



文字雲 in R 以PTT美食版(Food版)為例

主講人:莊舒媛



About me

莊舒媛 / Sharon

研究領域及專長

- » 文字探勘
- » 網路爬蟲
- » 關聯分析
- » 類神經網路預測分析
- » 主題模型分析

聯絡方式

sindy801230@gmail.com





Outline

- » 中文分詞介紹
- » 文字雲相關套件介紹
- » PTT-美食版應用案例分享
- » Q&A





1.


中文分詞介紹





中文分詞

在分詞過程中，英文與中文分詞有很大的不一樣！對英文來說，基本上以空白切分一個單字為詞；對中文來說卻沒有明顯的切分方式，可能的方式是透過字典進行分詞，如果在字典中找到一樣的詞彙可將之切分為一個詞，從最多字的詞彙開始搜尋再慢慢遞減字數。

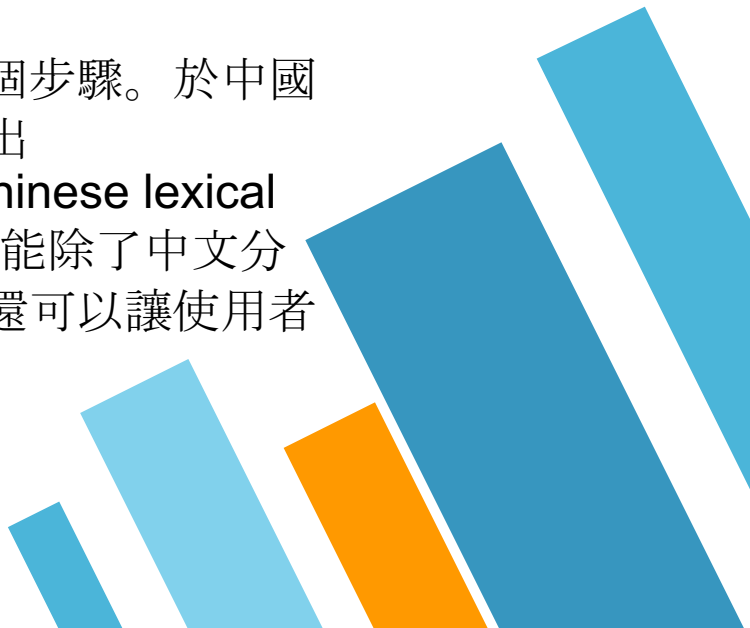




中文分詞

- **ICTCLAS**中文分詞

中文分詞方法是在中文文字處理中最關鍵的一個步驟。於中國科學院計算技術研究所在多年的研究後，研究出 **ICTCLAS**(institute of computing technology, chinese lexical analysis system)中文語法分析系統，其主要功能除了中文分詞以外，還有詞性標註與命名實體識別，甚至還可以讓使用者自行增加字典與使用繁體中文。

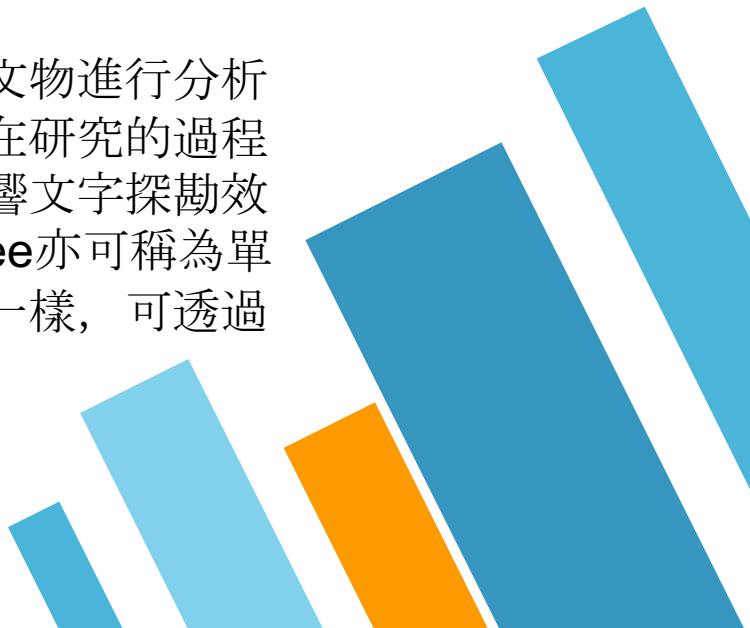




中文分詞

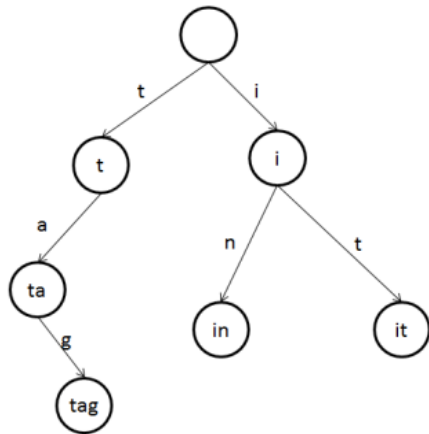
- **Trie Tree**結構

中文文字探勘是一個極重要的部分，可以針對文物進行分析來找出文物關鍵特性，對文章進行研究分析，在研究的過程中，分詞是一個很關鍵的步驟，它會直接的影響文字探勘效果。透過使用**Trie Tree**結構進行分詞，**Trie Tree**亦可稱為單詞搜尋樹，是一種樹狀資料結構，就如同字典一樣，可透過查詢字母找到字串。



中文分詞

- **Trie Tree**結構






中文分詞

- 詞項文件矩陣

依據本研究所建立的字庫，將其中的關鍵字與古文物進行索引後，透過tm套件建立詞項文件矩陣(Term-Document Matrix, TDM)，呈現古文物與關鍵字之間的關聯。其中以列(row)代表關鍵字，而行(column)表文件(document)。



中文分詞

- 詞項文件矩陣

document \ Term	Doc6	Doc7	Doc8	Doc9	Doc10
咖啡	0	1	1	1	1
拉麵	0	1	0	0	0
pizza	0	0	1	0	0
牛肉	0	0	0	0	1
麵包	0	0	0	1	0



2.


文字雲相關套件介紹





相關套件

套件	資料處理與分析方法
tm	建立語料庫與詞項文件矩陣
tmcn	簡字與繁體轉換、清除停止詞
Rwordseg	中文文字斷詞
wordcloud	文字雲繪製





3.

PTT-美食版應用案例分享



```
> d.corpus <- Corpus(DirSource("~/Desktop/Food/"), list(language = NA))
```

```
> content(d.corpus[[1]])
```

```
[1] "作者Dilbert (嘻)看板Food標題[公告] 板面刪文後處置措施定案時間Sat Oct 1 12:17:05 2005"
```

```
[2] "◆投票結果:(共有 220 人投票,每人最多可投 1 票)"
```

```
[3] "      選      項                                總票數  得票率  得票分布"
```

```
[4] "      刪文後於板面註記(原制度)                    88 票  40.00%  40.00%"
```

```
[5] "      ◎刪文移置資源回收桶(新方案)                132 票  60.00%  60.00%"
```

```
[6] ""
```

```
[7] " 投票結果已確定,未來板面刪文後處置措施將使用下列方案:"
```

```
[8] ""
```

```
[9] " 刪文後移置資源回收桶,板面不註記(新方案)"
```

```
[10] ""
```

```
[11] " 作法:"
```

```
[12] "      1.精華區內設置資源回收桶,子目錄則為各類刪文理由,投票結果已確定,未來板面刪文後處置措施將使用下列方案:
```

M.1128132666.A.0FD.txt

作者Dilbert (嘻)看板Food標題[公告] 板面刪文後處置措施定案時間Sat Oct 1 12:17:05 2005

◆投票結果:(共有 220 人投票,每人最多可投 1 票)

選 項	總票數	得票率	得票分布
刪文後於板面註記(原制度)	88 票	40.00%	40.00%
◎刪文移置資源回收桶(新方案)	132 票	60.00%	60.00%

刪文後移置資源回收桶,板面不註記(新方案)

作法:

- 1.精華區內設置資源回收桶,子目錄則為各類刪文理由,板面上的違規文章遭刪除後將置入資源回收桶內各子目錄。
- 2.版面配套措施:屆時被刪除的文章標題,將由板主手動改為全無內容的空洞,以與板友自刪的文章作區隔。
- 3.由於板面註記將改為空洞,故未來板主不再提供精華區路徑指引服務。
- 4.欲確認刪除者為哪一位板主,請檢視該文章的編選者便可得知。

編號	標 題	編 選 者	選 日 期
1.	◇ [請益] 請問高雄哪裡有不錯的合菜餐廳呢	bluefish	[09/17/05]
2.	◇ 請問一下-台北市哪裡有好吃的壓(捷運附近or23	Dilbert	[09/18/05]

Q & A 時間:

- 1.首先感謝許多板友們的鼓勵,在這個管理環境越趨艱難的當下,



```
#### Set the stopwords and remove them
```

```
myStopWords <- c(toTrad(stopwordsCN()), stopwords("english"), "編輯", "時間",  
"標題", "發信", "實業", "作者")
```

```
d.corpus <- tm_map(d.corpus, removeWords, myStopWords)  
content(d.corpus[[1]])
```

```
#### Building bag of words model(TF-IDF)
```

```
tdm <- TermDocumentMatrix(d.corpus,  
                           control = list(wordLengths = c(2, Inf),  
                                           weighting = function(x)  
                                             weightTfIdf(x, normalize = FALSE)))
```

```
tdm
```



```
#### Wordcloud
```

```
library(wordcloud)
```

```
m <- as.matrix(tdm)
```

```
v <- sort(rowSums(m), decreasing = TRUE)
```

```
d <- data.frame(word = names(v), freq = v)
```

```
#d <- data.frame(names(v))
```

```
#write.table(d, file = "Food.csv", sep = ",", row.names = F, col.names = F, fileEncoding = "big5")
```

```
par(family = "STKaiti") ## only for Mac OS
```

```
wordcloud(d$word, d$freq, min.freq = 50, random.order = F, ordered.colors = F,  
          colors = rainbow(length(row.names(m))))
```





Q & A

