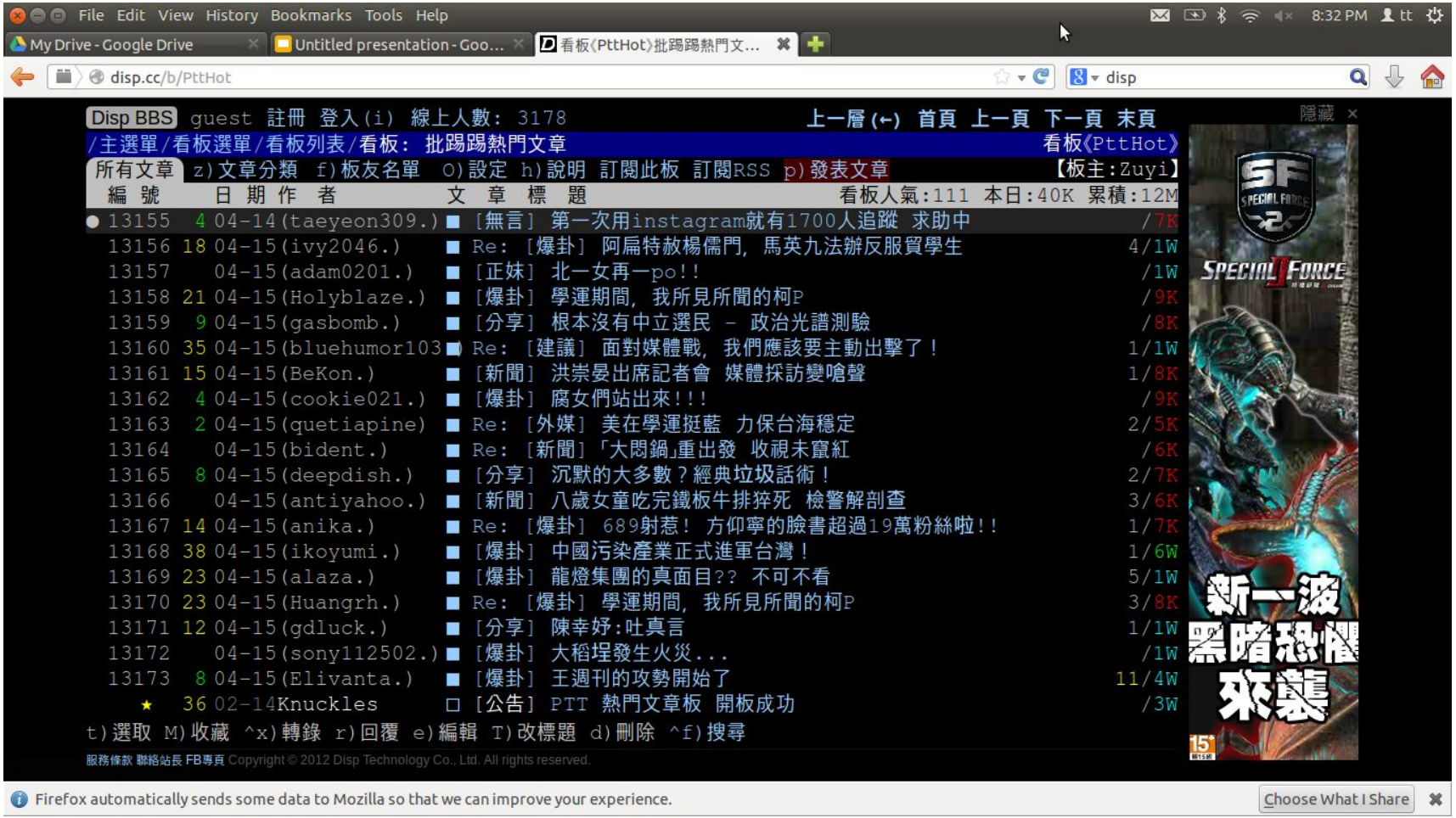
R語言推廣講座
Text Mining with R

陳嘉葳 04/19/2014

Text Mining 是什麼 ？

可以做什麼 ？


File Edit View History Bookmarks Tools Help
8:32 PM tt
My Drive - Google Drive
Untitled presentation - Goo...
看板《PttHot》批踢踢熱門文...
disp.cc/b/PttHot
disp
Disp BBS guest 註冊 登入(i) 線上人數: 3178
上一層(←) 首頁 上一頁 下一頁 末頁
/主選單/看板選單/看板列表/看板: 批踢踢熱門文章
看板《PttHot》
所有文章 z)文章分類 f)板友名單 O)設定 h)說明 訂閱此板 訂閱RSS p)發表文章
【板主:Zuyi】
編號 日期 作者 文章標題 看板人氣:111 本日:40K 累積:12M
13155 4 04-14(taeyeon309.) ■ [無言] 第一次用instagram就有1700人追蹤 求助中 /7K
13156 18 04-15(ivy2046.) ■ Re: [爆卦] 阿扁特赦楊儒門, 馬英九法辦反服貿學生 4/1W
13157 04-15(adam0201.) ■ [正妹] 北一女再一po!! /1W
13158 21 04-15(Holyblaze.) ■ [爆卦] 學運期間, 我所見所聞的柯P /9K
13159 9 04-15(gasbomb.) ■ [分享] 根本沒有中立選民 - 政治光譜測驗 /8K
13160 35 04-15(bluehumor103) ■ Re: [建議] 面對媒體戰, 我們應該要主動出擊了! 1/1W
13161 15 04-15(BeKon.) ■ [新聞] 洪崇晏出席記者會 媒體採訪變嗆聲 1/8K
13162 4 04-15(cookie021.) ■ [爆卦] 腐女們站出來!!! /9K
13163 2 04-15(quetiapine) ■ Re: [外媒] 美在學運挺藍 力保台海穩定 2/5K
13164 04-15(bident.) ■ Re: [新聞]「大悶鍋」重出發 收視未竄紅 /6K
13165 8 04-15(deepdish.) ■ [分享] 沉默的大多數?經典垃圾話術! 2/7K
13166 04-15(antiyahoo.) ■ [新聞] 八歲女童吃完鐵板牛排猝死 檢警解剖查 3/6K
13167 14 04-15(anika.) ■ Re: [爆卦] 689射惹! 方仰寧的臉書超過19萬粉絲啦!! 1/7K
13168 38 04-15(ikoyumi.) ■ [爆卦] 中國污染產業正式進軍台灣! 1/6W
13169 23 04-15(alaza.) ■ [爆卦] 龍燈集團的真面目?? 不可不看 5/1W
13170 23 04-15(Huangrh.) ■ Re: [爆卦] 學運期間, 我所見所聞的柯P 3/8K
13171 12 04-15(gdluck.) ■ [分享] 陳幸妤:吐真言 1/1W
13172 04-15(sony112502.) ■ [爆卦] 大稻埕發生火災... /1W
13173 8 04-15(Elivanta.) ■ [爆卦] 王週刊的攻勢開始了 11/4W
★ 36 02-14Knuckles □ [公告] PTT 熱門文章板 開板成功 /3W
t)選取 M)收藏 ^x)轉錄 r)回覆 e)編輯 T)改標題 d)刪除 ^f)搜尋
服務條款 聯絡站長 FB專頁 Copyright © 2012 Disp Technology Co., Ltd. All rights reserved.
隱藏
SF SPECIAL FORCE 2
SPECIAL FORCE
新一波
黑暗恐懼
來襲
15+
Firefox automatically sends some data to Mozilla so that we can improve your experience.
Choose What I Share

可以分析一下 ...

最近 ptt 提到服貿的文章在說些什麼 ？

最近 ptt 提到林非帆、陳為廷的文章在說些什麼 ？

第一步

先蒐集文章 ！！

把這幾天 ptt 的文章都抓下來


doc
8:49 PM
tt
Devices
系統保留
93 GB Filesy...
DATA
UDF Vol...
Computer
Home
Desktop
Documents
Downloads
Music
Pictures
Videos
File System
Trash
Network
Browse Net...
Home
Downloads
doc
Search
抓下來的一堆文章
我們稱為語料庫

```
> d.corpus <- Corpus(VectorSource(doc))
> d.corpus[[1]]   # 第1篇文章
```

永豐金(2890)晉升百億俱樂部，去年(2013)稅後淨利來到103億元，年增率超過1成，董事長何壽川表示，台灣金融市場結構產生變化，強勢貨幣存款占比明顯提高，做好商品和擴大規模間孰重孰輕需考慮清楚，不排斥合併，但重點是能否把品質做好。

對於開發金宣布購併萬泰銀，為金融整併開了第一槍，何壽川表示，永豐金當然有意願進行併購，但金融整併除了規模之外，更要注意到由於金融結構產生重大變化，強勢貨幣已占有非常重要的成分，因此，現階段做好商品與提高規模究竟孰重孰輕，業者必須慎重考量。

.....

要找出 提到 “服貿” 的文章在說些什麼

等於

找出 提到 ”服貿” 的這些文章, 經常出現哪些 ”詞彙”

ex. 立法院、30秒、投票 …

下面這句話有哪些詞彙 ？

官員表示昨天會議是就外界關切的特定議題範圍受影響產業等各個面向

ex. 官員、昨天、會議 ...

如何用程式 斷出這些詞彙 ？

官員 表示 昨天 會議 是 就 外界 關切 的 特定 議題 範圍 受影響 產業 等 各個 面向

中文斷詞套件 Rwordseg

```
> segmentCN('官員表示昨天會議是就外界關切的特定議題範圍受影響產業等各個面向')

[1] "官員" "表示" "昨天" "會議" "是" "就" "外界" "關切"
    "的" "特定" "議題" "範圍"
[13] "受" "影響" "產業" "等" "各個" "面向"
```

將蒐集來的 ptt 文章都經過中文斷詞

然後放到同一張 table (matrix)

方便一起做處理

```
tdm <- TermDocumentMatrix(d.corpus)
inspect(tdm[6050:6054, 1:10])
```

```
                    Docs
Terms          1 2 3 4 5 6 7 8 9 10
  黨中央       0 0 0 0 0 0 0 1 1  0
  黨代表       0 0 0 0 0 0 0 0 0  0
  黨外人士     0 0 0 0 0 0 0 0 0  0
  黨政軍       0 0 0 0 0 0 0 0 0  0
  齊頭並進     0 0 0 0 0 0 0 0 0  0
```

在文章中出現的次數

	Docs	
Terms	1 2 3 4 5 6 7 8 9 10	
黨中央	0 0 0 0 0 0 0 1 1 0	可以用 slam::row_sums 來算出每個字的總出現次數
黨代表	0 0 0 0 0 0 0 0 0 0	
黨外人士	0 0 0 0 0 0 0 0 0 0	
黨政軍	0 0 0 0 0 0 0 0 0 0	
齊頭並進	0 0 0 0 0 0 0 0 0 0	tm::findFreqTerms 直接找最常出現的字

	Docs									
Terms	1	2	3	4	5	6	7	8	9	10
黨中央	0	0	0	0	0	0	0	1	1	0
黨代表	0	0	0	0	0	0	0	0	0	0
黨外人士	0	0	0	0	0	0	0	0	0	0
黨政軍	0	0	0	0	0	0	0	0	0	0
齊頭並進	0	0	0	0	0	0	0	0	0	0

兩兩之間可以
算相關係數
or
歐幾里德距離

findAssocs(tdm, '黨代表', 0.2)

	黨代表
協議書	0.87
總動員	0.60
始作俑者	0.52
第二服	0.46
清清楚楚	0.33
白紙黑字	0.23
學生會	0.22

相關度大於 0.2 以上

所以可以來分析

“服貿” 新聞 經常出現哪些詞彙了嗎 ？

> findAssocs(tdm, '服貿協議')

Error in is.numeric(corlimit) : 'corlimit' is missing

> findAssocs(tdm, '服貿')

Error in is.numeric(corlimit) : 'corlimit' is missing

WHY ？

因為 ...

> segmentCN('反服貿學運退場，外界對於服貿協議的未來有許多討論')
 [1] "反"　"服"　"貿"　"學"　"運"　"退場" "外界" "對"　"於"　"服"　"貿"　"協議"
[13] "的"　"未來" "有"　"許多" "討論"

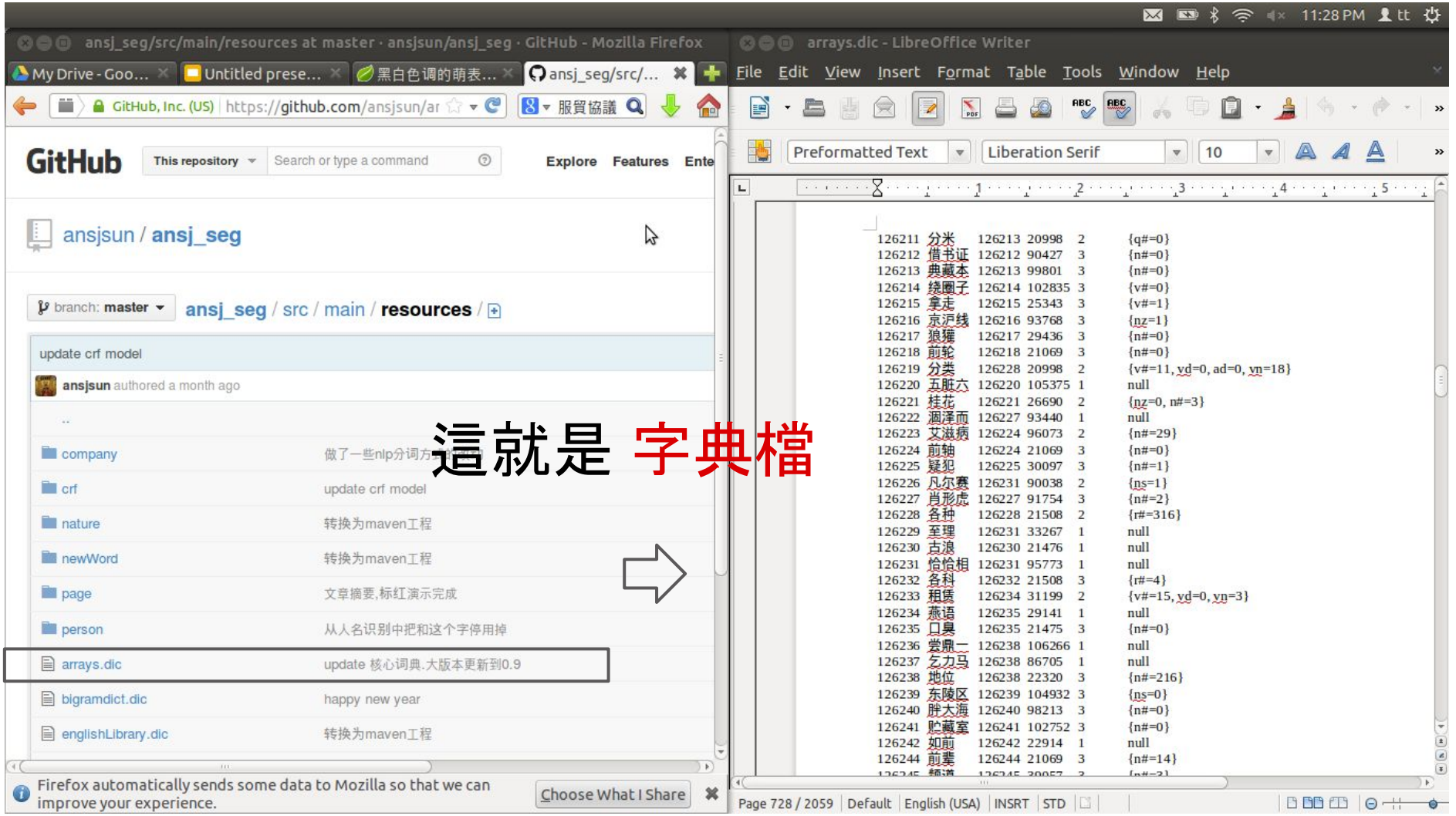
> segmentCN('台灣史上頭一遭，國會殿堂遭學生霸佔！你了解什麼是服貿嗎？')
 [1] "台灣" "史"　"上"　"頭"　"一遭" "國會" "殿堂" "遭"　"學生" "霸佔" "你"　"了解"
[13] "什麼" "是"　"服"　"貿"　"嗎"

Rwordseg 的字典檔 沒有 “服貿協議”、”服貿” 這些字

所以剛才統計的 詞彙 錯很大

想當然 …

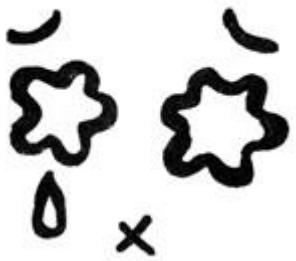
也關聯不出 林非帆、陳為廷、江宜樺 這些字


ansj_seg/src/main/resources at master · ansjsun/ansj_seg · GitHub - Mozilla Firefox
arrays.dic - LibreOffice Writer
GitHub
ansjsun / ansj_seg
branch: master
ansj_seg / src / main / resources /
update crf model
ansjsun authored a month ago
company
crf
update crf model
nature
转换为maven工程
newWord
转换为maven工程
page
文章摘要,标红演示完成
person
从人名识别中把和这个字停用掉
arrays.dic
update 核心词典.大版本更新到0.9
bigramdict.dic
happy new year
englishLibrary.dic
转换为maven工程
這就是 字典檔
126211 分米 126213 20998 2 {q#=0}
126212 借书证 126212 90427 3 {n#=0}
126213 典藏本 126213 99801 3 {n#=0}
126214 绕圈子 126214 102835 3 {v#=0}
126215 拿走 126215 25343 3 {v#=1}
126216 京沪线 126216 93768 3 {nz=1}
126217 狼獾 126217 29436 3 {n#=0}
126218 前轮 126218 21069 3 {n#=0}
126219 分类 126228 20998 2 {v#=11, vd=0, ad=0, vn=18}
126220 五脏六 126220 105375 1 null
126221 桂花 126221 26690 2 {nz=0, n#=3}
126222 涸泽而 126227 93440 1 null
126223 艾滋病 126224 96073 2 {n#=29}
126224 前轴 126224 21069 3 {n#=0}
126225 疑犯 126225 30097 3 {n#=1}
126226 凡尔赛 126231 90038 2 {ns=1}
126227 肖形虎 126227 91754 3 {n#=2}
126228 各种 126228 21508 2 {r#=316}
126229 至理 126231 33267 1 null
126230 古浪 126230 21476 1 null
126231 恰恰相 126231 95773 1 null
126232 各科 126232 21508 3 {r#=4}
126233 租赁 126234 31199 2 {v#=15, vd=0, vn=3}
126234 燕语 126235 29141 1 null
126235 口臭 126235 21475 3 {n#=0}
126236 赏鼎一 126238 106266 1 null
126237 乞力马 126238 86705 1 null
126238 地位 126238 22320 3 {n#=216}
126239 东陵区 126239 104932 3 {ns=0}
126240 胖大海 126240 98213 3 {n#=0}
126241 贮藏室 126241 102752 3 {n#=0}
126242 如前 126242 22914 1 null
126244 前辈 126244 21069 3 {n#=14}
Firefox automatically sends some data to Mozilla so that we can improve your experience.
Choose What I Share
Page 728 / 2059
Default
English (USA)
INSRT
STD

字典檔沒有 服貿相關 的辭彙

怎麼辦 ？？

自己手動把這些字加到字典檔吧 ..



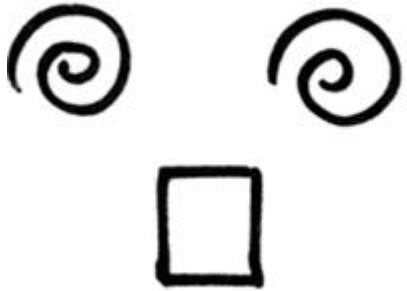
```
> insertWords(c("服贸协议", "服贸"))  # 要用簡體字才有用
```

```
> segmentCN('反服貿學運退場，外界對於服貿協議的未來有許多討論')
[1] "反"   "服貿"   "學"     "運"     "退場"   "外界"   "對"
[8] "於"   "服貿協議" "的"     "未來"   "有"     "許多"   "討論"
```

```
> segmentCN('台灣史上頭一遭，國會殿堂遭學生霸佔！你了解什麼是服貿嗎？')
[1] "台灣" "史"   "上"   "頭"   "一遭" "國會" "殿堂" "遭"   "學生" "霸佔" "你"   "了解"
[13] "什麼" "是"   "服貿" "嗎"
```

可是 服貿新聞 那麼多 新的關鍵字

要手動加到什麼時候阿 ！！！



我們來寫個程式

自動從一堆文章裡 挖掘新詞彙 吧 ！！

如何用程式從下面句子自動抓出各種不同詞彙 ？

'反服貿學運退場，外界對於服貿協議的未來有許多討論'

ex. 反服貿、學運、服貿協議 ...

```
> NGramTokenizer(x, Weka_control(min =2, max =4))
 [1] "反 服 貿 學" "服 貿 學 運" "貿 學 運 退" "學 運 退 場" "運 退 場 外" "退 場 外 界"
 [7] "場 外 界 對" "外 界 對 於" "界 對 於 服" "對 於 服 貿" "於 服 貿 協" "服_貿_協_議"
[13] "貿 協 議 的" "協 議 的 未" "議 的 未 來" "的 未 來 有" "未 來 有 許" "來 有 許 多"
[19] "有 許 多 討" "許 多 討 論" "反_服_貿" "服 貿 學"    "貿 學 運"    "學 運 退"
[25] "運 退 場"    "退 場 外"    "場 外 界"    "外 界 對"    "界 對 於"    "對 於 服"
[31] "於 服 貿"    "服 貿 協"    "貿 協 議"    "協 議 的"    "議 的 未"    "的 未 來"
[37] "未 來 有"    "來 有 許"    "有 許 多"    "許 多 討"    "多 討 論"    "反 服"
[43] "服 貿"   "貿 學"   "學_運"   "運 退"   "退 場"   "場 外"
[49] "外 界"   "界 對"   "對 於"   "於 服"   "服 貿"   "貿 協"
[55] "協 議"   "議 的"   "的 未"   "未 來"   "來 有"   "有 許"
[61] "許 多"   "多 討"   "討 論"
```

'反服貿學運退場 外界對於服貿協議的未來有許多討論'

將每一篇文章都做過 NGram 之後

再把這些 (候選)詞彙 加總 出現次數

服貿 、服貿協議、立法院 … etc

就會是排名很前面的詞彙 ！

可是 ！！

問題沒那麼簡單 ...

會遇到兩種狀況

第1種情況

是服貿　　協議的

出現次數非常多 ！

要怎麼消除 ？？

判斷 是服貿 是不是偶然才拼湊在一起

等於

判斷 是 + 服貿 為 機率 **獨立** ？

P(是) * P(服貿) 接近 P(是服貿)？

$$\frac{P(\text{是服貿})}{P(\text{是}) * P(\text{服貿})}$$

如果機率獨立的話會很接近

高於 門檻值 才保留

P(是服貿)、P(是) 、 P(服貿)

機率

前面的 NGram 計算相對次數得到

```
> words <- slam::row_sums(t.tdm)
> words_freq <- words / sum(words)
```

```
> tail(words_freq, 4)
     龐巴          龐巴迪         龐巴迪的       龐巴迪的關
3.687656e-06 3.687656e-06 3.687656e-06 3.687656e-06
```

```
segmentWord <- function(word){
    n <- nchar(word)-1
    seg <- lapply(1: n, function(i){
        w1 <- substr(word, 1, i)
        w2 <- substr(word,i+1, n+1)
        c(w1,w2)
    })
    return(seg)
}
```

```
> segmentWord('是服貿')
[[1]]
[1] "是"   "服貿"

[[2]]
[1] "是服" "貿"
```

第2種情況

服貿協　　貿協議　　貿協

出現次數也非常多 ！

要怎麼消除 ？？

右邊幾乎只能接
議

左邊幾乎只能接
服

完整詞彙的左右兩側，應該要能搭配很多其他詞彙 (訊息)

%$%^ + 服貿協議 + ^%$%^

$%# + 服貿協 + 議

需要一個衡量指標

衡量詞彙左右兩側 提供的訊息

Entropy

Entropy ⇨ 用來衡量訊息的資訊含量

Shannon

1948年 通訊的數學原理

下面哪個訊息你覺得比較有資訊含量 ??

1. 今天新聞會報導社會事件!

2. 你今天中了樂透！

1. 今天新聞會報導社會事件

⇨ 發生機率太高, 無感 !!

2. 你今天中了樂透！

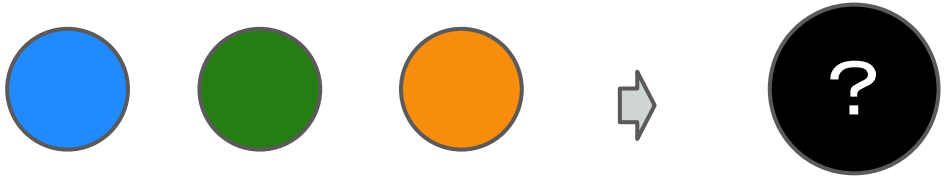
⇨ 因為不容易發生, 驚喜 !!

100 % 發生的事件, 資訊含量是 0

⇨ - log (1) = 0

不容易發生的事件, 資訊含量較高

⇨ - log (1/1024) = 10

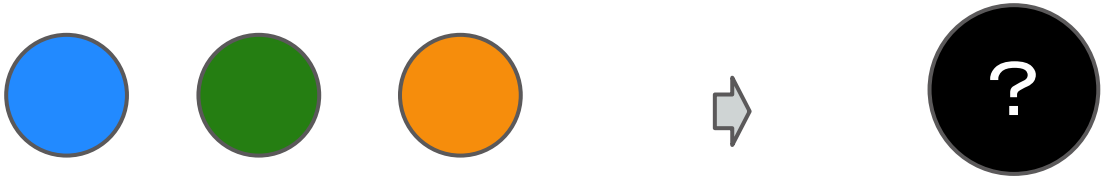


球的顏色數量這則訊息，提供多少資訊含量？

-(¼*log(¼) + 1/2 *log(1/2) + ¼*log(¼)) = 1.5

把球的例子換成一本書的辭彙數量

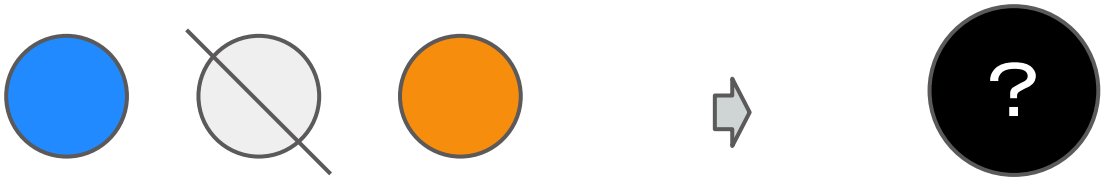
一本書的訊息含量



4　　8　　4

抽出來的球不是綠色 ！　　(補充)

這則訊息的資訊含量是多少？



4 4

$$-(½*\log(½) + ½*\log(½)) = 1$$

抽出來的球不是綠色

抽出來的球不是綠色

降低了不確定性

其中蘊含的資訊量為 0.5

```
w1 <- grep(paste('^', 服貿協, sep=''), doc, value=TRUE)
pre <- mean( -log2(words_freq[w1]) )  # 計算 Entropy

w2 <- grep(paste(貿協議, '$', sep=''), doc, value=TRUE)
post <- mean( -log2(words_freq[w2]) )  # 計算 Entropy

return( min(pre ,post) )  # 取較小的 Entropy
```

每個 (候選)詞彙都算出左右兩邊的 entropy

取較低那一側的值

若高於門檻值才保留這個詞彙 ！

> topEntropyWords # 最後結果

[1] "陳其邁" "服務貿易" "20" "召委" "立法" "對台" "表示" "朝野" "協定"
[10] "兩岸服貿" "支持" "國民黨" "台灣" "民進黨" "兩岸" "一定" "完成" "政治"
[19] "經濟" "談判" "張慶忠" "江宜樺" "公聽會" "30" "國家" "競爭" "協商"
[28] "貿易" "開放" "進入" "對台灣" "市場" "立委" "行政院" "合作" "實質"
[37] "服務" "希望" "台北" "中國" "加入" "自由" "政院" "指出" "tpp"
[46] "中央" "影響" "如果" "民眾" "大陸" "一個" "人民" "相關" "王金平"
[55] "沒有" "發展" "強調" "媒體" "中國大陸" "簽署" "國民黨團" "認為" "立院"
[64] "投資" "目前" "今天" "不能" "我們" "政府" "國際" "金融" "問題"
[73] "不是" "廣告" "對於" "國內" "香港" "最後" "總統" "去年" "產業"
[82] "政策" "可能" "進行" "可以"

把這些詞彙加入字典後

才能正確算出詞彙之間的關聯 ！！

做出來的分析也比較正確

```
writeLines(insWrods, 'ecfa.txt')
loadDict('/home/tt/ecfa.txt')
```

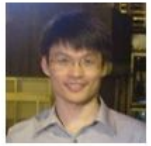
可以每隔幾天計算一次

和上次的結果做 差集

就能得到 最新出現 的 關鍵字 ！

有適當的新詞彙之後

就能來分析一些 關鍵字之間的關聯



CHENG-JUN WANG

东风夜放花千树：对宋词进行主题分析初探

邱怡轩在统计之都中展示了对宋词进行的分析（参见http://cos.name/tag/%E5%AE%8B%E8%AF%8D/），因为当时缺乏中文分词的工具，他独辟蹊径，假设宋词中任意两个相邻的汉字构成一个词语，进而找到了宋词当中的高频词。本文则尝试使用他所提供的宋词语料（http://cos.name/wp-content/uploads/2011/03/SongPoem.tar.gz），分析一下使用R进行中文分词、构建词云、高频词语聚类以及主题模型分析。

首先要载入使用的R包并读入数据。

```
library(Rwordseg)
require(rJava)
library(tm)
library(slam)
library(topicmodels)
library(wordcloud)
library(igraph)
setwd("D:/github/text mining/song") # 更改为你的工作路径，并存放数据在此。
txt=read.csv("SongPoem.csv",colClasses="character")
```

然后进行对数据的操作。当然，第一步是进行中文分词，主要使用Rwordseg这个R包，其分词效果不错。分词的过程可以自动去掉标点符号。

```
poem_words <- lapply(1:length(txt$Sentence), function(i) segmentCN(txt$Sen
```

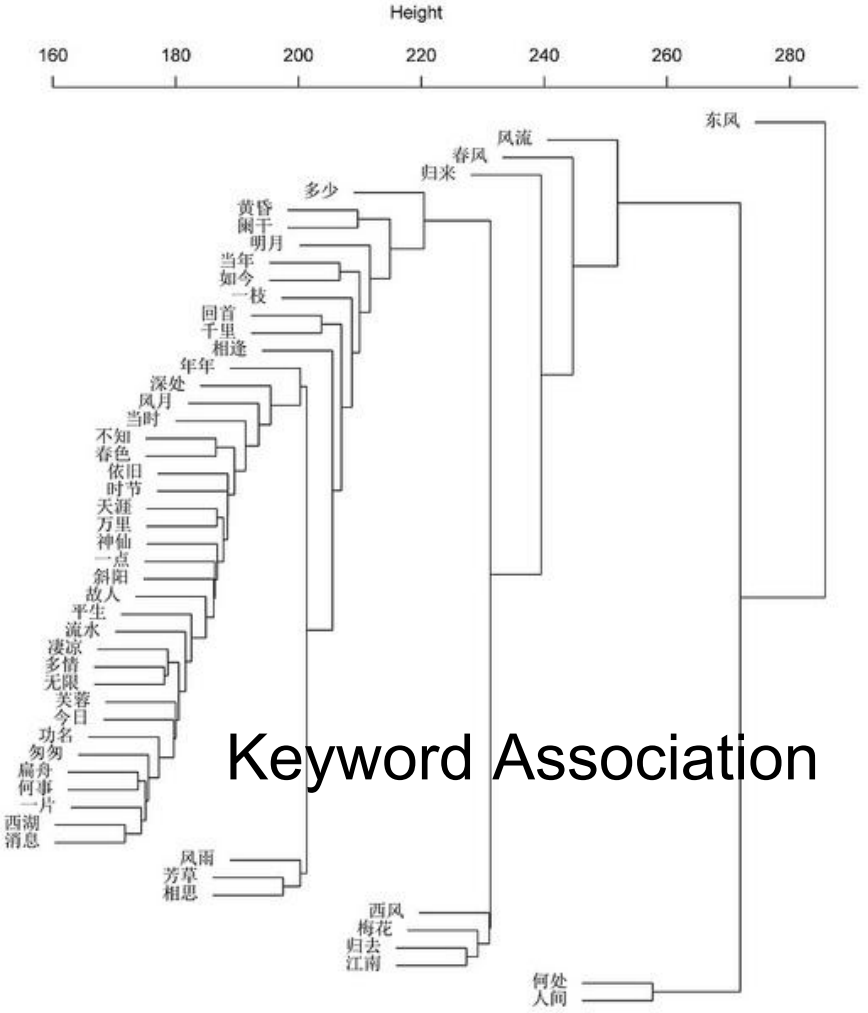
然后，我们将数据通过tm这个R包转化为文本-词矩阵（DocumentTermMatrix）。

wordcorpus <- Corpus(VectorSource(poem_words), encoding = "UTF-8") # 组成语料库格式

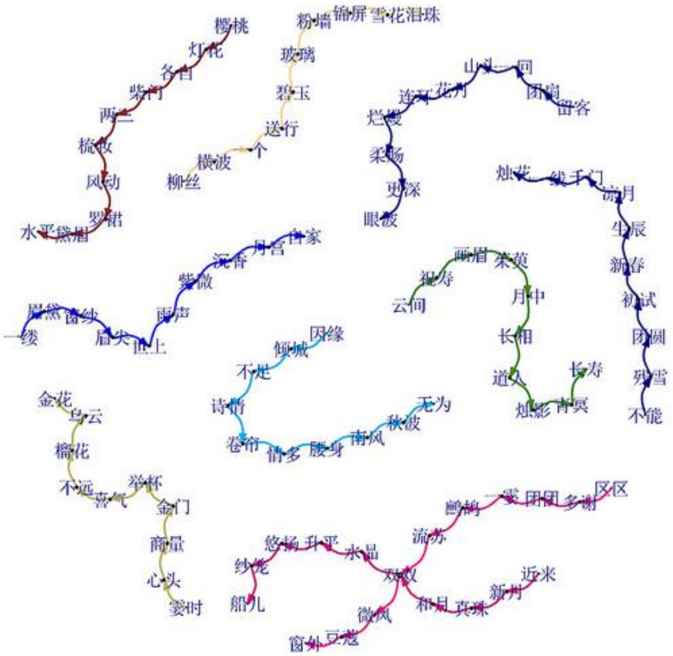
```
Sys.setlocale(locale="Chinese")
dtm1 <- DocumentTermMatrix(wordcorpus,
```

Keyword Cloud

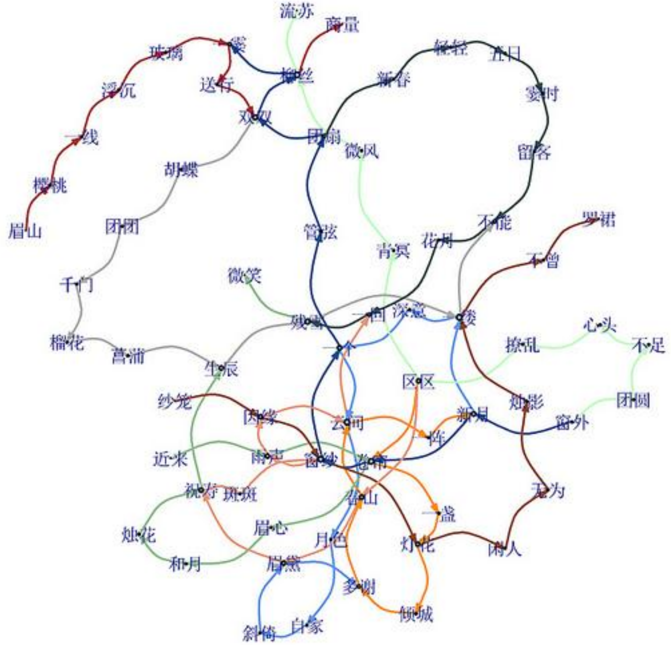
Keyword Association



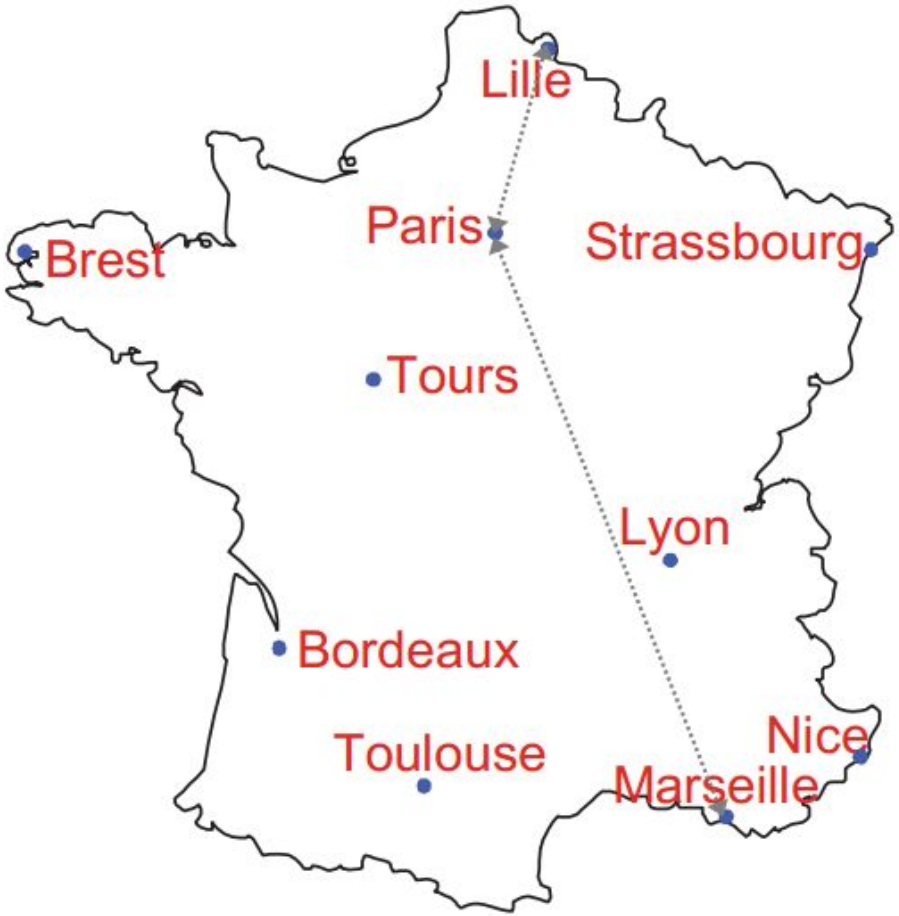
Topic Model



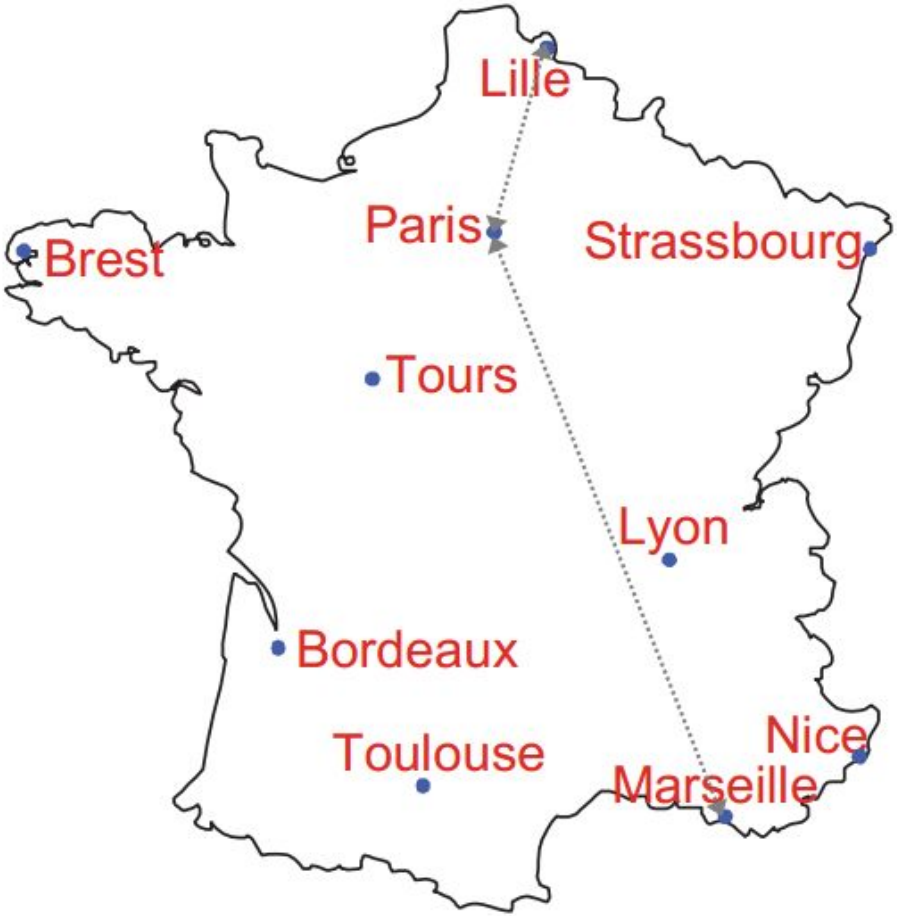
樱桃
灯花
梳妆
风动
罗裙
玻璃
碧玉
送行
横波
柳丝
团扇
留客
柔肠
更深
眼波
烛花
新月
生辰
新春
团圆
残雪
不能
紫薇
雨声
眉尖
一缕
因缘
倾城
不足
诗情
卷帘
南风
秋波
无为
云间
月中
长相
道人
长寿
金花
乌云
榴花
举杯
金门
商量
心头
霎时
悠扬
水晶
流苏
双双
纱笼
船儿
微风
窗外
新月
近来
区区

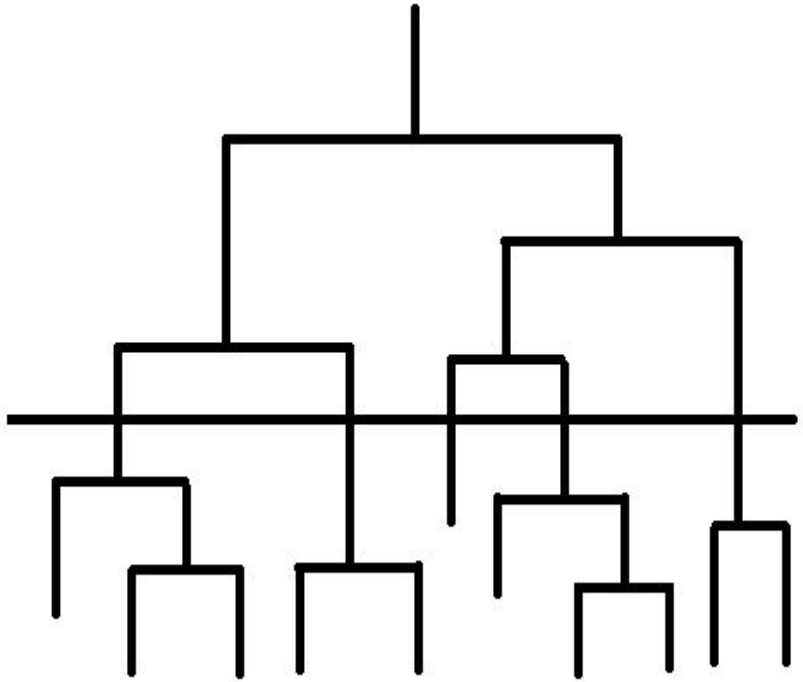


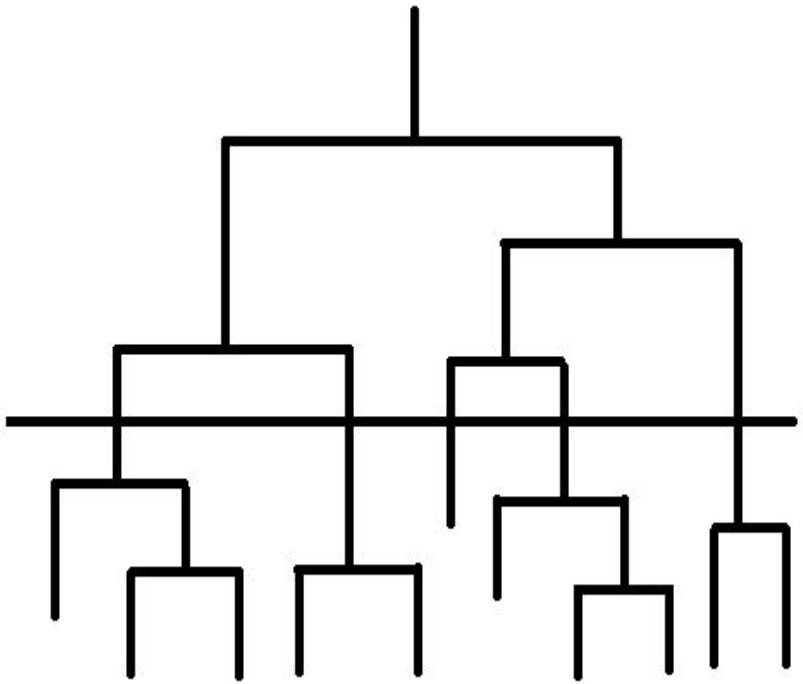
流苏
商量
玻璃
浮沉
送行
双双
团扇
新春
留客
霎时
樱桃
眉山
胡蝶
团团
千门
榴花
微风
管弦
花月
不能
罗裙
不曾
微笑
心头
不足
撩乱
团圆
窗外
烛影
新月
区区
纱笼
近来
生辰
烛花
和月
眉心
月色
灯花
无为
闲人
多谢
倾城
斜倚
白家

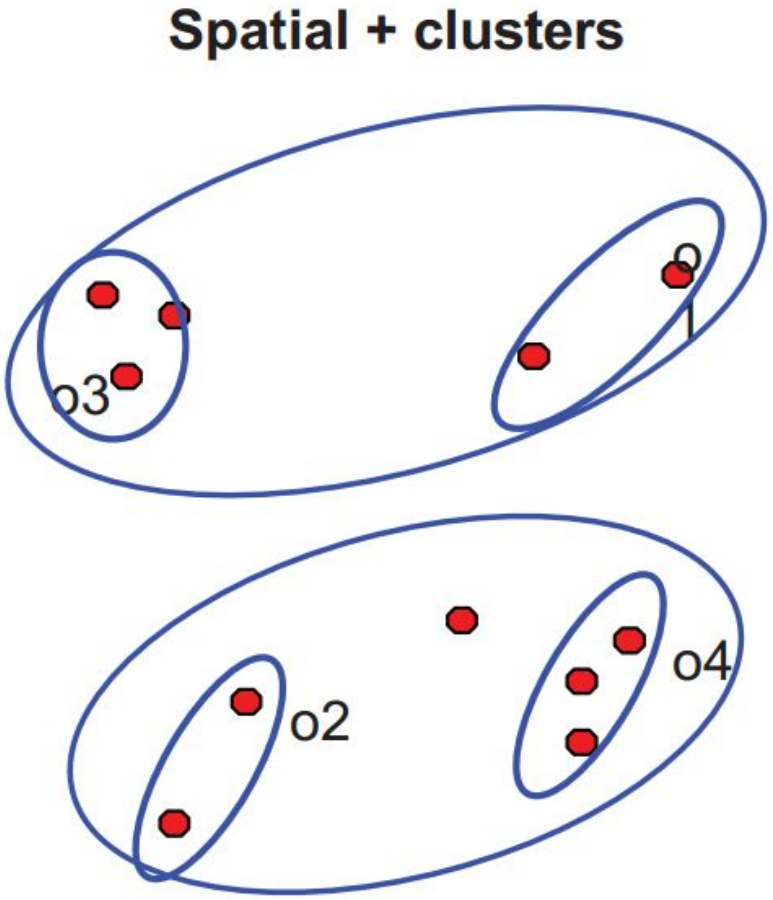

Lille
Paris
Brest
Strassbourg
Tours
Lyon
Bordeaux
Nice
Toulouse
Marseille

	Bor-deaux	Brest	Lille	Lyon	Mar-seille	Nice	Paris	Strassb ourg	Tou-louse	Tours
Bordeaux	0									
Brest	9:58	0								
Lille	6:39	7:11	0							
Lyon	8:05	7:11	4:52	0						
Marseille	5:47	8:49	6:12	1:35	0					
Nice	8:30	13:36	8:20	4:33	2:26	0				
Paris	2:59	4:17	1:04	2:01	3:00	5:52	0			
Strassbourg	8:08	10:16	6:54	4:36	7:04	11:15	4:01	0		
Toulouse	2:02	13:52	9:42	4:25	3:26	6:29	5:14	10:56	0	
Tours	2:36	5:38	4:17	4:21	5:13	9:04	1:13	6:03	6:06	0


Lille
Paris
Brest
Strassbourg
Tours
Lyon
Bordeaux
Toulouse
Nice
Marseille






Spatial + clusters
o3
o
1
o2
o4

最後的小叮嚀

儘量用 sparse matrix 來進行計算

ex. slam, Matrix 套件

進行 NGram 相關運算時，可用平行化套件加速

ex. snowfall 套件

Thank You !!

Taiwan R User Group

MLDM monday

Meetup

http://www.meetup.com/Taiwan-R/

Youtube

https://www.youtube.com/user/TWuseRGroup

高雄 R 社群

Kaohsiung useR! Meetup

https://www.facebook.com/groups/Kaohsiung.R.Users/

參考資料

服貿新聞資料
https://github.com/johnsonhsieh/ecfa

对宋词进行主题分析初探

http://chengjun.github.io/cn/2013/09/topic-modeling-of-song-peom/

用 R 進行中文斷詞
https://www.youtube.com/watch?v=TcMao3r6jYY

用 R 進行中文Text mining
http://rstudio-pubs-static.s3.amazonaws.com/12422_b2b48bb2da7942acaca5ace45bd8c60c.html