

從字裡行間中找出數據價值- 文字探勘

David Chiu
2016/07/06

文字探勘的重要性

星星代表的意思？



Books ▾

🔍



NEW & INTERESTING FINDS
FROM ACROSS AMAZON

Departments ▾ Browsing History ▾ David's Amazon.com Today's Deals Gift Cards & Registry Sell Help

Hello, David
Your Account ▾ Try Prime ▾ Lists ▾  Cart

Books Advanced Search New Releases Best Sellers The New York Times Best Sellers Children's Books Textbooks Textbook Rentals Sell Us Your Books Best Books of the Month

Books > Computers & Technology > Databases & Big Data



Machine Learning with R Cookbook

Explore over 110 recipes to analyze data and build predictive models with the simple and easy-to-use R code.

Yu-Wei Chiu (David Chiu)

[Flip to back](#)

Machine Learning With R Cookbook - 110 Recipes for Building Powerful Predictive Models with R

Paperback – March 26, 2015

by Chiu (David Chiu) Yu-Wei ▾ (Author)

★★★★☆ ▾ 10 customer reviews

[See all 2 formats and editions](#)

Kindle \$31.99	Paperback \$39.99
-------------------	----------------------

[Read with Our Free App](#) 9 Used from \$43.64
16 New from \$39.99

Key Features

- Apply R to simplify predictive modeling with short and simple code
- Use machine learning to solve problems ranging from small to big data
- Build a training and testing dataset from the churn dataset, applying different classification methods

Share     <Embed>

Buy New \$39.99

Qty: 1 ▾

In Stock.

Ships from and sold by Amazon.com.
Gift-wrap available.

 Add to Cart

This item ships to **Taipei, Taiwan; Republic of China.** [Learn more](#)

 Pay in TWD with 1-Click

Price in TWD: **1,363.80**
[Change 1-Click payment to USD](#)

Ship to:
David Chiu- Taipei ▾

內部的評論反應使用者的真正感覺

- But the serious problem is that there's **no navigating bar supported for Mac OsX**

Showing 1-1 of 1 reviews (1 star). [Show all reviews](#)

★☆☆☆☆ Decent contents but poor book structure

By [Taz](#) on August 9, 2015

Format: Kindle Edition | **Verified Purchase**

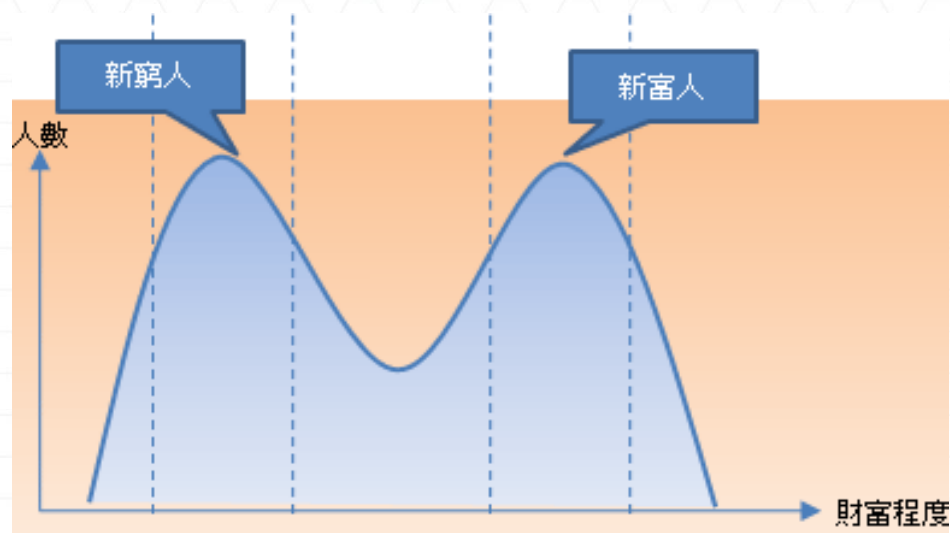
This book has pretty decent contents and well curated examples. But the serious problem is that there's no navigating bar supported for Mac OsX. I have contacted the customer service in many emails but circulating discussion, which made me send the screenshot for the diagnostic several times.

Missing the navigating bar is not tiny one as you need to find information yourself and back to the page in that you don't know where you are reading now. The publisher should fix this issue.

► [2 comments](#) | 2 people found this helpful. Was this review helpful to you? [Report abuse](#)

過往了解使用者的反應

- 讓使用者評分
- 用五星量級評分
- 問卷調查
- 舉辦焦點團體



當抽樣方法有誤，所得到推論也有誤

搜尋引擎找出跟流病的關係

- 當感冒的人多了，使用 Google 查詢「發燒」或「咳嗽」的民眾也跟著變多，讓**特定關鍵字的搜尋熱門度**，成了疫情變化的指標



Google 的預測

- 預測的數值和真實的病情呈現超高度的正相關（相關係數高達 0.85）
- <https://www.google.org/flutrends/about/>

正相關係數（介於1~0之間）	等級
≥ 0.8	超高度相關（excellent correlation）
0.8~0.6	高度相關（good correlation）
0.6~0.4	中度相關（moderate correlation）
< 0.4	低度或無相關（poor correlation）

使用Google Trend

■ <https://www.google.com.tw/trends/>



但你又沒有Google 後面的資料

社群媒體的興起

facebook

電子郵件或電話密碼登入忘記密碼？

Facebook，讓你和親朋好友保持聯繫，隨時分享生活中的每一刻。

註冊

永遠免費！

姓氏名字

手機號碼或電子郵件

重新輸入手機號碼或電子郵件地址

新密碼

生日

年 月 日 為什麼需要提供出生日期的資料？

☐ 女性 ☐ 男性

一旦點擊註冊，即表示你同意使用條款，而且你也閱讀了資料政策，包括 Cookie 的使用。

註冊

看看此刻正在發生什麼事。
尋找與你所喜愛事物相關的社群、對話和靈感。

註冊Log in

Q精選新聞時事名人影視音樂體育更多

大中轉>



台灣蘋果日報 Taiwan News @TW_nextmedia 22 小時
「少轉」 #李潤 抵台賣 #秒收 上百粉絲接機

演藝>



JJ Lin @JJ_Lin 24 小時

你會愛好蔡依林的創作過程？
也好奇他的音樂都從哪裡來？
走進他的音樂，聽見林俊傑
這是一段 JJ 生命中最重要的歷程
也是一部關於音樂與人生的紀錄片
獻給所有音樂人以及熱愛音樂的人
... fb.me/5JHPZmbet

29 293

大中轉>

新加入 Twitter?
立即註冊，取得你的個人化時間軸！
註冊

批踢踢實業坊 · Gossiping

精選區

最新 上頁 下頁 最新

4 [問卦] L4D2 是最經典的殺殭屍遊戲嗎?
7/06 hell3266

1 Re: [新聞] 加國 29.5 小時只給 3 小時加班費 高雄銀行
7/06 putpi2007

Re: [新聞] 談大巨蛋爭議 柯：法律保障應得起律師的人
7/06 unclefucka

Re: [問卦] 人類以外的動物交配會爽嗎?
7/06 Meursault

! [公告] 八卦板板規(2016.02.16)
2/16 seabox

爆 M [爆料] 前市府葛基法爭議整理
6/28 superlighter

56 Fw: [盜錄] 請大家幫我找大伯公橋南新屋鄉埔頂村(代PO)
7/02 kun0616

爆 M [公告] 七月份單次開闢文
7/01 Bignana

本網站已依台灣網站內容分級規定處理。此區域為限制級，未滿十八歲者不得瀏覽。

9

使用輿情分析了解民意



觀測清單

- PTT
- Facebook粉絲團
- 網路新聞媒體

每五分鐘可即時蒐集所有頻道評論



- 蘋果日報、時報資訊、NOW News、聯合新聞網、TVBS、中央通訊社、中廣新聞網、鉅亨網、新頭殼、民報、風傳媒、優活新聞網、健康醫療網、實況新聞網、Match 生活網、自由時報

每日統計報表

[首頁](#) [關鍵字](#) [監控網頁](#) [反向追蹤](#)

設定查詢條件

起始日期

2014-10-01

終止日期

2014-11-15

監控來源

ptt

排序

date

設定關鍵字

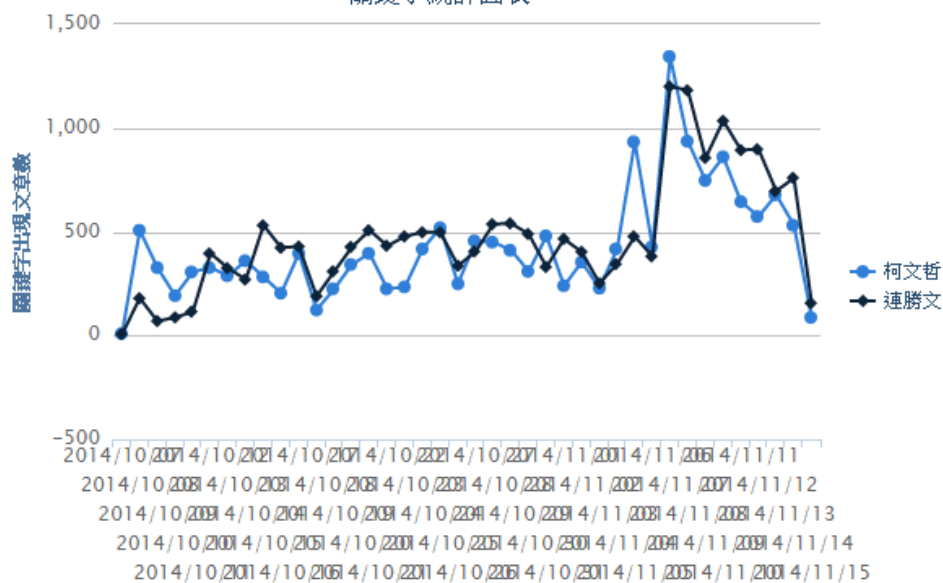
× 連勝文 × 柯文哲 × MG149

※ 台北市長選舉

查詢

統計圖表

關鍵字統計圖表



貝貝絃之音

檢視關聯文章

關聯文章

[討論] 如果神豬被慘電，吱吱還能

HatePolitics

💬 comments:5 📈 likes:4 📅 2014/11/15

Re: [新聞] 支持者若不投票 連：以後遇捷運殺人只能

Gossiping

💬 comments:2 📈 likes:3 📅 2014/11/15

Re: [新聞] 零安樂死流浪狗安置地 連勝文：雲林、嘉

PublicIssue

💬 comments: 📈 likes:1 📅 2014/11/15

[新聞] 連勝文：藉外界抹黑訓練抗壓能耐

HatePolitics

💬 comments:3 📈 likes:1 📅 2014/11/15

Re: [心情] 政治獻金--柯文哲做到了

HatePolitics

💬 comments:3 📈 likes:2 📅 2014/11/15

Re: [問卦] 有沒有連勝文選輸後下一步的八卦

Gossiping

💬 comments:3 📈 likes:12 📅 2014/11/15

看聲量說故事？



引用自壹週刊 2015/09/02 封面故事：
<http://www.nextmag.com.tw/magazine/news/20150902/25253358>

文字資料蒐集

數據分析的步驟

- 使用不同工具與方法，以從資料中找出有意義的結果
- 目的在於從資料中找出規律或固定模式
- 幫助從過去的資料，了解業務未來可能碰到的不確定因素與未來走向



蒐集資料

整理資料

分析資料

資料抽取、轉換、儲存 (Data ETL)



原始資料

Raw Data



ETL腳本

ETL Script



結構化資料

Tidy Data



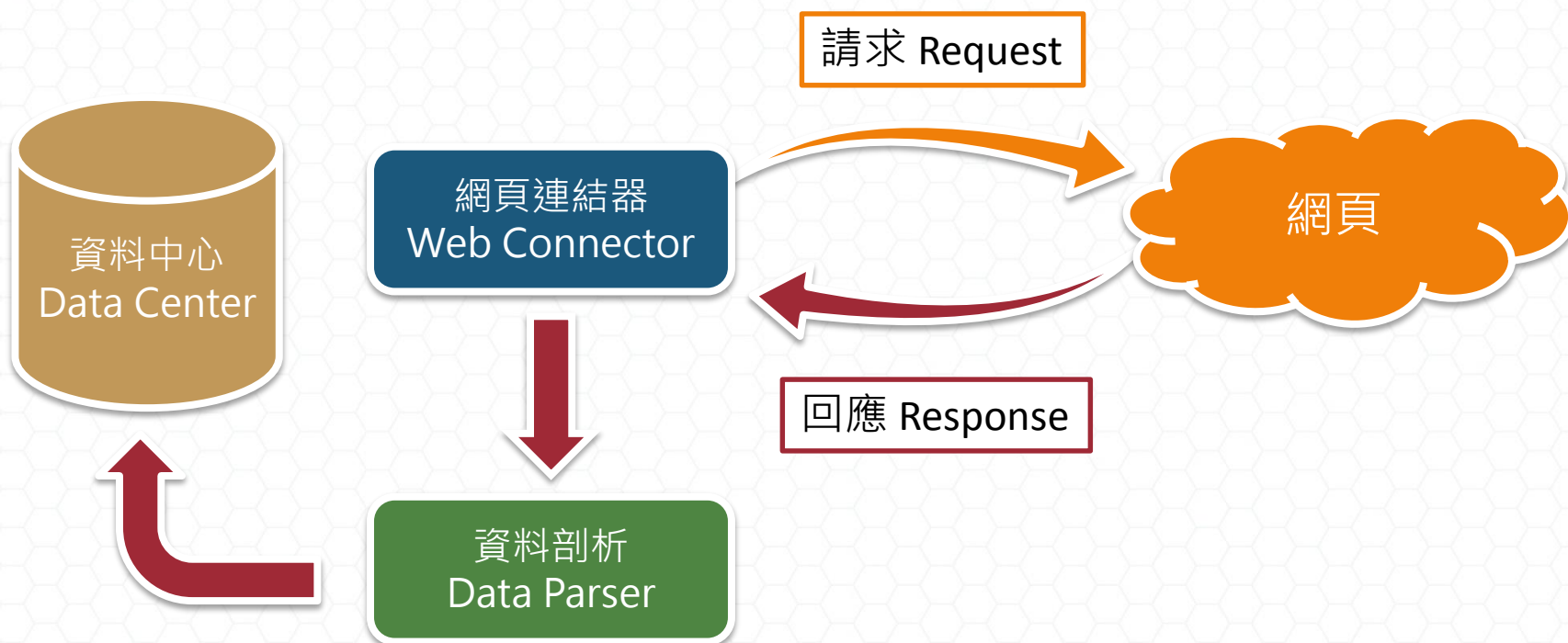
網路爬蟲



將非結構化的網頁資料 轉成結構化資訊

	time	title	category
1	16:55	印度天橋坍塌 已知2死多人被困(0)	國際
2	16:52	異議人士高瑜北京住處 驚傳遭強拆(0)	國際
3	16:50	【更新】50元偽幣流竄 台南警查獲6萬枚(157224)	社會
4	16:50	【特企】一份用愛傳遞的禮物 萬安生命契約 (2911)	特企
5	16:50	【法廣RFI】安倍參加核峰會欲與習近平會...(0)	國際
6	16:48	喬治亞州死刑犯 周四注射毒針(45)	國際
7	16:46	【更新】銷燬贖物卻炸燬焚化爐員工 188...(2428)	社會
8	16:43	挺翁啟惠 學者增批評者：會畫龍的結構式...(166)	生活
9	16:40	【TOMO雙語爆】「德翔台北輪」斷裂漏油...(253)	國際
10	16:40	駁打運將標鐵難除 湘瑩嘆「努力還是不夠」(2171)	娛樂
11	16:39	【有片】身體因素無法返台？ 翁啟惠以色列...(38642)	政治
12	16:35	【10種可樂魅力大！】花錢卻買不到(582)	財經
13	16:33	【法廣RFI】富家女被綁案 港府被指拱手...(330)	國際
14	16:32	檢討新藥發展條例？經長：暫時沒必要(135)	財經

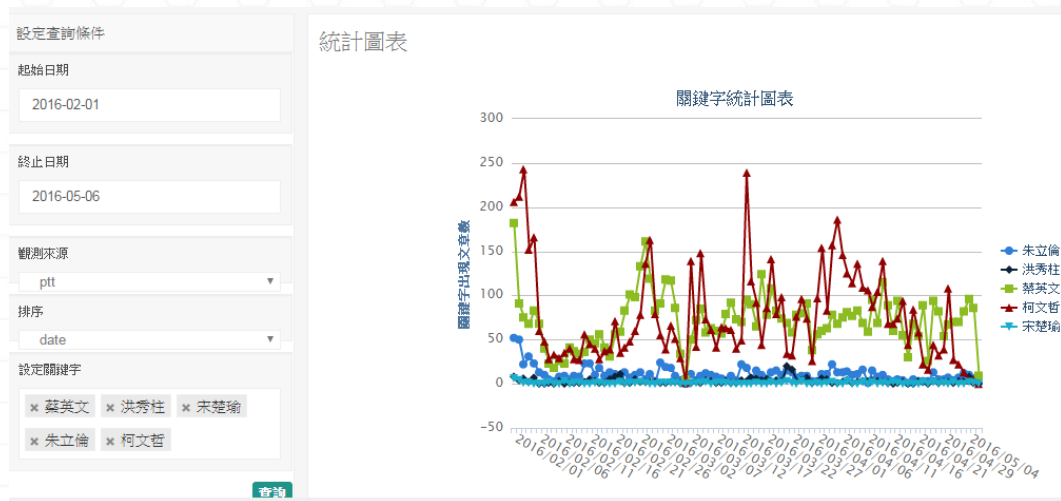
網路爬蟲架構



文字資料處理

文字探勘

- 傳統的資料分析著重於結構化的資料
- 文字探勘的重點在於如何從非結構化的文字中，萃取出有用的重要資訊或知識
- 普遍應用
 - 民意調查
 - 事件追蹤
 - 找出關聯議題
 - 找出文章正負評
 - 文章摘要 (Summly)



文字探勘步驟

文字處理

- 斷詞
- 斷句

資料量化

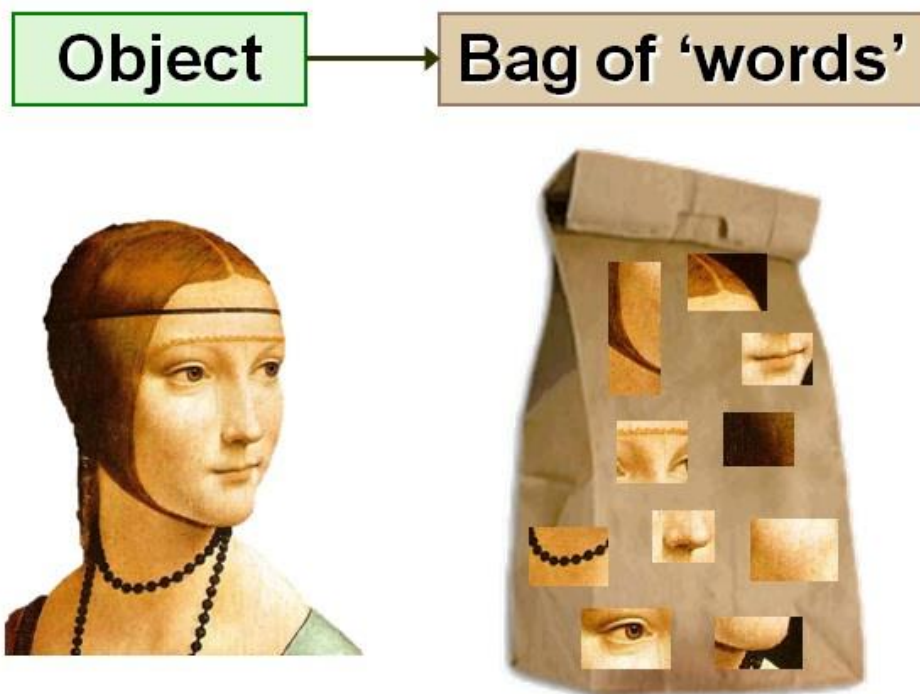
- 詞頻計算
- 文字矩陣
- 計算TF-IDF

探勘分析

- 文字雲
- 文章分群
- 文章分類
- 關聯分析

詞袋模型 (Bag of Words)

- 將文章斷詞以後，可以用向量表示文字。這種表示方式如同將文字變成在袋子中零散且獨立的物件。



中文與英文斷詞

- 英文只要用空白就可以斷詞

'this is a book'

['this', 'is', 'a', 'book']

- 中文要如何斷詞?

酸民婉君也可以報名嗎?

字串比對的斷詞方法

■ 字串比對的斷詞方法

- ❑ 將待分析的漢字串與詞典中的詞進行比對，若在詞典中找到某個字串，則比對成功
- ❑ 按照比對方向的不同，比對斷詞方法可以分為正向和逆向
- ❑ 按照不同長度優先比對的情況，可以分為最長優先和最短優先
- ❑ 按照是否與詞性標注過程相結合，又可以分為單純斷詞方法和斷詞與標注相結合的方法

基於語意的斷詞方法

■ 基於語意的斷詞方法

- 斷詞的同時進行句法、語義分析，利用句法資訊和語義資訊來處理歧義現象
- 模擬人對句子的理解過程，以根據詞、句子等的句法和語義資訊來對斷詞歧義進行判斷

基於統計的斷詞方法

■ 基於統計的斷詞方法

- 在內文(Context)中，相鄰的字同時出現的次數越多，就可能構成單詞
- 根據隱藏馬可夫模型(Hidden Markov Model)所建立的斷詞系統
- 根據條件隨機域 (CRF)所建立的斷詞系統

n-gram 方法屬於基於統計的斷詞方法

如何找出有意義的詞彙？

■ 使用n-gram 做中文斷詞

■ 假設 $n = 2$

- 統計所有 2-gram 的出現次數
- 可表示成機率：出現次數除以總次數。

■ n-gram 的缺點

- 沒有參考中文文法

大巨 2
巨蛋 2

=
大巨蛋 2

酸民婉君也可以報名嗎



酸民 民婉 婉君 君也 也可 可以 以報 報名 名嗎

產生 2-gram 與 3-gram

■ 2-gram

[1] "那 我" "我 們" "們 酸" "酸 民" "民 婉" "婉 君" "君 也" "也 可"
[9] "可 以" "以 酸" "酸 民" "民 嗎"

■ 3-gram

[1] "那 我 們" "我 們 酸" "們 酸 民" "酸 民 婉" "民 婉 君" "婉 君 也"
[7] "君 也 可" "也 可 以" "可 以 酸" "以 酸 民" "酸 民 嗎"

最多切到4-gram 就好

長詞優先法

- 最普遍被廣泛使用的斷詞方法
- 從句子的一端開始，取最長的詞串逐一比對辭典內的詞，若找到就把它當作斷詞的結果，再把句子中比對到的詞去除，剩下的部份再重複剛剛的動作，直到整句都斷詞完畢
- 通常若有夠大的辭典，長詞優先法的正確率可高達90%

長詞優先演算法

- 給定一個連續句子S
- 給定常用詞典D
- 從最長詞 $n = 4$ 到 $n=2$:
 - 從左到右掃描S
 - 檢查S中是否有關鍵詞在D中
 - 如是則移除該關鍵詞
 - 回傳移除關鍵詞的句子s'
 - 用n gram 將s'斷句
 - 將出現超過**最小閾值**的字加到字典D

現有中文斷詞系統 (一)

■ 中研院中文斷詞系統

<http://ckipsvr.iis.sinica.edu.tw/>

- ▣ 準確率最高、速度較慢
- ▣ 細分四十多種詞性，如名詞可細分為地方名詞、普通名詞，專有名詞等。
- ▣ 中文分詞達95%準確，詞性標記達90%準確

現有中文斷詞系統 (二)

■ Stanford Word Segmenter

<http://nlp.stanford.edu/software/segmenter.shtml>

- ▣ 轉成簡體效果較佳
- ▣ 可下載單機版，自己訓練繁體模型
- ▣ 支援多種程式語言：JAVA, Python, Ruby, PHP
- ▣ 詞性有十幾種

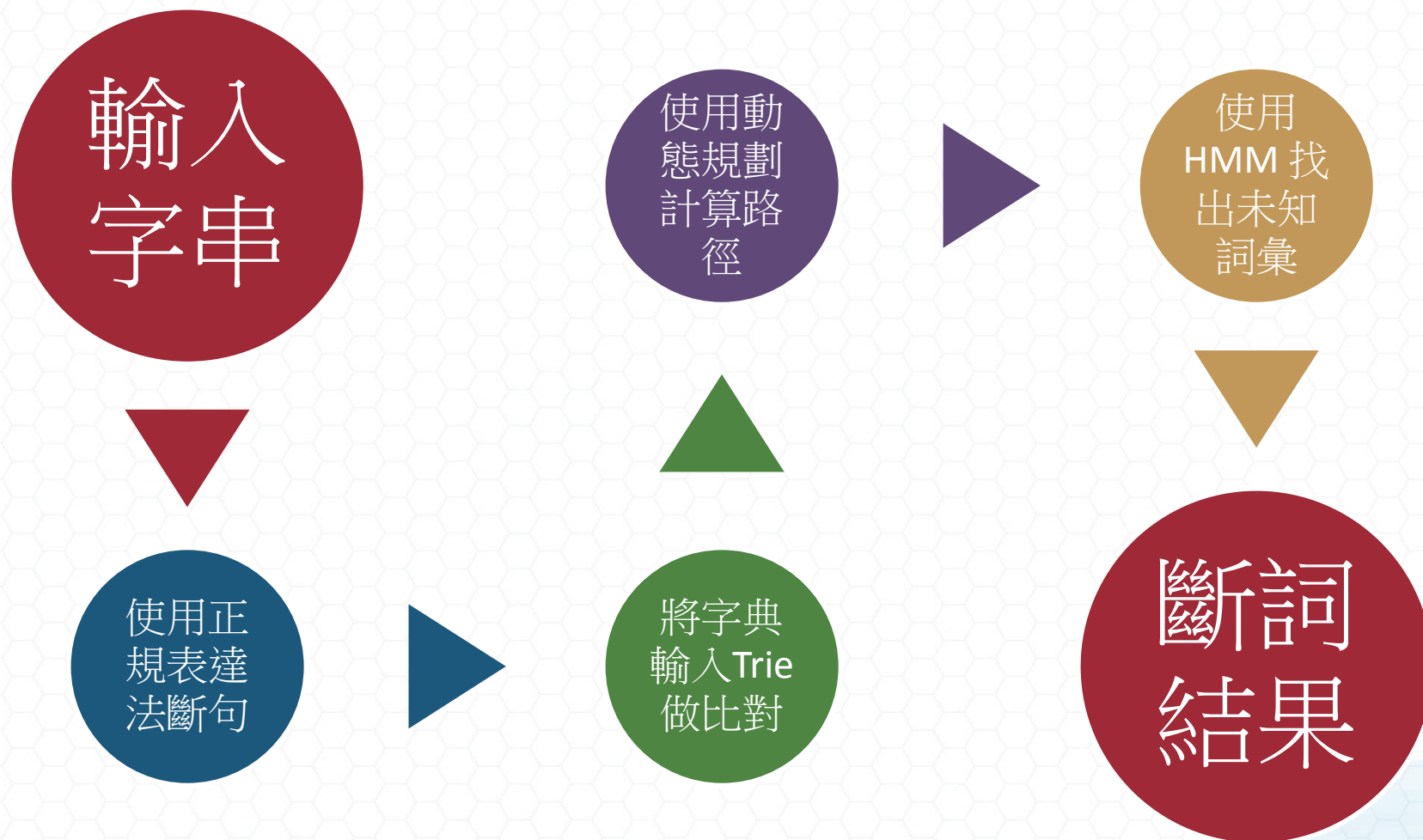
現有中文斷詞系統 (三)

■ Python Jieba中文分詞

<https://pypi.python.org/pypi/jieba/>

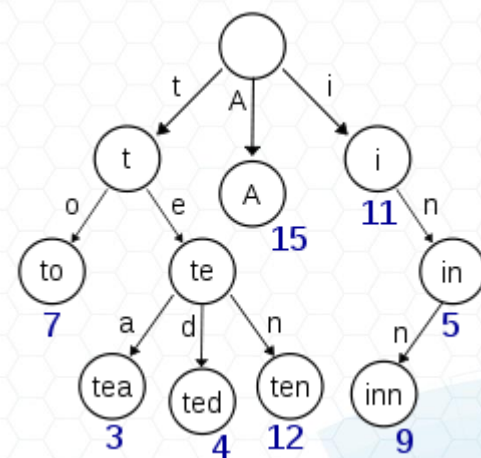
- ▣ 使用Trie 生成句子中字所有可能成詞的情況，然後使用動態規劃依詞頻來找出最大機率的路徑
- ▣ 辨識新詞使用 HMM 模型 (Hidden Markov Model) 及 Viterbi 算法來辨識出來

Jieba 演算法



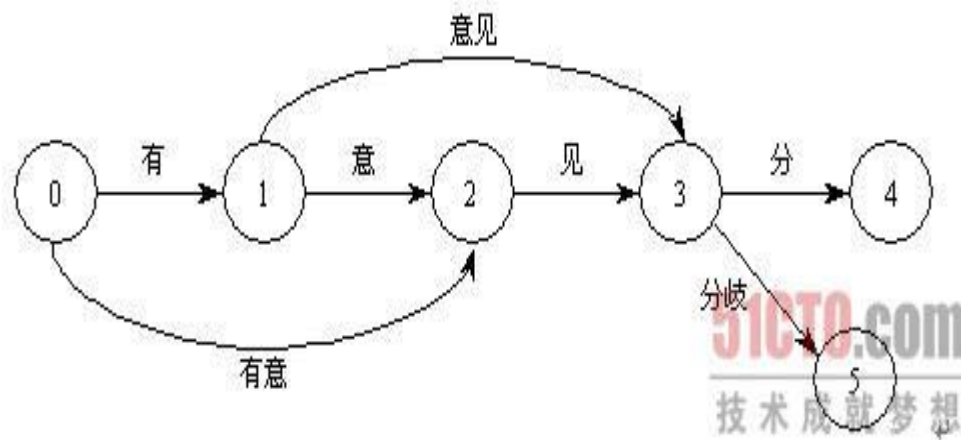
將字典輸入Trie 做比對

- Jieba 內建dict.txt的詞典, 裡面有2萬多條詞, 包含了詞條出現的次數和詞性
- 把這2萬多條詞語, 放到一個trie樹中, 而trie樹是有名的首碼樹, 也就是說一個詞語的前面幾個字一樣, 就表示他們具有相同的首碼, 就可以使用trie樹來存儲, 具有查找速度快的優勢



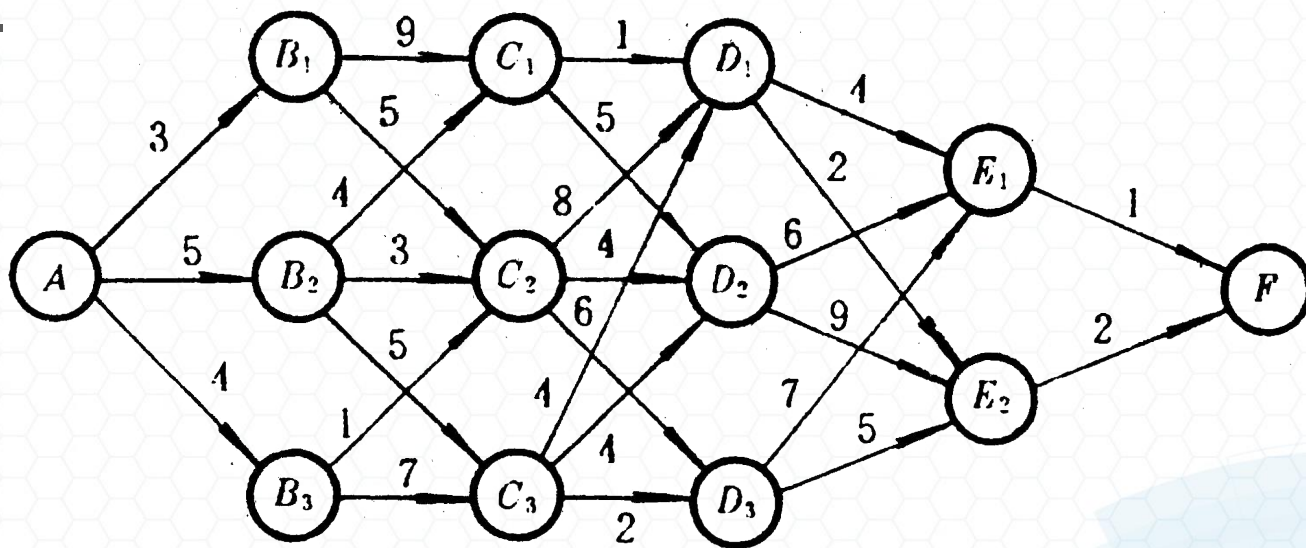
生成DAG

- 根據dict.txt生成trie樹
- 根據trie樹, 生成DAG根據比對到的字樣, 產生幾種可能的端詞方法



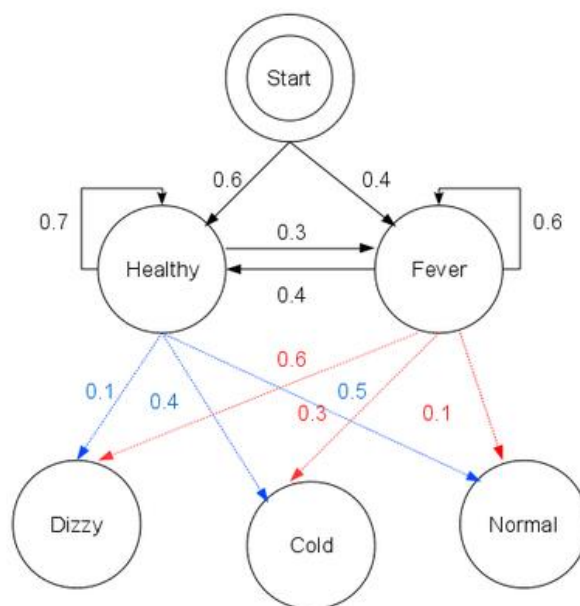
使用動態規劃計算路徑

- 動態規劃中, 先查找待句子中已經切分好的詞語, 對該詞語查找該詞語出現的頻率(次數/總數)
- 然後根據動態規劃查找最大機率路徑的方法, 對句子從右往左反向計算最大機率得到最大機率路徑的組合.



隱馬爾可夫模型

- 用來描述一個含有隱含未知參數的馬爾可夫過程
- 目的是從可觀察的參數中確定該過程的隱含參數。
。然後利用這些參數來作斷詞



誰是馬可夫？

- Andrey Markov

(14 June 1856 N.S. – 20 July 1922)

- Calculated letter sequences of the Russian language



問題描述

■ States -> "F", "L"

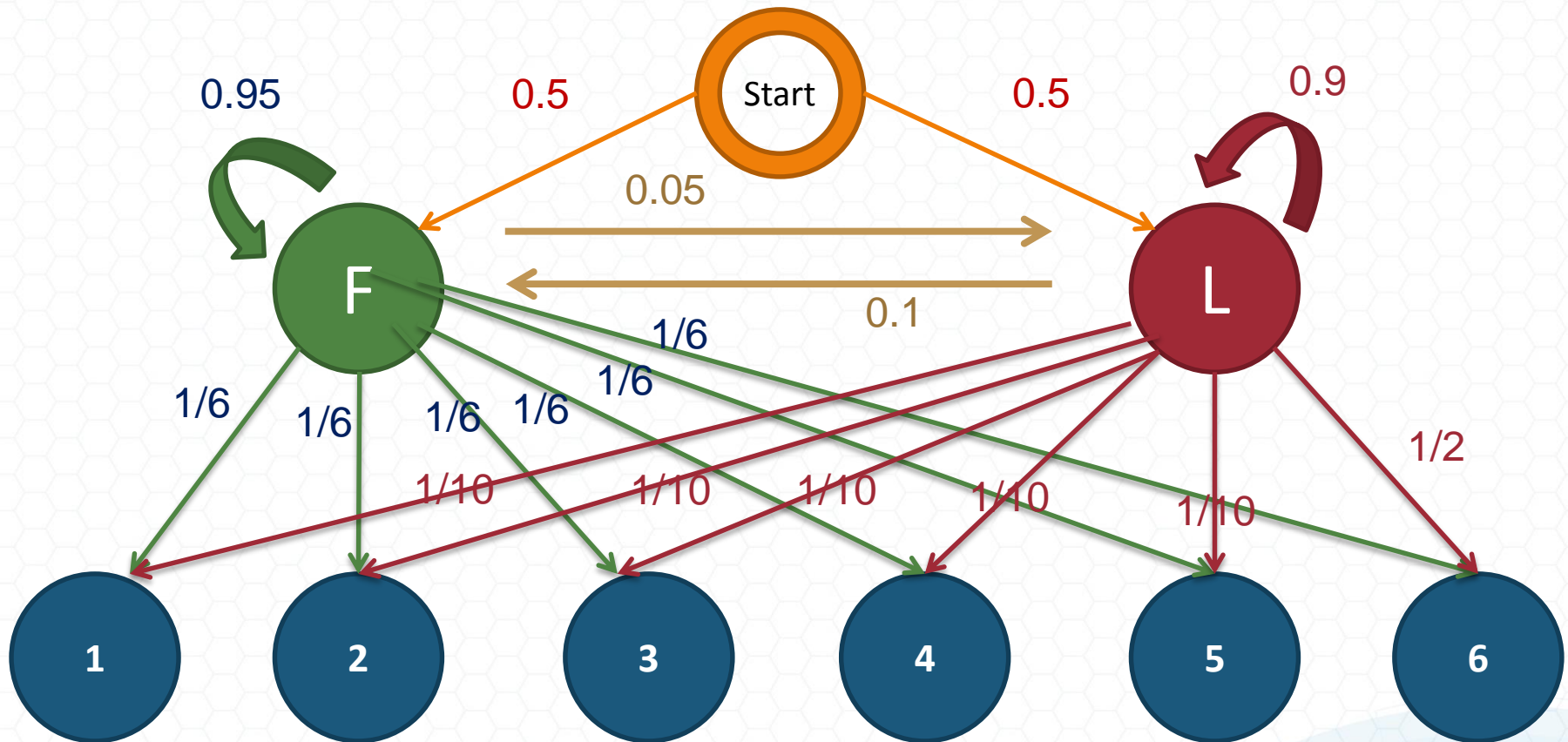
■ Transition Matrix

	Fair	Loaded
Fair	0.95	0.05
Loaded	0.1	0.9

■ Emission Matrix

	Fair	Loaded
1	1/6	1/10
2	1/6	1/10
3	1/6	1/10
4	1/6	1/10
5	1/6	1/10
6	1/6	1/2

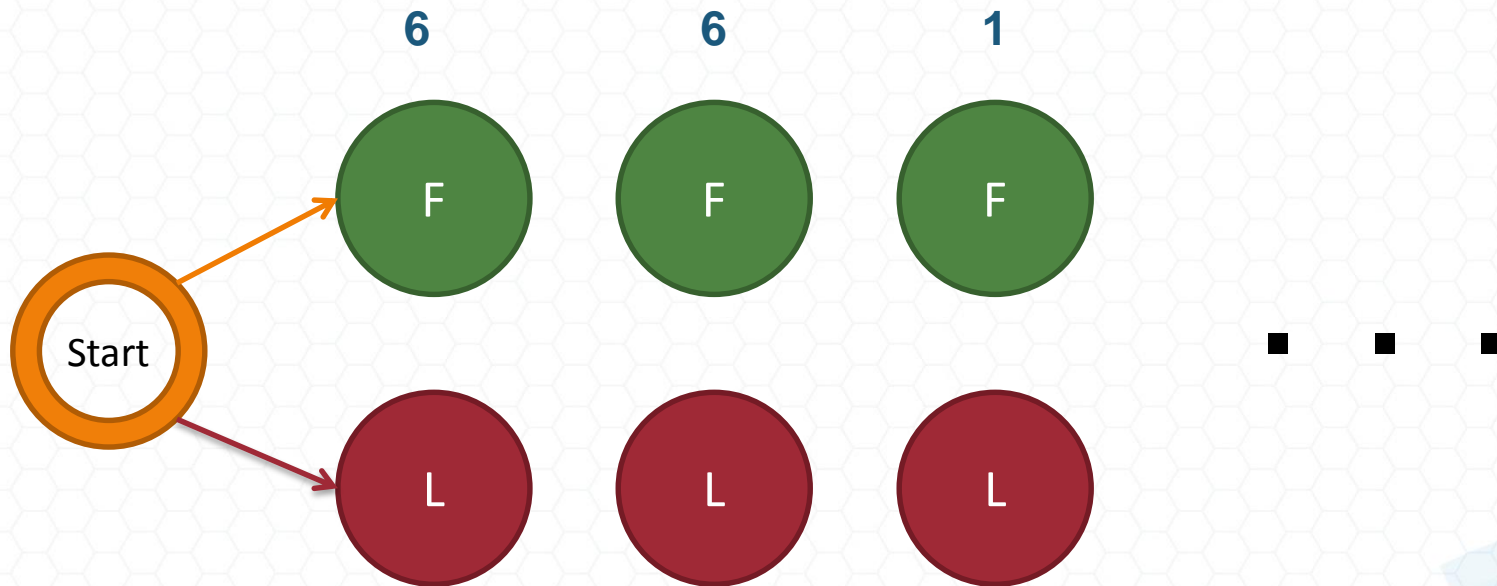
Problem Description



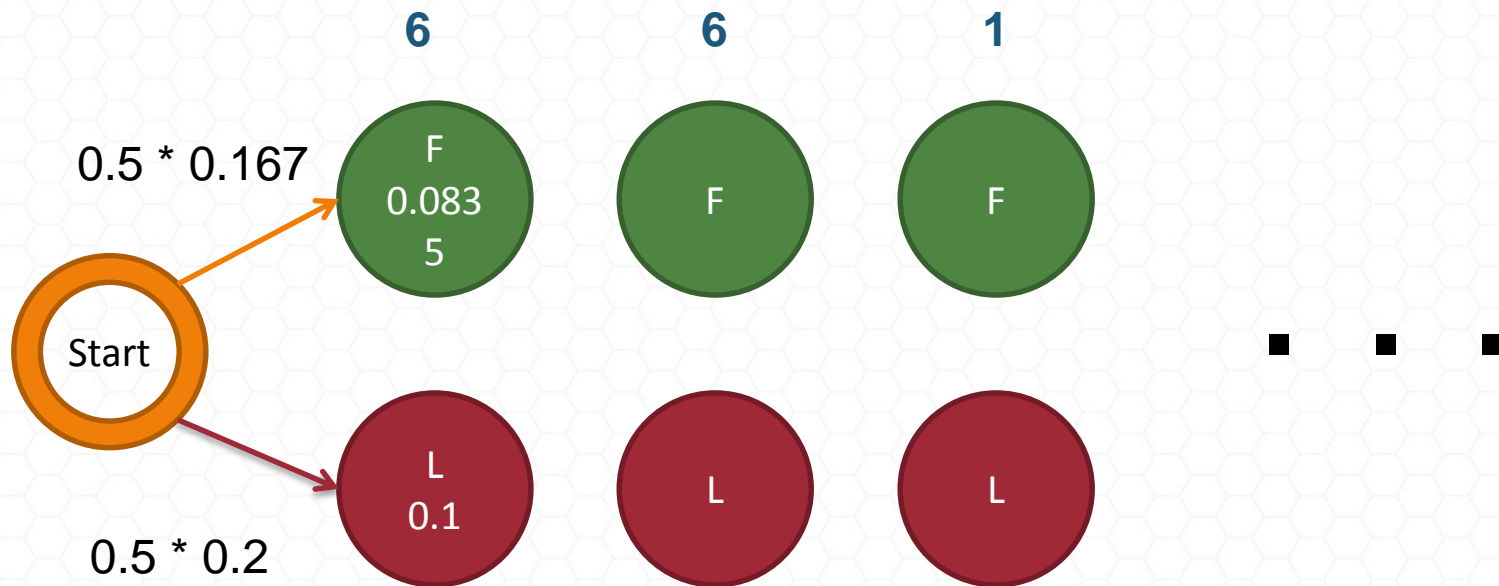
Algorithm

■ Given Observable Sequence:

6,6,4,1,5,3,2...

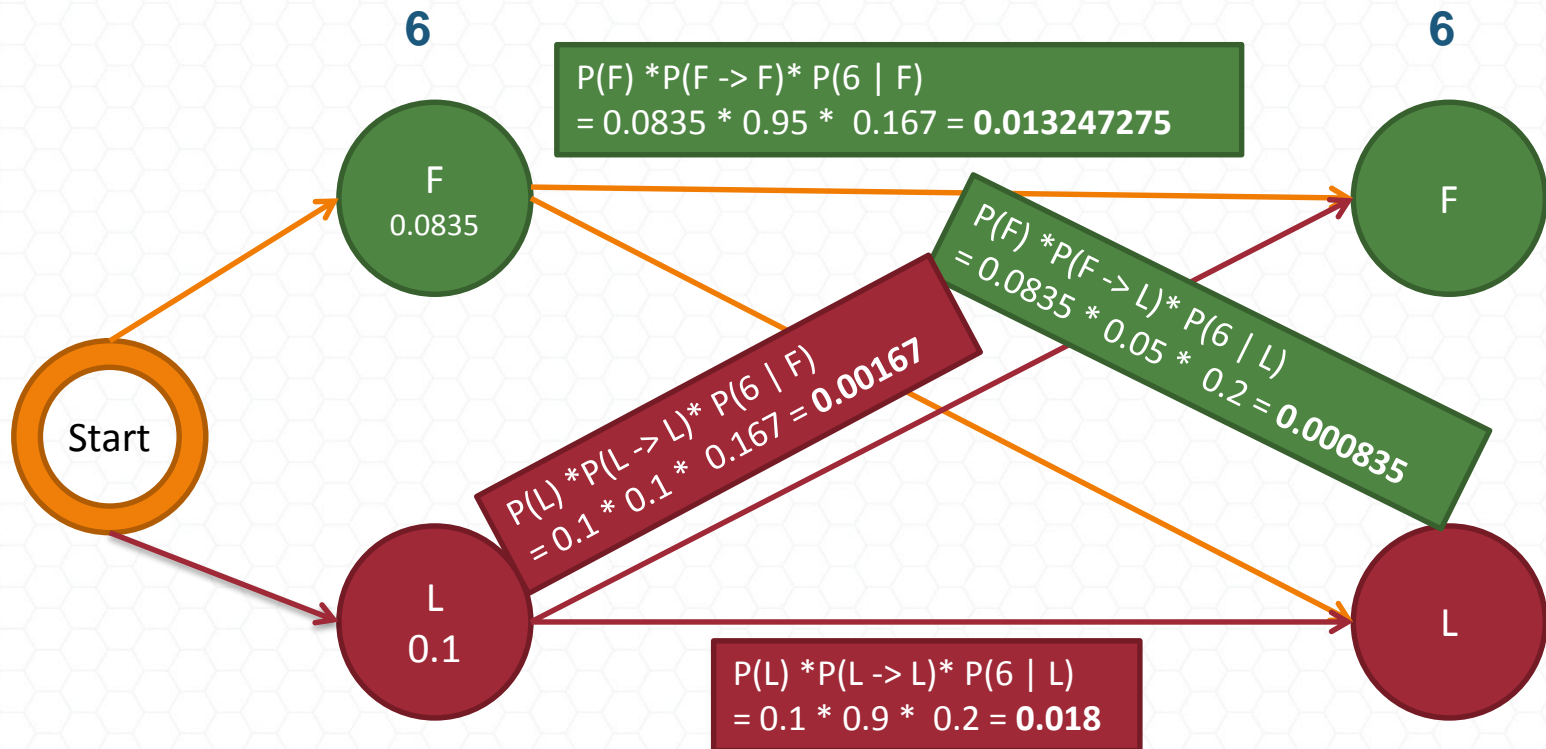


Start to Step 1



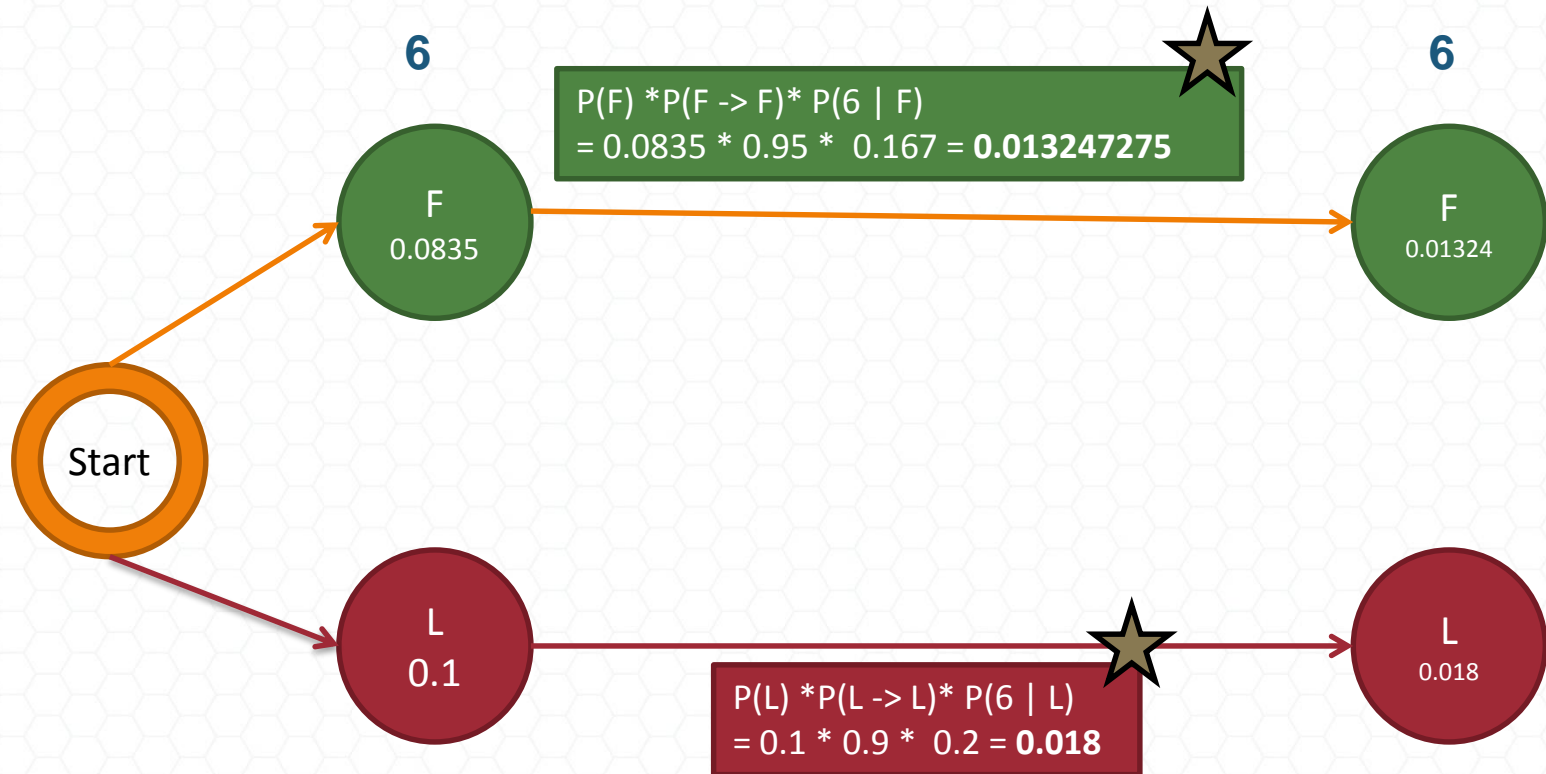
$$P_{\text{Start}}(\text{State}) * P_{\text{Observe}}(6)$$

Step 1 to Step 2

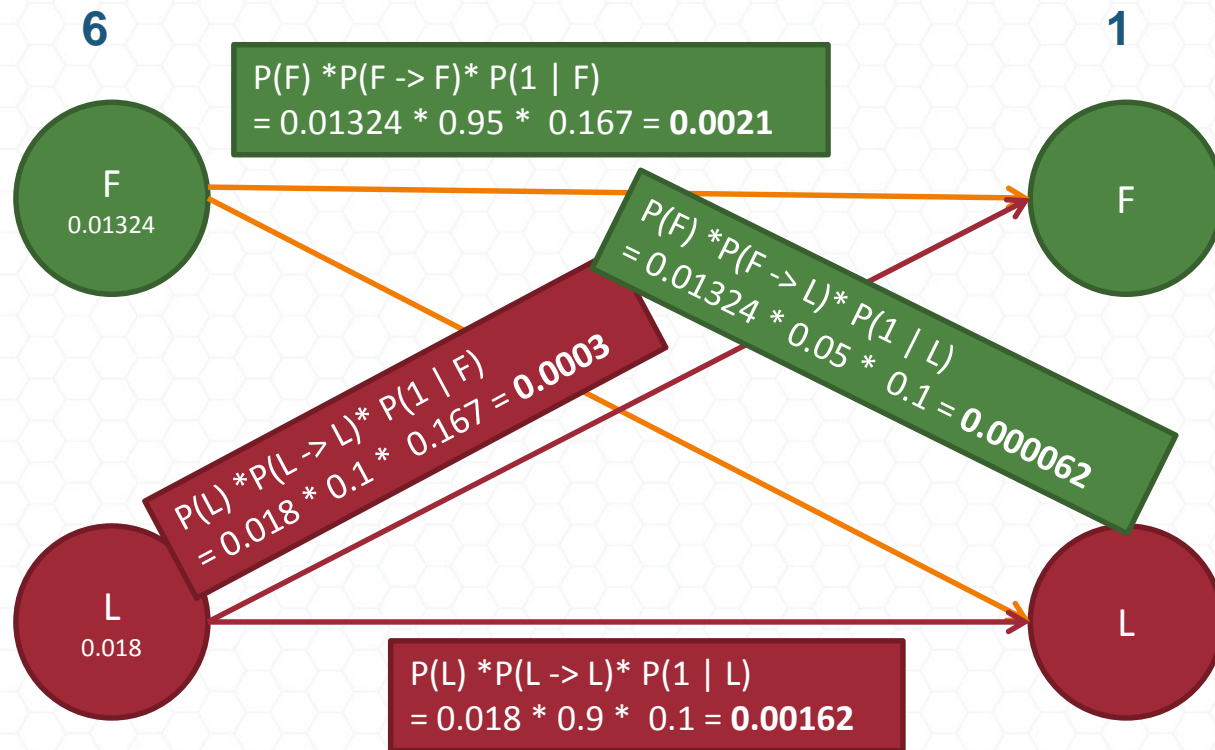


$P_OldState(State) * P_Trans(Old_State \rightarrow New_State) * P_Observe(6 | New_State)$

Most Likely Path

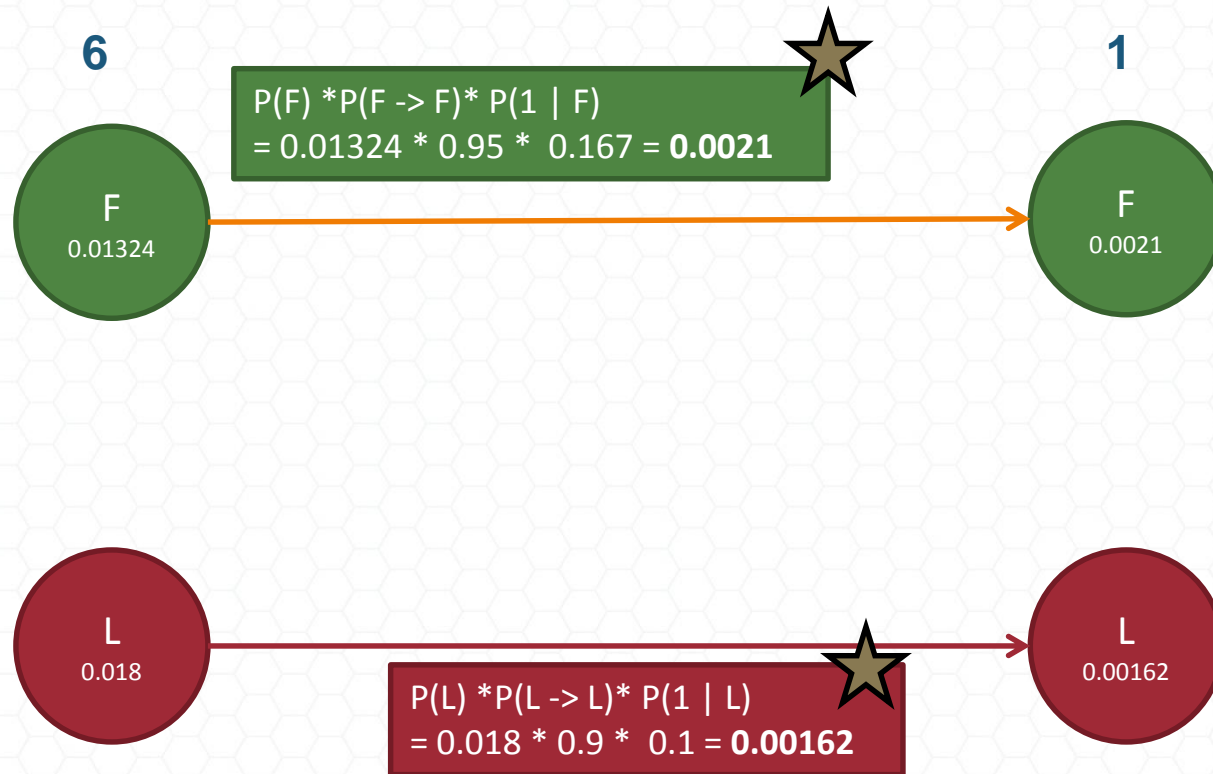


Step 2 to Step 3



$P_OldState(State) * P_Trans(Old_State \rightarrow New_State) * P_Observe(1 | New_State)$

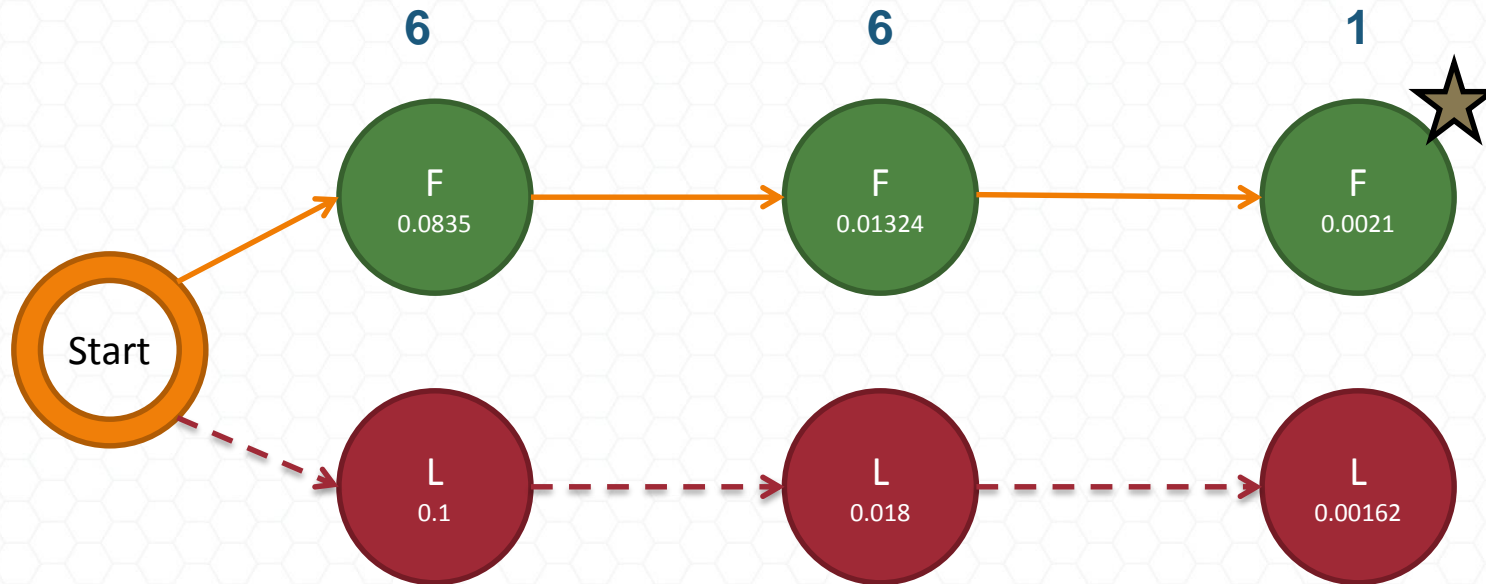
Most Likely Path



$P_OldState(State) * P_Trans(Old_State \rightarrow New_State) * P_Observe(1 | New_State)$

Path Construction

Max(Probability) = 0.021



State Sequence: Fair, Fair, Fair

切詞範例

■ 柯文哲相關資訊與新聞

- 輸出的狀態序列為

- BMEBEBESBE

■ 可以切詞為

- BME/BE/BE/S/BE

- 柯文哲/相關/資訊/與/新聞

■ B後面只可能接(M or E)，不可能接(B or S)，而 M後面也只可能接(M or E)，不可能接(B, S)

機率矩陣

■ 初始機率 InitStatus

#B -0.26268660809250016

#E -3.14e+100

#M -3.14e+100

#S -1.4652633398537678

趨近於0

■ 轉移矩陣機率 TransProbMatrix

B

E

M

S

B
E
M
S

-3.14e+100	-0.510825623765990	-0.916290731874155	-3.14e+100
-0.5897149736854513	-3.14e+100	-3.14e+100	-0.8085250474669937
-3.14e+100	-0.33344856811948514	-1.2603623820268226	-3.14e+100
-0.7211965654669841	-3.14e+100	-3.14e+100	-0.6658631448798212

EmitProbMatrix 矩陣

■ $P(\text{Observed}[i], \text{Status}[j]) = P(\text{Status}[j]) * P(\text{Observed}[i]|\text{Status}[j])$

#B	柯 PB1, 文PB2 哲PB3
#E	柯 PE1, 文PE2 哲PE3
#M	柯 PM1, 文PM2 哲PM3
#S	柯 PS1, 文PS2 哲PS3

從既有詞組發現
每個單字出現的
機率

求出可能路徑

- EBSEBEBEMB

- 倒回來變成

 - BMEBEBESBE

- 可以切詞為 BME/BE/BE/S/BE

 - 柯文哲/相關/資訊/與/新聞

文字分析

文字雲

- 針對使用者輸入文章，分析其文字詞出現的頻度值，以「詞彙地圖」的方式展示
 - ▣ 可使用n-gram 產生詞頻
 - ▣ 或使用jieba 關鍵詞 (有去掉標點符號)



找出文章關鍵詞

■ 如何判斷一個詞是不是關鍵詞

- 如果某個詞比較少見，但是在這篇文章中多次出現，那麼該詞很反映了這篇文章的特性
- 可用來評估該詞對於該文件的重要程度
- 使用 $TF * IDF$ ，假設單詞對文章的重要性越高， $TF-IDF$ 值就越大

TF-IDF

■ TF (Term Frequency)

- 單詞在該文件的出現次數
- 單詞 w 在文檔 d 中出現的次數: $\text{count}(w, d)$
- 文檔 d 中總詞數: $\text{size}(d)$
- $\text{tf}(w, d) = \text{count}(w, d) / \text{size}(d)$

■ IDF (Inverse Document Frequency)

- 一個詞語普遍重要性的度量
- 設文檔總數為 n
- 設詞 w 所出現檔數 $\text{docs}(w, D)$
- $\text{idf} = \log(n / \text{docs}(w, D))$

詞頻矩陣 (document-term matrix)

s = "大巨蛋案對市府同仁下封口令？柯P否認"

s1 = "柯P市府近來飽受大巨蛋爭議"

單詞



	"大巨蛋"	"案"	"對"	"市府"	"同仁"	"下"	"封口令"	"柯P"	"否認"	"近來"	"飽受"	"爭議"
[s]	1	1	1	1	1	1	1	1	1	0	0	0
[s1]	1	0	0	1	0	0	0	1	0	0	1	1

文章

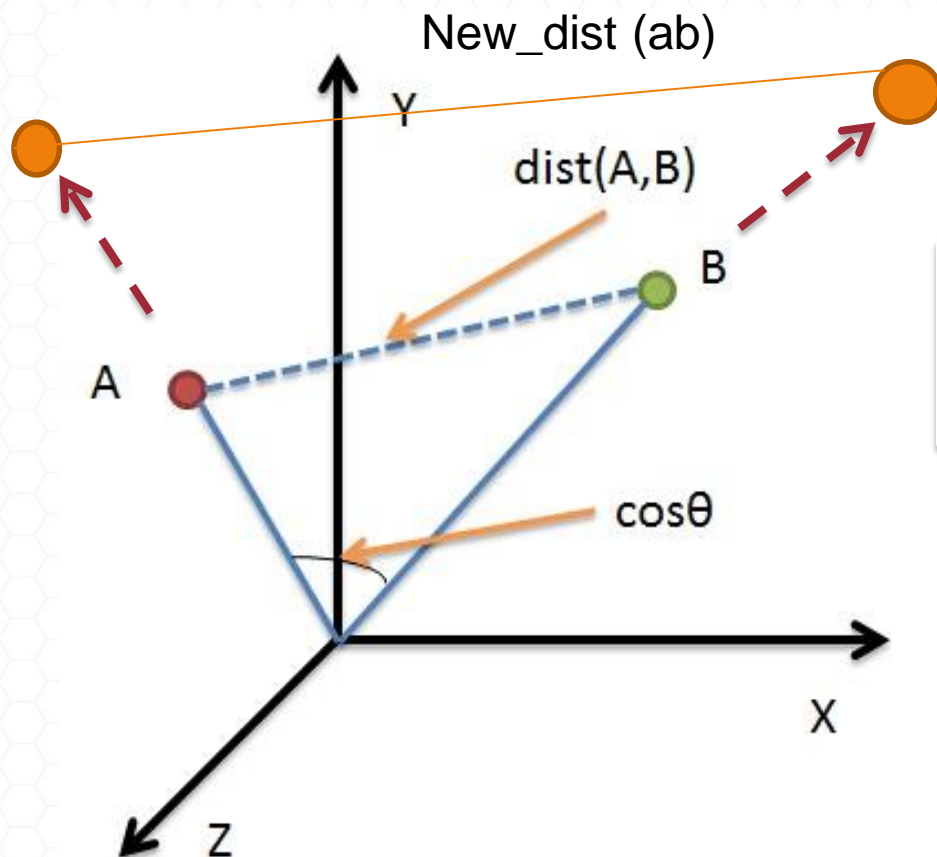


詞頻矩陣

文章相關的相似度

1. 使用TF-IDF演算法，找出兩篇文章的關鍵字
2. 每篇文章各取出若干個關鍵字（比如20個），合併成一個集合，計算每篇文章對於這個集合中的詞的詞頻
3. 生成兩篇文章各自的詞頻向量
4. 計算兩個向量的余弦相似度，值越大就表示越相似。

Euclidean Distance v.s. Cosine Distance



計算相對向量距離
而非絕對距離

實際範例

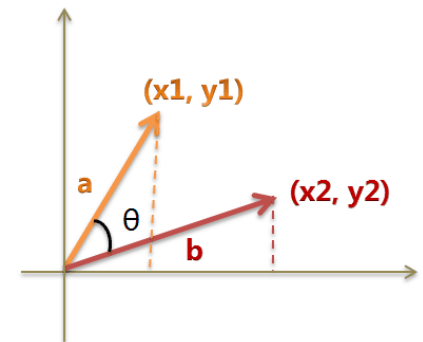
句子A：我 1，喜歡 2，看 2，電視 1，電影 1，不 1，也 0。

句子B：我 1，喜歡 2，看 2，電視 1，電影 1，不 2，也 1。

句子A：[1, 2, 2, 1, 1, 1, 0]

句子B：[1, 2, 2, 1, 1, 2, 1]

$$\begin{aligned}\cos\theta &= \frac{1 \times 1 + 2 \times 2 + 2 \times 2 + 1 \times 1 + 1 \times 1 + 1 \times 2 + 0 \times 1}{\sqrt{1^2 + 2^2 + 2^2 + 1^2 + 1^2 + 1^2 + 0^2} \times \sqrt{1^2 + 2^2 + 2^2 + 1^2 + 1^2 + 2^2 + 1^2}} \\ &= \frac{13}{\sqrt{12} \times \sqrt{16}} \\ &= 0.938\end{aligned}$$



■ 找出文章主題



新聞|擋不住的風暴 川普主宰美總統選情
問卦|川普贏了美國人會崩潰嗎?
Re: 討論|川普 大危機!
新聞|川普又贏 美國人恐傷出走潮
Re: 討論|川普 大危機!
Re: 討論|川普 大危機!
新聞|共和黨「反川普」陣線成立 羅姆尼親上火
新聞|擔心川普式外交 60共和黨大老發公開信
新聞|共和黨前總統候選人羅姆尼稱川普是騙子
新聞|川普致勝秘訣 把選舉當交易
Re: 新聞|擔心川普式外交 60共和黨大老發公開信
討論|川普當選美國總統對台灣未必是壞事
Re: 討論|川普當選美國總統對台灣未必是壞事
Re: 討論|川普 大危機!
Re: 討論|川普當選美國總統對台灣未必是壞事

[新聞】未完待續 三姊弟布丁風暴尚未完結](#)
[新聞】三姊弟布丁要關了！ 臉書PO告別文：從今](#)

人怎麼分類新聞？



【狗仔偷拍】
陳柏霖「分
手」宋智孝
事實竟是...

【動新聞】
接見外賓唸
稿卡住 蔡
英文：稿子
借你唸唸看

下列三篇新聞該怎麼分類？

- 鴻海收購夏普正式簽約郭台銘：全球高科技產業最棒的一天
- 嘉玲採果郭台銘美人柑到手
- 憶起「馬習會」 郭台銘爆氣飆罵：Stupid！

使用Naïve Bayes 分類器

■ Bayes 分類器源自Bayes 理論



Drew Barrymore



Drew Carey

What is the probability of being called
“*drew*” given that you are a **male**?

What is the probability
of being a **male**?

$$p(\text{male} | \text{drew}) = \frac{p(\text{drew} | \text{male}) p(\text{male})}{p(\text{drew})}$$

What is the probability of
being named “*drew*”?

(actually irrelevant, since it is

Naïve Bayes 分類器

- 假設每個特徵(Feature)都為獨立

$$p(d|c_j) = p(d_1|c_j) * p(d_2|c_j) * \dots * p(d_n|c_j)$$

根據Feature d 所產生類別c 的機率

在文章中每個詞可以視為獨立
因此使用貝氏分類法即可以做文章分類

文字分析應用案例

用文字探勘歸納事件

- <http://vtwvtv.net/forum.php?mod=viewthread&tid=78600&tpage=7>



用文字探勘濃縮文章

■ Summly

□ <http://summly.com/>



After launching in November, the Summly app has had more than

500,000

users in 4 weeks

100,000

news articles are summarised by Summly every day



Sources:
Metro/Future Foundation,
Summly

Of British adults who live or work in a city:



want a service which hosts their favourite news in one place



are interested in a service which monitors the news content read by people they know



check their favourite social networking site several times a day...



... of which
do so on their smartphone

今日頭條

■ <http://toutiao.com/>

下载APP

头条号

图虫

反馈

更多

今日头条

大家都在搜：张侧朱颖模恋情



登录

推荐

热点

视频 HOT

图片 NEW

社会

娱乐

科技

汽车

体育

财经

搞笑



火葬场捡了个手机，看到里面的录像时，胆寒不已

老周家住在廿四里殡仪馆附近，这天路过殡仪馆门口的时候，看到地上有一个手机，捡起来后，发现还是个品牌新机，一看就是个高级货，于是老周笑纳了。

老苏 · 3860评论 · 刚刚



在马桶上面贴上保鲜膜，绝对有你意想不到的，回家试...

在马桶上面贴保鲜膜分三个步骤，接下来我看看都是怎样贴的，又会有什么意想不到的事情发生呢？是不是亲生的啊这难道是一种新的面试方法？

表情系列专题 · 399评论 · 10分钟前



我53岁娶了26岁的她，一个月后我就被折腾地离婚

【文图无关】我不够否认，在我和前妻离婚之后，我不服老，想要找一个岁数小一点的妻子。这是每个男人都有的虚荣心，说出来没有什么好丢脸的。然而人有时候真的不...
生活感悟日记 · 6017评论 · 20分钟前



直播：武汉大水围城启动排渍红色预警 待产孕妇水中被困

湖北举水河流域1日晚发生特大洪水，漢口附近200居民已安全转移，还有1.2万人正在转移。

热 我在现场 · 30分钟前



贵的牙膏真的比便宜的牙膏好吗？

（作者：徐明磊，国家一级公共营养师，原创作品，转载请注明出自知识就是力量微信

24小时热闻



今年洪涝灾害会超过1998年吗？

2 伊恩·拉什：葡萄牙的弱点是防守

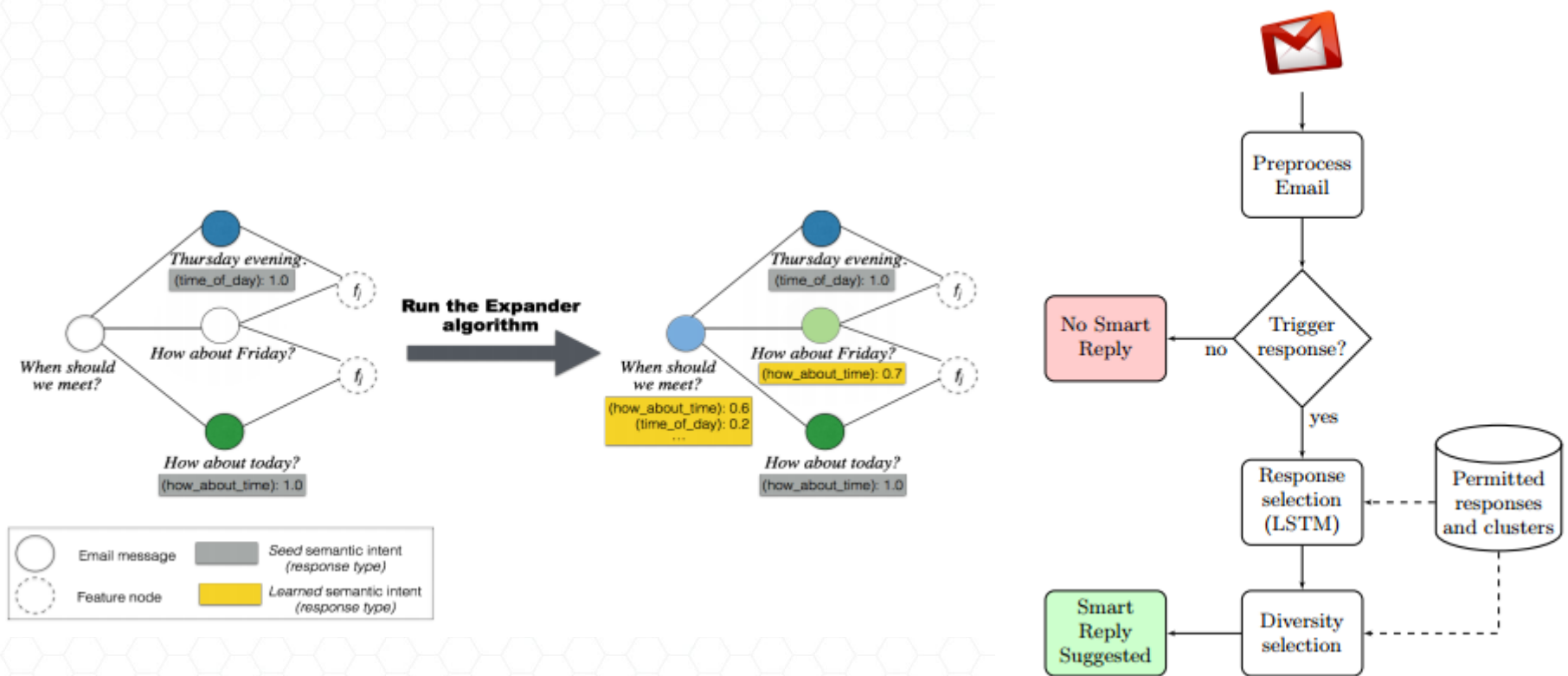
3 贝尔C罗半决赛聚首 皇马两大将基情四射情同手足

4 戴秉国：美国10个航母战斗群都开进南海也吓

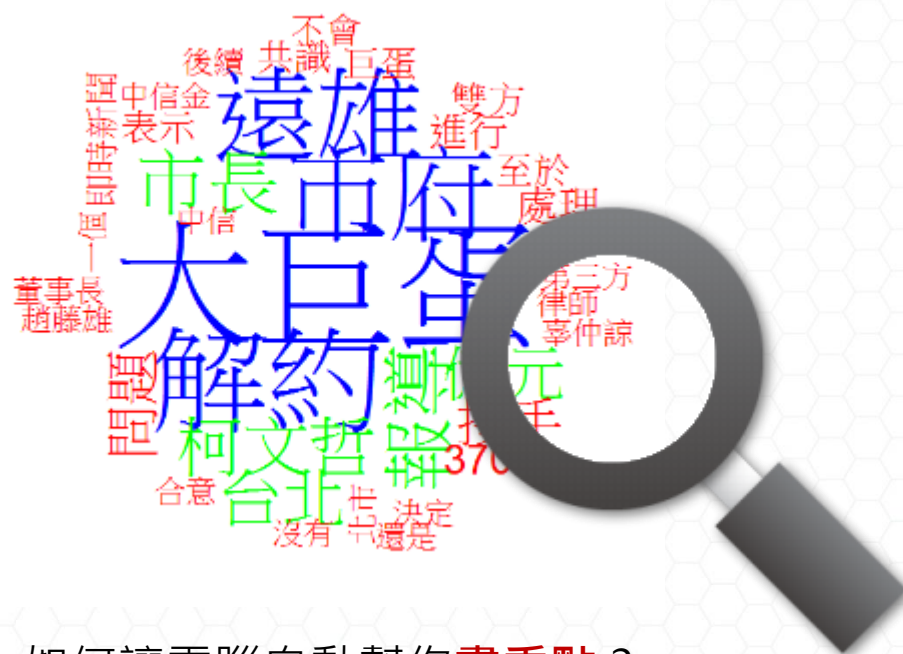


Smart Reply: Automated Response Suggestion for Email

■ 運用深度學習自動回信



文章內的重點到底是什麼？



如何讓電腦自動幫你畫重點？

探勘素材



	Comment
1	內鍋很難清洗
2	電鍋打開鍋蓋時瞬間蒸氣很燙，每次都掙扎許久。且不太確...
3	鍋內髒汙難清洗、水量不好測
4	每次要拿電鍋的東西時，如果沒有ㄇ型夾，常常被燙到，有...
5	煮東西時電鍋外常常都有水灑出來，造成地上都濕濕的，很讓...
6	電鍋好難洗
7	連接電鍋處的插頭總是易脫線、斷掉，恐有走火疑慮
8	電鍋用久後，產生的髒汙很怕洗了會讓電鍋壞掉。另外，...
9	用久後內鍋有異味
10	電鍋鍋蓋沒有地方擺
11	每種食物不確定要放多少水，放多了還
12	每個食物要加的水量跟時間都很難
13	異味、難清、水量不好掌握、忘記
14	鍋底很難清理



歸納相似評論

- 透過分群方法找出相似言論

德不孤，必有鄰

進行文字分群

使用dependency parser

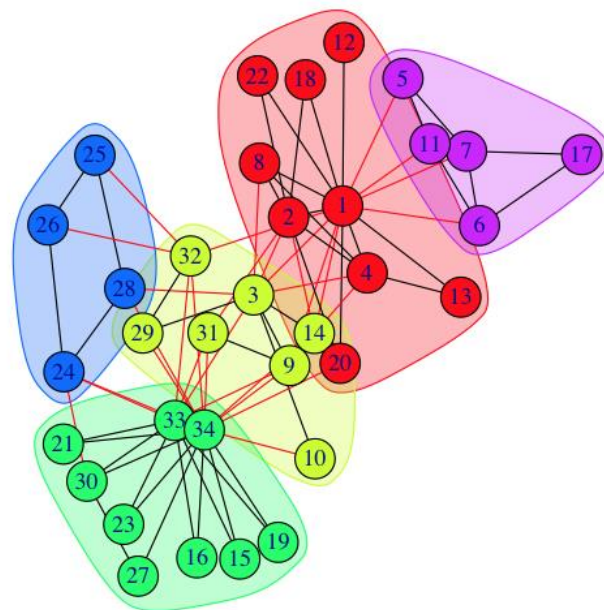
鍋內難清理
鍋內難清理，不知放幾瓢水
電鍋內外常常沾油污很難清潔
電鍋裡的汙垢難清理
鍋內髒了好難清理

清潔



清理

找出同義詞?



最佳解...

■ 人工智慧打不贏工人智慧？

104人力銀行 | 學生打工

f 分享



學生打工 > 工讀精選 > 檔期工讀



f 讚 91

熱門：晚班 隨時工 時薪150 工讀生 6月展場

職務、公司	地區	上班時段	時薪	查詢
-------	----	------	----	----

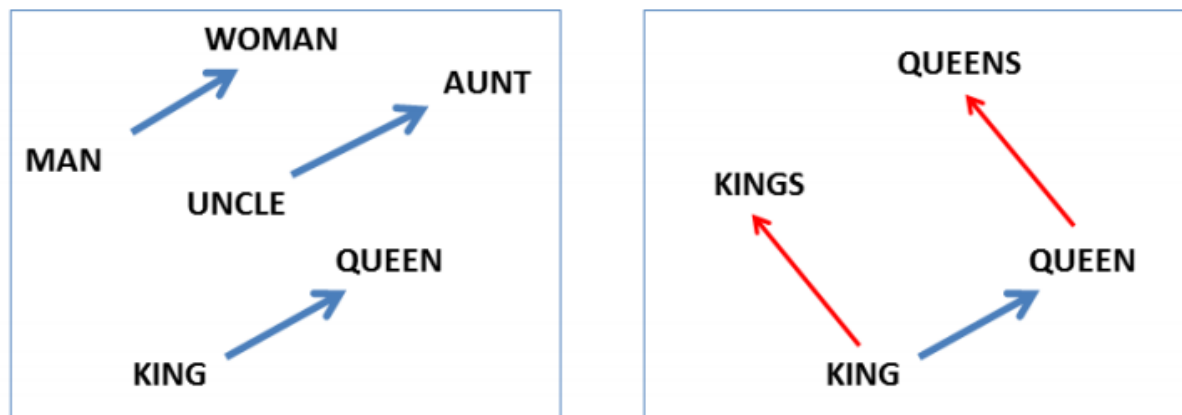
查詢條件：全部工作

儲存工作

共 854 筆，第 1 頁

Word2Vec

■ 聽說深度學習除了會下圍棋也會找同義字？



(Mikolov et al., NAACL HLT, 2013)

Word2Vec 根據文字的脈絡，將文字轉換到一 K 維度空間，並藉由判斷文字於該向量空間的距離，判斷字詞之間的相似度

建立領域詞彙網路

■ 使用分類方法產生群聚敘述

你會說鍋蓋很漂亮嗎？

鍋蓋

不知道放哪裡

燙手

水氣堆積

電鍋

空間不夠

異味

清理

句子貼標

■ 使用分類模型為句子貼標

鍋內難清理

鍋內難清理，不知放幾瓢水

電鍋內外常常沾油污很難清潔

電鍋裡的汙垢難清理

鍋內髒了好難清理



鍋內難清理

統計標籤

■ 抓出使用者的真實意見

敘述	計數
鍋內難清理	62
不知道加多少水	26
蓋子不知道要放哪裡	24
電鍋異味	22
蒸東西水氣	6
蒸東西空間不夠	2
食物還是涼的	2
鍋蓋水蒸氣到處都是	2
蒸食物味道混	2
鍋蓋發出聲音	2

The background features a light blue hexagonal grid pattern. Overlaid on this is a large, faint, circular graphic composed of concentric rings and radial lines, resembling a stylized sun or a target. The text "THANK YOU" is centered in a bold, dark blue, sans-serif font.

THANK YOU