

# 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

牽涉到線性代數、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

```
1 #Accept default parameter and bulit cutter engine
2 wk = worker()
3
4 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：

寶釵迎風擺袖，說：「你走，你走。」寶釵見了，略定些神，央告道：「姐姐，怎麼你也來鬧起我來了？那人道：『你們弟兄沒有一個好人，敗人名節，破人婚姻。』今兒你到這裡，是不講你的了。寶釵聽去，話頭不好，正自着急，只聽後面有人叫道：『姐姐，快快攔住，不要放他走了。』尤三姐道：『我奉紀子之命，等候已久。今兒見了，必定要一劍斬斷你的塵緣。』寶釵聽了，益發着忙，又不信這些話，到底是什麼意思，只得回頭要詢，豈知身後說話的並非別人，卻是晴雯。寶釵一見，悲喜交加，便說：『我一個人走迷了道兒，遇見仇人，我要逃回，卻不見你們。』一人跟着我，如今好了。晴雯姐姐，快快的帶我回家去罷。晴雯道：『侍者不必多疑，我非晴雯，我是奉紀子之命，特來請你。』一會並不勝為你。寶釵滿腹狐疑，只得問道：『姐姐，說是紀子叫我，我那紀子，究竟是何人？』晴雯道：『此時不必問了，那裡自然知道。寶釵沒法，只得跟着走。細看那人，背後舉動，恰是晴雯，那面目聲音，是不錯的。』怎麼她說不是我？此時心裡模稜，且別管她，到了那邊，見了紀子，就有不是，那時再求她，到底女人的心腸，是慈悲的，必是怨我，我失正想，不多時到了，一個所在，只見殿宇精緻，色彩輝煌，庭中一叢翠竹，戶外數本蒼松，廊簷下立着幾個侍女，都是宮妝打扮，見了寶釵進來，便悄悄的說道：『這就是神瑛侍者，怎麼引着寶釵的？』說道：『就是你快進去通報罷。』有一侍女笑着，招手，寶釵便跟着進去，過了幾層房舍，見一正房，珠簾高掛，那侍女說：『站着候着。』寶釵聽了，也不敢則聲，只得在外等着。那侍女進去不多時，出來，說：『侍者，參見。』又有一人，擡起珠簾，只見一女子，頭戴花冠，身穿縐紗，端坐在內。寶釵略一抬頭，見是黛玉，的形容，便不識的，說道：『妹妹，在這裡，叫我，我好想。』那屋外的侍女，悄說道：『侍者，無禮，快快出去。』說猶未了，又見一個侍兒，將珠簾放下。寶釵此時，欲待進去，又不放，走又不捨，待要問明，見那些侍女，並不認得，又被驅逐，無奈出來，心想：『要問晴雯，回頭，四顧並不見有晴雯，心下狐疑，只得快快出來，又無人引着，正欲找原路而去，卻又找不出舊路了。正在為難，見鳳姐站在一所房簷下，招手，寶釵看見，喜歡，道：『可好了，原來回到自己家裡了。我怎麼一時迷亂，如此急弄。』前來說：『姐姐，在這裡，我被這些人捉弄，到這個份兒，林妹妹又不肯見我，不知何原故。』說着，走到鳳姐站的地方，細看起來，並不是鳳姐，原來卻是寶釵的前妻秦氏。寶釵只得立住，那要問鳳姐，在那裡，那秦氏也不管，竟自往屋裡去了。寶釵恍惚惚的，又不敢跟進去，只得呆呆的站着，說道：『我今兒得了什麼，不是，眾人都說，我便痛哭起來，見有幾個黃巾力士，執鞭趕來，說是何處男人，敢闖入我們這天仙福地，來快走出去。』寶釵聽得，不敢言語，正要尋路出來，遠遠望見一群女子，說笑，前來，寶釵一看，時，像有迎春等一千人，走來，心裡喜歡，叫道：『我迷住，在這裡，你們快來救我。』正着，後面力士趕來，寶釵急得，往前亂跑，忽見那一群女子，都變作鬼怪形象，也來追捕。寶釵正在情急，只見那送玉來的和尚，手裡拿着一面鏡子，一面說道：『我奉元妃娘娘旨意，特來救你。』登時鬼怪全無，仍是一片荒郊。寶釵拉着和尚，說道：『我記得是你，領我到這裡，你一時又不見了，看見了好些人，只是都不理我。』忽又變作鬼怪，到底是夢，是真？望老師，明白指示。那和尚道：『你到這裡，曾偷看什麼東西？』寶釵一想，道：『他既能帶我到天仙福地，自然也是神仙了。如何瞞得他？況且正要問個明白，便道：『我倒見了好些冊子，來着。』那和尚道：『可又來，你見了冊子，還不解，塵世上的情緣，都是那些魔障，只要把歷過的事情，細細記着，將來我與你說明，說着，把寶釵狠命的一推，說回去罷。』寶釵站不住，腳一交，跌倒，口裡嚷道：『喇嘛王夫人等，正在吳拉，聽見寶釵，蘇來，連忙叫喚。』寶釵睜眼看時，仍躺在炕上，見王夫人，寶釵等哭的眼泡，紅腫，定神一想，心裡說道：『是了，我是死去過來的，迷把神魂，所歷的事，呆呆的，細想，幸喜多遠，記得便，哈哈的，笑，是了，是了，王夫人，只道舊病復發，便好延醫，調治，即命丫頭，婆子，快去告訴，寶釵，說是寶釵，回過來了。』寶釵迷住了，如今說出話來，不用催，辦妥了，寶釵聽了，即忙進來，看視，果見寶釵，蘇來，便道：『沒的，嚇兒，你要唬死誰？』說着，眼淚也不，不知不覺，滾下來了，又歡了幾口氣，仍出去，叫人請醫生，診脈，服藥。這裡，月正，思自，盡見寶釵，一過來，也放了心，只見王夫人，叫人，端了，桂圓湯，叫他，吃了，幾口，漸漸的，定了，神，王夫人等，放心，也沒有，說，月，只叫人，仍把那玉，交給寶釵，給他，帶上，想起，那和尚，來，這玉，不知，哪裡，找來的，也是，古怪，怎麼，一時，要，一時，又不見了，莫非，是神仙，不成？寶釵，道：『說起，那和尚，來的，說，辭，去，的影響，那玉，並不是，找來的，頭，丟的時候，必是，那和尚，取去的。』王夫人，道：『玉在家裡，怎麼，能取的，了去？』寶釵，道：『既可，送來，就可，取去，襲人，那月，道：『那年，丟了，玉，林大，鬱，了，個字，後來，二奶奶，過了門，我，還告訴，過二奶奶，說，別的那字，是什麼？』寶釵，道：『二奶奶，還記得，麼？』寶釵，道：『是了，你們，說，別的是，當鋪，裡，找去，如今，才，明白了，竟是個和尚，的，尚字，在上頭，可不是和尚，取了。』

接著我用tm函式庫將分詞後的結果建立成文本：

```
1 library(tm)
2 mycorpus <- Corpus(DirSource("RedDream.segment.2016-11-27_17_49_03"),
3 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：

```

1 #Setting Chinese font on word cloud
2 par(family=("Heiti TC Light"))
3
4 wordcloud(cloud1$word, cloud1$freq, min.freq = 100, random.order = F,
5 colors = rainbow(length(row.names(mtr1))))
6
7 wordcloud(cloud2$word, cloud2$freq, min.freq = 100, random.order = F,
8 colors = rainbow(length(row.names(mtr2))))

```



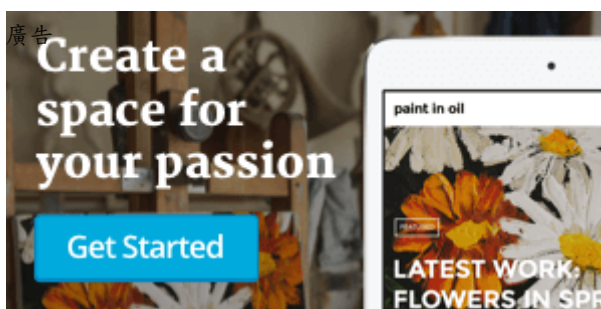
第一頁是前八十回的結果，第二頁是後四十回的結果。排除出現頻率最高的“你、我、他、的、了”等等這些主受詞或語助詞，可以看出來在橘色部份的连接詞或介系詞，這種牽涉到個人寫作習慣的詞在分佈上有些許的不同，因此其實不難看出兩者的差異。

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 ↑