Stranity Blog

R 文字探勘－中文斷詞

2016 年 11 月 22 日2017 年 03 月 08 日 • long.jian

文字探勘 (Text mining) 通常用在擷取非結構化資料，我們的生活中存在著許多非結構化的資料，像是新聞、網路論壇 PTT、Mobile01 、社群網站 Facebook、 Twitter 等等文字相關的資料。在以往難以去使用這些資料進行統計分析，而文字探勘則是為了將這些非結構化資料轉換成結構化資料以便進行分析。這篇介紹在進行中文的文字探勘面對斷詞時目前的解決方法，及 R 套件 jiebaR 的推薦。

非結構化信息

結構化信息是可以數字化的數據信息，可以方便地通過電腦和資料庫技術進行管理。無法完全數字化的信息稱為非結構化信息，如文檔文件、圖片、圖紙資料 、縮微膠片等。這些資源中擁有大量的有價值的信息。

斷詞

在進行中文的文字探勘時，與英文有極大不同的地方，英文有空白作為詞與詞的自然斷詞工具，而中文的詞與詞是相連在一起，所以在進行分析前，需要對文章或句子進行所謂的分詞動作，將詞與詞拆開以便進行分析。以下列出在中文斷詞較常使用的方法。

- 字符串匹配分詞法：按照不同的掃描方式，逐個查找詞庫進行分詞。這個方法有個缺點就是當詞庫沒有這個詞的話，則無法精準斷出這個詞。舉個簡單的例子，來說明如何進行字符串匹配法。
  - Eg. 我們在野生動物園玩：
    - 正向最大匹配法：”我們/ 在野/ 生動/ 物/ 園/ 玩”
    - 逆向最大匹配法：”我們/ 在/ 野生動物園/ 玩”
    - 雙向最大匹配法：兩種算法都算一遍，取顆粒最大。
      非字詞典：正向(1) > 逆向(0) (越少越好)。
      單字字詞典：正向(2) = 逆向(2) (越少越好)。
      總詞數：正向(6) > 逆向(4) (越少越好)。
      因此最後輸出為逆向。

- 全切分方法：切分出與詞庫匹配的所有可能詞，再運用統計語言模型決定最優切分結果。與字符串匹配法不同的，這個方法較可能找出新的用語或詞彙。下面列出常用的斷詞統計模型。
  - 最大概率法 (MPSegment)
  - 隱式馬爾科夫模型 (HMMSegment)
  - 混合模型 (MixSegment)
  - 索引模型 (QuerySegment)

R 套件: 斷詞套件 jiebaR (https://github.com/qinwf/jiebaR)

小編習慣使用 R 進行資料的分析，在這裡推薦 jiebaR 進行中文文章的斷詞，這個套件已經建立好各個全切分方法的模型，可以直接使用。另外也可以引用外部的詞庫。在下面示範使用川普的當選宣言使用 jiebaR 進行中文的斷詞，資料來自 Ettoday 東森新聞雲 (http://www.ettoday.net/news/20161109/808308.htm)。

```
install.packages("jiebaR")

library("jiebaR")

#將R環境設定成中文

Sys.setlocale(category = "LC_ALL", locale = "cht")

#在 worker() 內可以設定各種不同的全切分法模型與引用外部詞庫，在這裡直接使用預設的全切分

cc = worker()

text <- "謝謝，謝謝，謝謝大家。很抱歉讓大家久等了，我剛和希拉蕊通話完，
        他非常恭喜我們，我們所有的人，大家都非常努力，這段路非常漫長，
        感謝她為美國人的努力，不管是共和黨民主黨或任何人，現在是時候讓我們團結。
        現在我將成為美國總統，這一刻非常重要。我們可以一起為國家努力，
        為每個愛我們家園的人努力。政府將會盡力照顧所有宗教、生長背景的人，
        這就是我要為我們國家所做的。
        每一個美國人都有機會，都能了解到自己的潛力，讓自己更完美，
        我們將會重建任何不完善的體制。在這18個月我體會到非常多，遇見不少美好的人。
        我們將會重建強健的經濟體制，我們要有夢想，對國家有期望，一起邁向努力，
        沒有任何種族衝突，和平前進。我要感謝我的父母，我從他們身上學到許多。
        感謝我的兄弟姊妹及好友，他們對我的支持；還有我的妻子以及所有支持的我家人，
        我愛你們，這段時間很難熬，政治是噁心的但是感謝妳們的陪伴。
        你們給我很棒的支持，看看我的團隊們，我非常非常的感謝他們，
        下一個4年我還是會繼續站在這跟大家說話，我愛這個國家。"

cc[text]
```

```
  [1] "謝謝"       "謝謝"       "謝謝"       "大家"       "很"         "抱歉"       "讓"
 [11] "我剛"       "和"         "希拉"       "蕊"         "通話"       "完"         "他"
 [21] "我們"       "所有"       "的"         "人"         "大家"       "都"         "非常"
 [31] "非常"       "漫長"       "感謝"       "她"         "為"         "美國"       "人"
 [41] "是"         "共和黨"     "民主黨"     "或"         "任何人"     "現在"       "是"
 [51] "團結"       "現在"       "我"         "將成"       "為"         "美國"       "總統"
 [61] "重要"       "我們"       "可以"       "一起"       "為"         "國家"       "努力"
 [71] "我們"       "家園"       "的"         "人"         "努力"       "政府"       "將會"
 [81] "宗教"       "生長"       "背景"       "的"         "人"         "這"         "就是"
 [91] "國家"       "所"         "做"         "的"         "每"         "一個"       "美國"
[101] "機會"       "都"         "能"         "了解"       "到"         "自己"       "的"
[111] "更"         "完美"       "我們"       "將會"       "重建"       "任何"       "不"
[121] "在"         "這"         "18"         "個"         "月"         "我"         "體會"
[131] "遇見"       "不少"       "美好"       "的"         "人"         "我們"       "將會"
[141] "經濟體制"   "我們"       "要"         "有"         "夢想"       "對"         "國家"
[151] "邁向"       "努力"       "沒有"       "任何"       "種族"       "衝突"       "和平"
[161] "我"         "的"         "父母"       "我"         "從"         "他們"       "身上"
[171] "我"         "的"         "兄弟"       "姊妹"       "及"         "好友"       "他們"
[181] "支持"       "還有"       "我"         "的"         "妻子"       "以及"       "所有"
[191] "人"         "我愛你"     "們"         "這段"       "時間"       "很"         "難熬"
[201] "的"         "但是"       "感謝"       "妳們"       "的"         "陪伴"       "你們"
[211] "支持"       "看看"       "我"         "的"         "團隊"       "們"         "我"
```

```
 [221] "感謝"   "他們"   "下"     "一個"   "4"      "年"     "我"
 [231] "站"     "在"     "這跟"   "大家"   "說話"   "我"     "愛"
```

可以發現因為詞庫並沒有希拉蕊這個詞，所以 jieba 將希拉蕊拆成“希拉"“蕊"，但是在其他詞彙則是效果很不錯的。所以在使用 jieba 適時的增加詞庫也是很重要的。

結論

在進行中文的文字探勘時， jiebaR 提供更精簡且迅速的方式進行分詞，並且在小編的建議下，在開發者版本的 jiebaR 新增對 list 格式的中文斷詞功能，方便開發者可以一次性對多個文本進行斷詞。在這裡特別謝謝 Qin Wenfeng 開發並且開放這麼強大的套件，還有他對小編在文字探勘上的幫忙。

在wordpress.com寫網誌. 佈景主題：lyretail，發表者：wordpress.com。