詩詞，遺世的作品都是詩集與詞集，和章回小說的用字遣詞有極大落差，程偉元則是科場失意，一生未仕，記載甚少。於是似乎頂多只能驗證出後面四十回是不是曹雪芹所著，對於後四十回眾說紛紜可能是出自誰手的真相，想必只能永遠被塵封在不得而知的歷史了吧！

如果對程式內容有興趣，我將詳細的code操作內容放在Github上供參考：
https://github.com/poweihuang/TheDreamofTheRedChamber

發表者：poweihuang

- 檢視 poweihuang 的所有文章

- **2016-12-06**

data science

向上 ↑