中文斷詞：斷句不要悲劇

Head first Chinese text segmentation

Fukuball Lin @ iThome TechTalk

關於我

Fukuball / 林志傑

KKbox

幕僚工程師

CodeTengu　iNDIEVOX

中文斷詞是什麼？

- 讓電腦把詞彙以「意義」為單位切割出來
- 例如：塵世中一個迷途小書僮
  - X 塵 / 世 / 中 / 一 / 個 / 迷 / 途 / 小 / 書 / 僮
  - O 塵世 / 中 / 一個 / 迷途 / 小 / 書僮

中文斷詞與英文斷詞不同

- 我們在野生動物園玩 vs We play at the wildlife park
- We / play / at / the / wildlife / park
- 我們 / 在野 / 生動 / 物 / 園 / 玩 or 我們 / 在 / 野生 / 動物園 / 玩

中文斷詞的用處

- 文本分析研究
- 問答系統、自動摘要、文件檢索、機器翻譯、語音辨識

中文斷詞技術的難題

- 新詞識別
  - 特有名詞：人名、地名，魯蛇、溫拿
- 歧異詞識別
  - 我們 / 在野 / 生動 / 物 / 園 / 玩 or 我們 / 在 / 野生 / 動物園 / 玩
- 表情符號識別
  - XDDDD、:)、0rz

常用解法

- 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
- 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

中文斷詞：首選 Jieba

中研院也有中文斷詞系統啊？

曾經我也使用中研院斷詞系統，
直到我膝蓋中了一箭

Jieba 結巴

open source

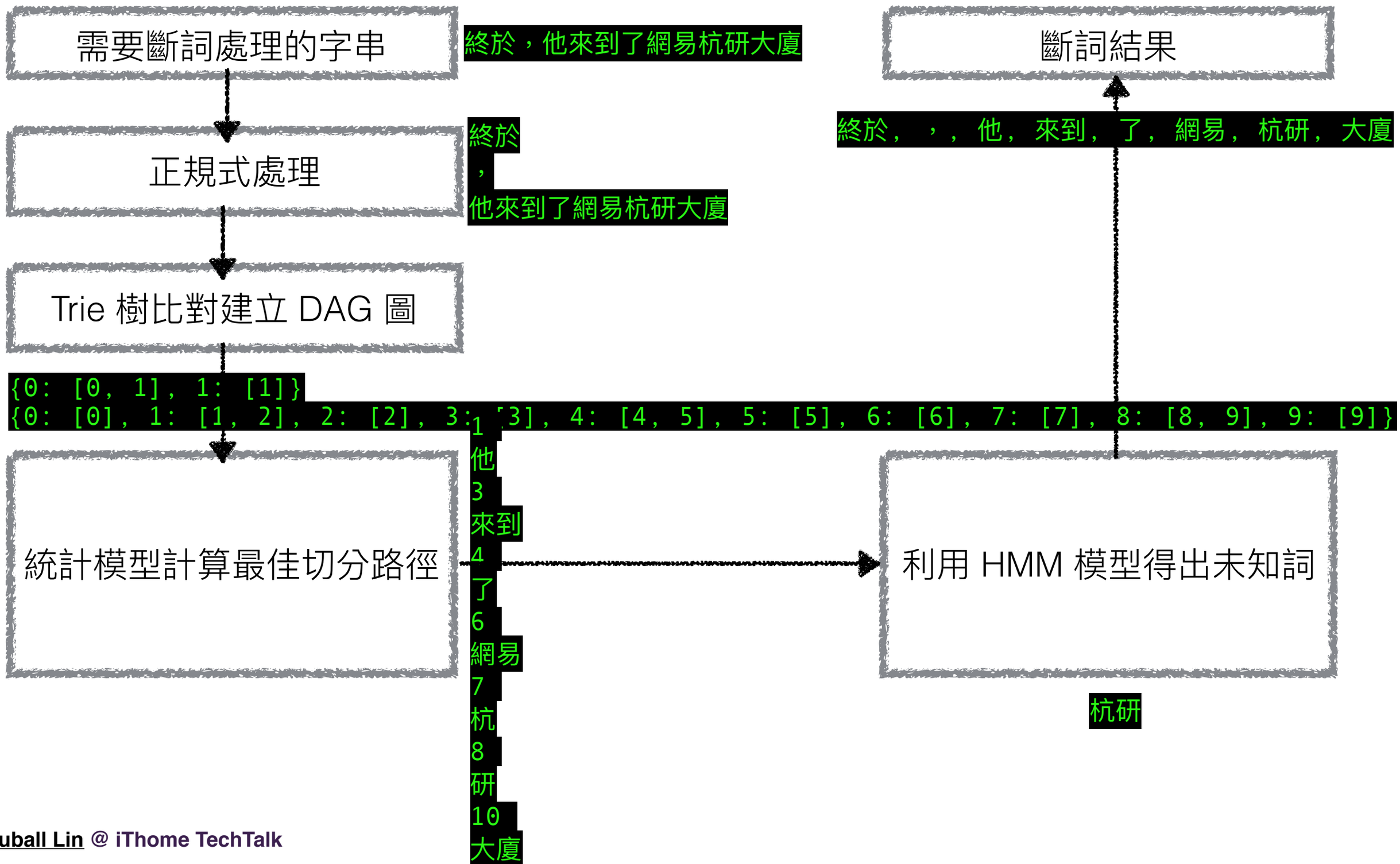
擁抱開源碼

Jieba 斷詞演算法

Jieba 斷詞演算法

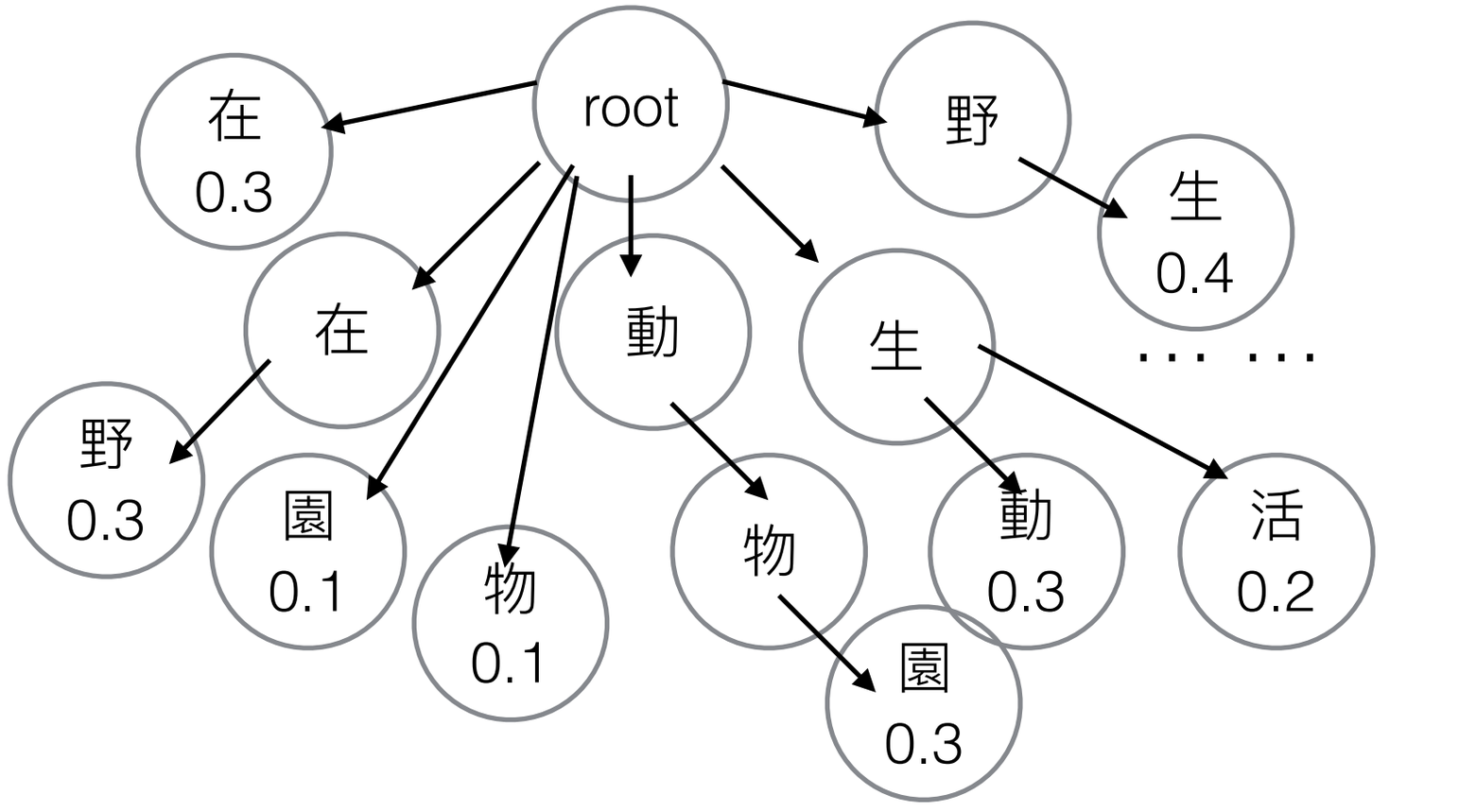
- 建立 Trie DAG，使用全切分方法，統計模型計算最佳結果
- 未知詞（新詞）使用 HMM 模型計算辨識出來

Jieba 結巴斷詞演算法概觀

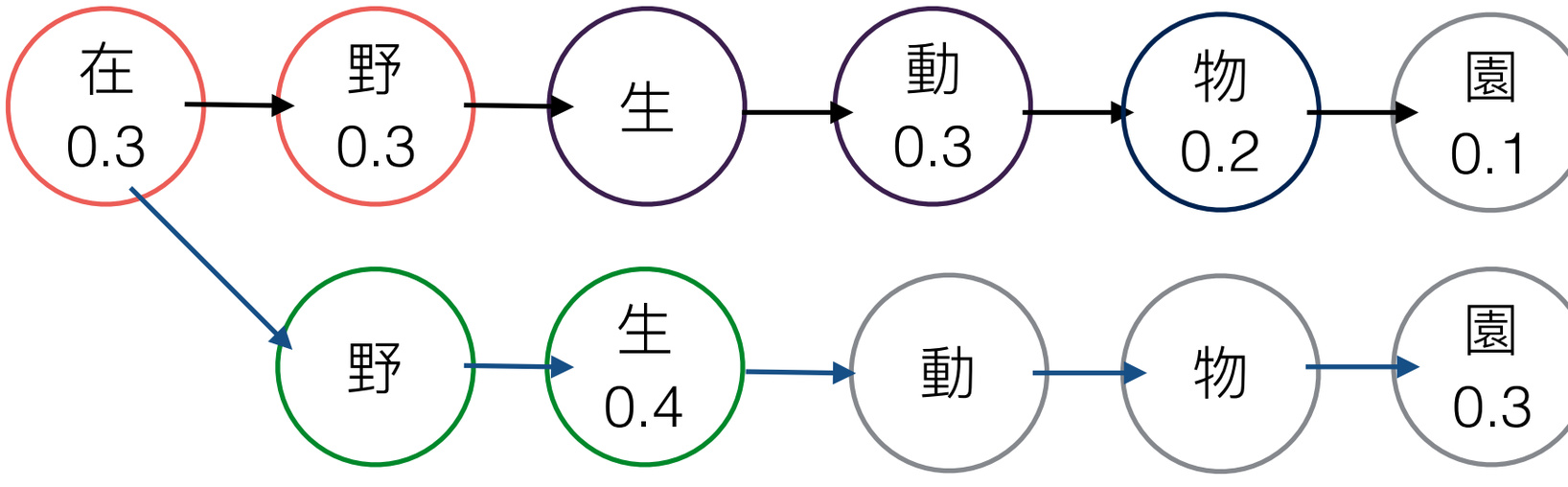


Trie DAG 計算最佳切分路徑（1）

- Trie 樹 - 前綴樹、字典樹，增加比對速度

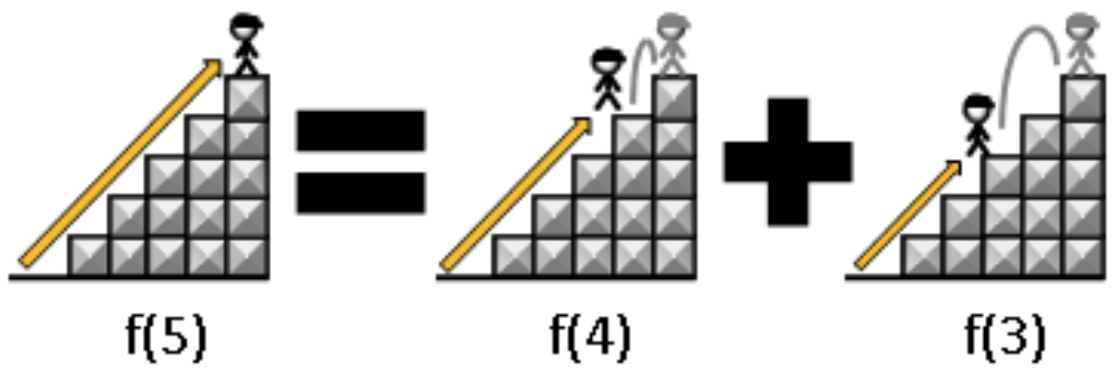


- DAG 有向無環圖

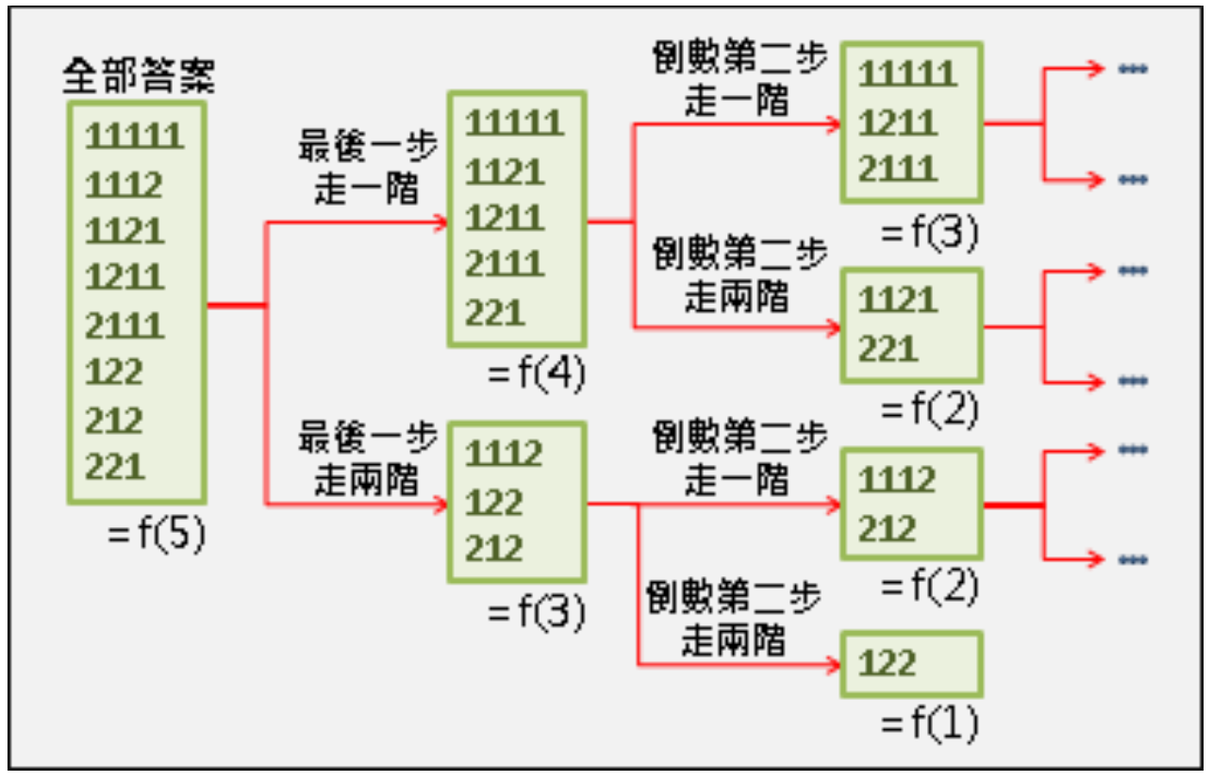


Trie DAG 計算最佳切分路徑（2）

- 使用動態規劃計算斷詞的切分組合（加快計算速度）

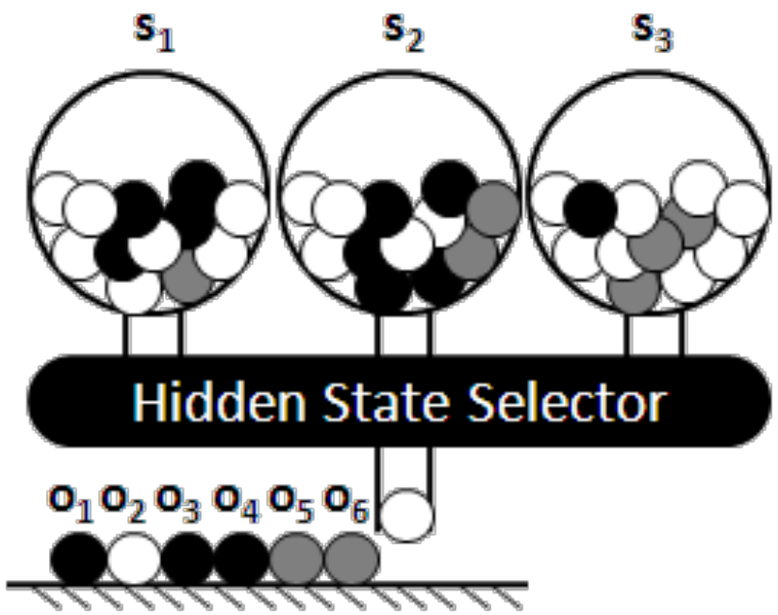


- 舉例：斷詞就像爬樓梯



HMM 模型及 Viterbi 算法（1）

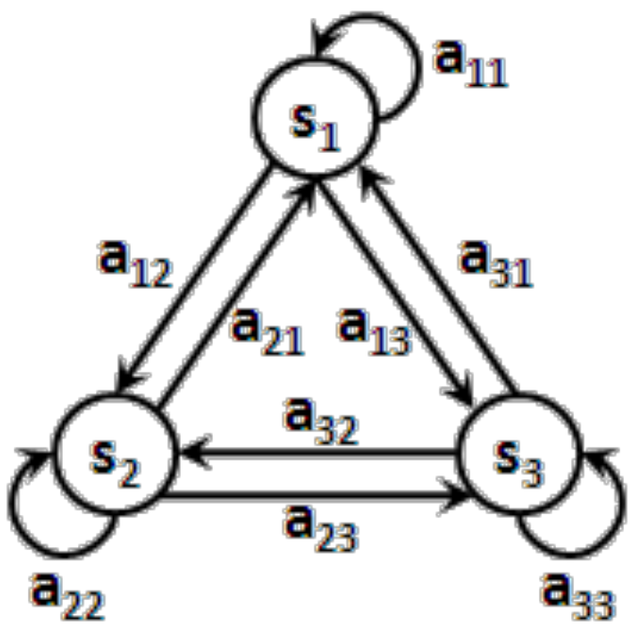
- 什麼是 HMM 隱馬可夫模型（Hidden Markov Model）
- 只能觀察到觀察序列 O（果），無法觀察到狀態序列 S（因）



馬可夫模型補充（1）

- 馬可夫模型：選一個狀態作為起點，然後沿著邊隨意走訪任何一個狀態，一直走一直走，沿途累積機率，走累了就停在某狀態。

- 舉例：猜天氣，可直接觀察到天氣狀態及轉移機率



馬可夫模型補充（2）

- 有一名旅客，三天後想到台南遊玩，由氣象報告得知今天的降雨機率為 0.2，也知道晴天雨天的轉移機率如下，則此遊客三天後到台南遇到下雨的機率為多少？

	晴天	雨天
晴天	0.9	0.8
雨天	0.1	0.2

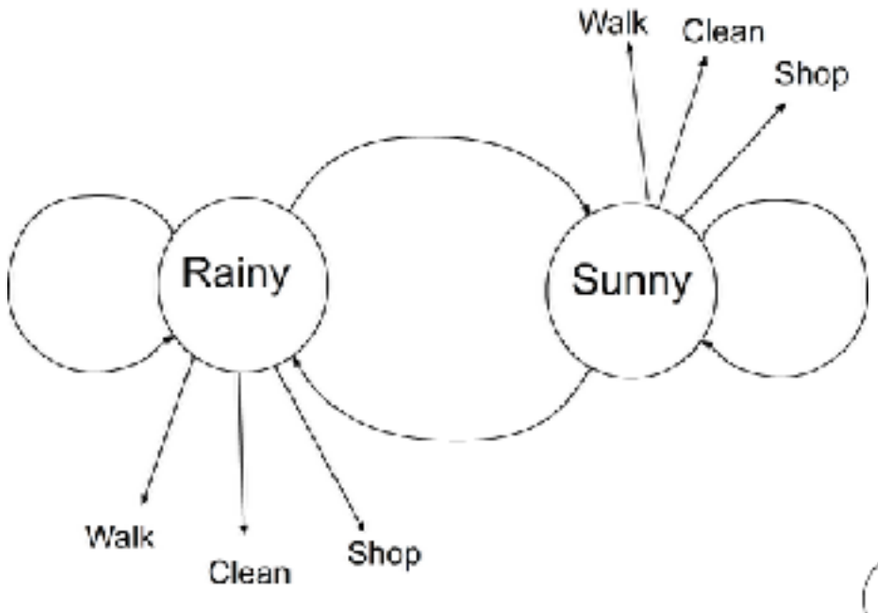
$$T^2 = \begin{bmatrix}0.9 & 0.8\\ 0.1 & 0.2\end{bmatrix}\begin{bmatrix}0.9 & 0.8\\ 0.1 & 0.2\end{bmatrix} = \begin{bmatrix}0.89 & 0.88\\ 0.11 & 0.12\end{bmatrix}$$

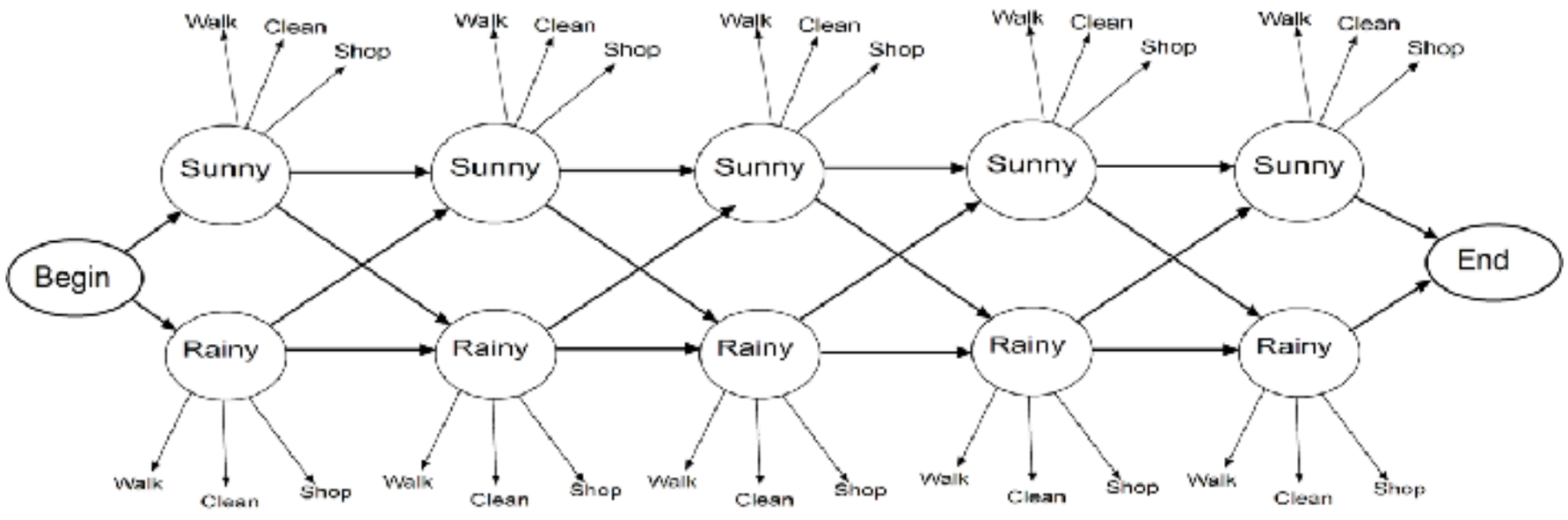
$$T^3 = \begin{bmatrix}0.89 & 0.88\\ 0.11 & 0.12\end{bmatrix}\begin{bmatrix}0.9 & 0.8\\ 0.1 & 0.2\end{bmatrix} = \begin{bmatrix}0.889 & 0.888\\ 0.111 & 0.112\end{bmatrix}$$

$$X_3 = T^3 X_0 = \begin{bmatrix}0.889 & 0.888\\ 0.111 & 0.112\end{bmatrix}\begin{bmatrix}0.8\\ 0.2\end{bmatrix} = \begin{bmatrix}0.8888\\ 0.1112\end{bmatrix}$$

HMM 模型及 Viterbi 算法（2）

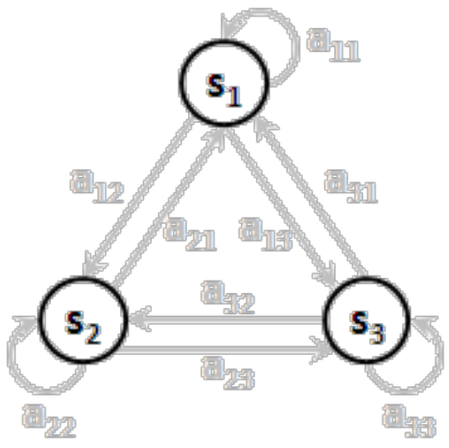
- 舉例：猜天氣，只能看到人們的行為，但看不到天氣狀態，所以由觀察行為來估算實際天氣情況





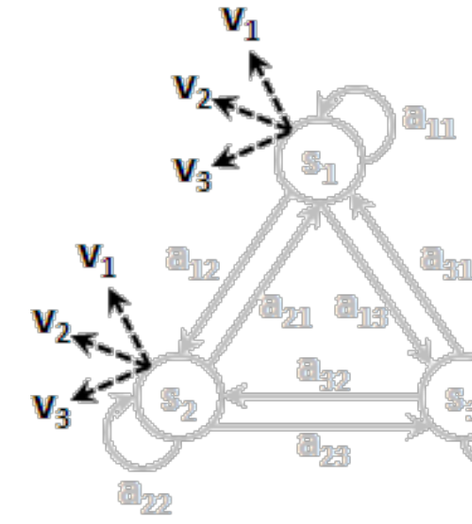
HMM 模型及 Viterbi 算法（3）

- 隱藏狀態



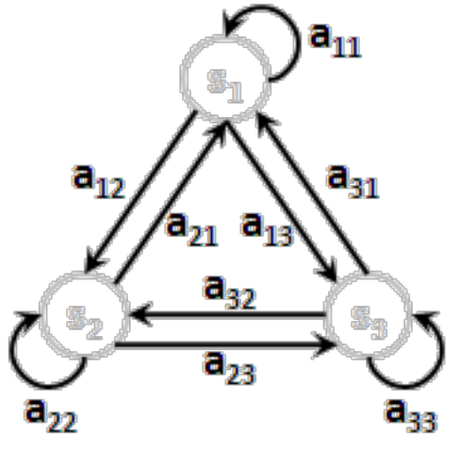
$S = \{s_1 , s_2 , s_3\}$

- 觀察狀態



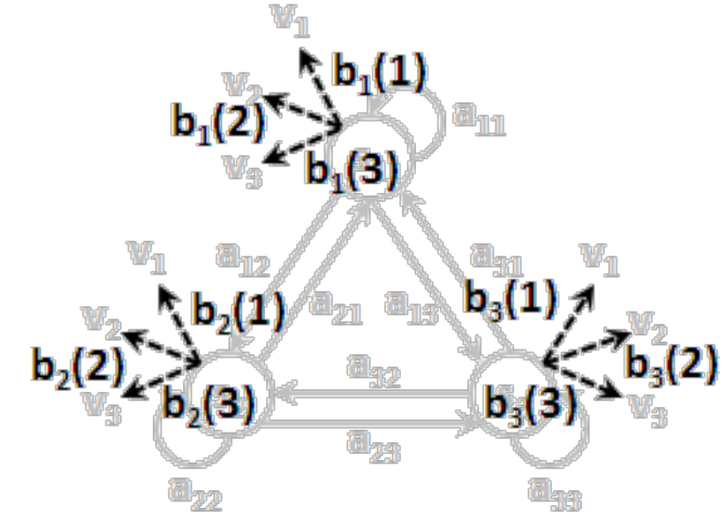
$V = \{v_1 , v_2 , v_3\}$

- 轉移機率



$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

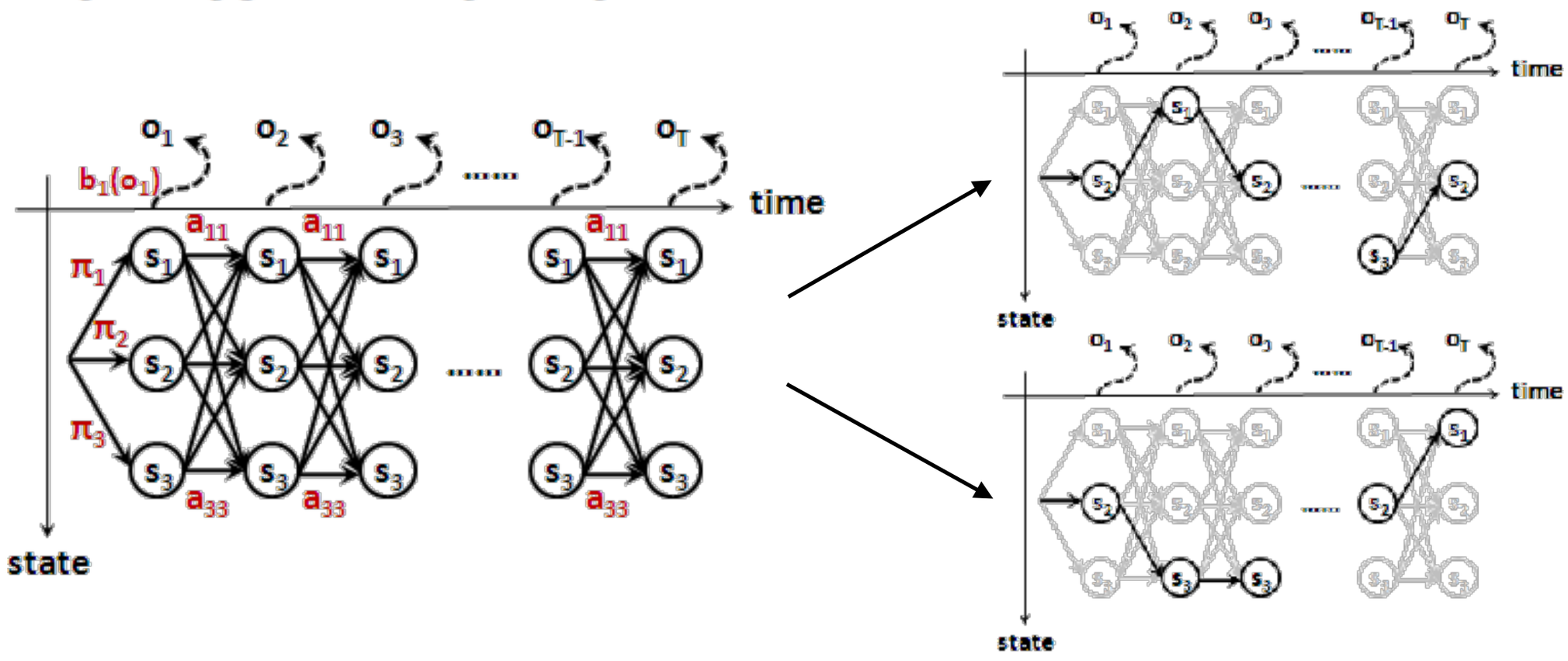
- 觀察狀態機率



HMM 模型及 Viterbi 算法（4）

- 其中一條路徑的算法

$$P(s_{q_1}) \times P(v_{o_1} \mid s_{q_1}) \times P(s_{q_2} \mid s_{q_1}) \times P(v_{o_2} \mid s_{q_2}) \times \ldots\ldots \times P(s_{q_T} \mid s_{q_{T-1}}) \times P(v_{o_T} \mid s_{q_T})$$



HMM 模型及 Viterbi 算法（5）

```
states = ('Rainy', 'Sunny')

observations = ('walk', 'shop', 'clean')

start_probability = {'Rainy': 0.6, 'Sunny': 0.4}

transition_probability = {
   'Rainy' : {'Rainy': 0.7, 'Sunny': 0.3},
   'Sunny' : {'Rainy': 0.4, 'Sunny': 0.6},
}

emission_probability = {
   'Rainy' : {'walk': 0.1, 'shop': 0.4, 'clean': 0.5},
   'Sunny' : {'walk': 0.6, 'shop': 0.3, 'clean': 0.1},
}
```

觀察狀態：走 走 走

晴 晴 晴 0.4*(0.6)*0.6*(0.6)*0.6*(0.6) = 0.031104
雨 晴 晴 0.6*(0.1)*0.3*(0.6)*0.6*(0.6) = 0.003888
雨 雨 晴 0.6*(0.1)*0.7*(0.1)*0.3*(0.6) = 0.000756
雨 晴 雨 0.6*(0.1)*0.3*(0.6)*0.4*(0.1) = 0.000432
晴 雨 雨 0.4*(0.6)*0.4*(0.1)*0.7*(0.1) = 0.000672
雨 雨 雨 0.6*(0.1)*0.7*(0.1)*0.7*(0.1) = 0.000294

最大機率組合：晴晴晴！！！

HMM 模型及 Viterbi 算法（6）

- 轉換到斷詞（看原始碼幫助理解）
    - 隱藏狀態：BMES，B(開頭) M(中間) E(結尾) S(獨立成詞)
    - 觀察狀態：所有可以看到的字
- 由觀察到的字詞序列，計算出最大的 BMES 機率組合
- 在野生動物園：SBEBME

Jieba 結巴實作

結巴實作

Virtualenv 安裝與使用

安裝

```
$ [sudo] pip install virtualenv
```

創建虛擬環境

```
$ virtualenv ENV
```

進入虛擬環境資料夾

```
$ cd ENV
```

啟動虛擬環境

```
$ source bin/activate
```

退出虛擬環境

```
$ deactivate
```

範例程式碼下載

- 全部檔案位址
  - https://bit.ly/chinese-seg

結巴實作

斷詞精確模式 (demo01)

```
#encoding=utf-8
import jieba

jieba.set_dictionary("data/dict.txt.big")

seg_list = jieba.cut("塵世中一個迷途小書僮")
print(" / ".join(seg_list))

seg_list = jieba.cut("我們在野生動物園玩")
print(" / ".join(seg_list)) # 歧異詞辨識

seg_list = jieba.cut("林志傑是結巴 PHP 的作者")
print(" / ".join(seg_list)) # 新詞辨識
```

斷詞精確模式執行結果

```
塵世 / 中 / 一個 / 迷途 / 小 / 書僮
我們 / 在 / 野生 / 動物園 / 玩
林志傑 / 是 / 結巴 / PHP / 的 / 作者
```

斷詞全模式 (demo02)

```
#encoding=utf-8
import jieba

jieba.set_dictionary("data/dict.txt.big")

seg_list = jieba.cut("我來到北京清華大學")
print(" / ".join(seg_list))

seg_list = jieba.cut("我來到北京清華大學",
cut_all=True)
print(" / ".join(seg_list))
```

斷詞全模式執行結果

```
我 / 來到 / 北京 / 清華大學

我 / 來到 / 北京 / 清華 / 清華大學 / 華大 / 大學
```

結巴實作

斷詞返回原文的起止位置 (demo03)

```
#encoding=utf-8
import jieba

jieba.set_dictionary("data/dict.txt.big")

result = jieba.tokenize(u'圖畫裡，龍不吟，虎不嘯，小小書僮可笑可笑')
for tk in result:
    print("word %s\t\t start: %d \t\t end:%d"
% (tk[0],tk[1],tk[2]))
```

結巴實作

斷詞返回原文的起止位置執行結果

```
word 圖畫    start: 0      end:2
word 裡     start: 2      end:3
word ，     start: 3      end:4
word 龍不吟   start: 4      end:7
word ，     start: 7      end:8
word 虎不嘯   start: 8      end:11
word ，     start: 11     end:12
word 小小    start: 12     end:14
word 書僮    start: 14     end:16
word 可笑    start: 16     end:18
word 可笑    start: 18     end:20
```

詞性標注 (demo04)

```
#encoding=utf-8
import jieba
import jieba.posseg as pseg

jieba.set_dictionary("data/dict.txt.big")

words = pseg.cut("颱風就是要泛舟啊不然要幹嘛")
for word, flag in words:
    print('%s %s' % (word, flag))
```

詞性標注執行結果

```
颱風 x
就是 d
要 v
泛舟 nz
啊 zg
不然 c
要 v
幹嘛 x
```

詞性列表：https://gist.github.com/luw2007/6016931

使用實例一 (demo05)

回聲樂團

座右銘

我沒有心
我沒有真實的自我
我只有消瘦的臉孔
所謂軟弱
所謂的順從一向是我
的座右銘

我沒有力
我沒有滿腔的熱火
我只有滿肚的如果
所謂勇氣
所謂的認同感是我
隨便說說

而我
沒有那海洋的寬闊
我只要熱情的撫摸
所謂空洞
所謂不安全感是我
的墓誌銘

而你
是否和我一般怯懦
是否和我一般矯作
是否對你來說
只是一場遊戲
雖然沒有把握

而你
是否和我一般怯懦
是否和我一般矯作
和我一般囉唆

而你
是否和我一般退縮
是否和我一般肌迫
是否對你來說

結巴實作

使用實例：中文歌詞斷詞，使用預設詞庫

```
#encoding=utf-8
import jieba

content = open('data/lyric1.txt', 'rb').read()

print "Input：", content

words = jieba.cut(content)
print(" / ".join(words))
```

使用實例：中文歌詞斷詞，使用預設詞庫執行結果

```
我 / 沒 / 有心 / 我 / 沒 / 有 / 真實 / 的 / 自我 / 我 / 只有 /
消瘦 / 的 / 臉孔 / 所謂 / 軟弱 / 所謂 / 的 / 順 / 從 / 一向 / 是
/ 我 / 的 / 座 / 右銘 / 而 / 我 / 沒有 / 那 / 海洋 / 的 / 寬闊 /
我 / 只要 / 熱情 / 的 / 撫 / 摸 / 所謂 / 空洞 / 所謂 / 不安全感 /
是 / 我 / 的 / 墓誌 / 銘 / 而 / 你 / 是否 / 和 / 我 / 一般 / 怯
懦 / 是否 / 和 / 我 / 一般 / 矯作 / 和 / 我 / 一般 / 囉 / 唆 /
而 / 你 / 是否 / 和 / 我 / 一般 / 退縮 / 是否 / 和 / 我 / 一般 /
肌迫 / 一般 / 地 / 困惑 / 我 / 沒 / 有力 / 我 / 沒 / 有 / 滿腔 /
的 / 熱火 / 我 / 只有 / 滿肚 / 的 / 如果 / 所謂 / 勇氣 / 所謂 /
的 / 認 / 同感 / 是 / 我 / 隨便 / 說 / 說 / 而 / 你 / 是否 /
和 / 我 / 一般 / 怯懦 / 是否 / 和 / 我 / 一般 / 矯作 / 是否 / 對
/ 你 / 來 / 說 / 只是 / 一場 / 遊戲 / 雖然 / 沒 / 有把握 / 而 /
你
 / 是否 / 和 / 我 / 一般 / 退縮 / 是否 / 和 / 我 / 一般 / 肌
迫 / 是否 / 對 / 你 / 來 / 說 / 只是 / 逼不得已 / 雖然 / 沒有 /
藉口
```

中文歌詞斷詞，使用預設詞庫結果分析

- 「座右銘」被斷成了「座 / 右銘」
- 「墓誌銘」被斷成了「墓誌 / 銘」
- 預設詞庫是簡體中文

結巴實作

使用實例：中文歌詞斷詞，使用繁體詞庫 (demo06)

```
#encoding=utf-8
import jieba

jieba.set_dictionary("data/dict.txt.big")

content = open('data/lyric1.txt', 'rb').read()

print "Input：", content

words = jieba.cut(content)
print(" / ".join(words))
```

使用實例：中文歌詞斷詞，使用繁體詞庫執行結果

我 / 沒有 / 心 / 我 / 沒有 / 真實 / 的 / 自我 / 我 / 只有 / 消瘦 / 的 / 臉孔 / 所謂 / 軟弱 / 所謂 / 的 / 順從 / 一向 / 是 / 我 / 的 / 座右銘 / 而 / 我 / 沒有 / 那 / 海洋 / 的 / 寬闊 / 我 / 只要 / 熱情 / 的 / 撫摸 / 所謂 / 空洞 / 所謂 / 不安全感 / 是 / 我 / 的 / 墓誌銘 / 而 / 你 / 是否 / 和 / 我 / 一般 / 怯懦 / 是否 / 和 / 我 / 一般 / 矯作 / 和 / 我 / 一般 / 囉唆 / 而 / 你 / 是否 / 和 / 我 / 一般 / 退縮 / 是否 / 和 / 我 / 一般 / 肌迫 / 一般 / 地 / 困惑 / 我 / 沒有 / 力 / 我 / 沒有 / 滿腔 / 的 / 熱火 / 我 / 只有 / 滿肚 / 的 / 如果 / 所謂 / 勇氣 / 所謂 / 的 / 認同感 / 是 / 我 / 隨便說說 / 而 / 你 / 是否 / 和 / 我 / 一般 / 怯懦 / 是否 / 和 / 我 / 一般 / 矯作 / 是否 / 對 / 你 / 來說 / 只是 / 一場 / 遊戲 / 雖然 / 沒有 / 把握 / 而 / 你 / 是否 / 和 / 我 / 一般 / 退縮 / 是否 / 和 / 我 / 一般 / 肌迫 / 是否 / 對 / 你 / 來說 / 只是 / 逼不得已 / 雖然 / 沒有 / 藉口

結巴實作

中文歌詞斷詞，使用繁體詞庫結果分析

- 「座右銘」成功斷成「座右銘」
- 「墓誌銘」也成功斷成「墓誌銘」

結巴實作

使用實例：取出文章中的關鍵詞 (demo07)

```
#encoding=utf-8
import jieba
import jieba.analyse

jieba.set_dictionary("data/dict.txt.big")

content = open('data/lyric1.txt', 'rb').read()

print "Input：", content

tags = jieba.analyse.extract_tags(content, 10)
print "Output："
print(" / ".join(tags))
```

結巴實作

使用實例：取出文章中的關鍵詞執行結果

所謂 / 沒有 / 是否 / 一般 / 矯作 / 來說 / 怯懦 / 墓誌銘 / 退縮 / 寬闊

結巴實作

TF-IDF 關鍵詞算法

- 某個詞在一篇文章中出現的頻率高，且在其他文章中很少出現，則此詞語為具代表性的關鍵詞
- Term Frequency

$$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

- Inverse Document Frequency

$$\mathrm{idf_i} = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

- TF-IDF

$$\mathrm{tfidf_{i,j}} = \mathrm{tf_{i,j}} \times \mathrm{idf_i}$$

結巴實作

使用實例：關鍵詞去除停用字 (demo08)

```
#encoding=utf-8
import jieba
import jieba.analyse

jieba.set_dictionary("data/dict.txt.big")
jieba.analyse.set_stop_words("data/stop_words.txt")

content = open('data/lyric1.txt', 'rb').read()

print "Input：", content

tags = jieba.analyse.extract_tags(content, 10)
print "Output："
print(" / ".join(tags))
```

使用實例：關鍵詞去除停用字執行結果

所謂 / 一般 / 矯作 / 來說 / 怯懦 / 墓誌銘 / 退縮 / 寬闊 / 順從 / 熱情

如何再提高斷詞的準確性？

- 調整文本資料，如 HMM 模型，字典詞頻
- 調整演算法（使用現在號稱最準的 Deep Learning）
- 使用自定義詞典

結巴實作

Jieba 自定義詞典用法

```
#encoding=utf-8
import jieba

jieba.set_dictionary("data/dict.txt.big")
jieba.load_userdict("data/userdict.txt")
```

結巴實作

Jieba 動態新增詞典 (demo09)

```
#encoding=utf-8
import jieba

jieba.set_dictionary("data/dict.txt.big")
jieba.add_word(word, freq=None, tag=None)
```

使用實例二 (demo10)

滅火器

島嶼天光

親愛的媽媽
請你毋通煩惱我
原諒我
行袂開跤
我欲去對抗袂當原諒
的人

歹勢啦
愛人啊
袂當陪你去看電影
原諒我
行袂開跤
我欲去對抗欺負咱的
人

天色漸漸光
遮有一陣人
為了守護咱的夢
成做更加勇敢的人

已經袂記
是第幾工
請毋通煩惱我
因為阮知道
無行過寒冬
袂有花開的一工

天色漸漸光
天色漸漸光
已經是更加勇敢的人

天色漸漸光
咱就大聲來唱著歌
一直到希望的光線
照光島嶼每一個人

天色漸漸光
咱就大聲來唱著歌
日頭一爬上山

結巴實作

使用實例：台語歌詞斷詞，使用繁體詞庫

```
#encoding=utf-8
import jieba

jieba.set_dictionary("data/dict.txt.big")

content = open('data/lyric2.txt', 'rb').read()

print "Input：", content

words = jieba.cut(content)
print(" / ".join(words))
```

使用實例：台語歌詞斷詞，使用繁體詞庫執行結果

親愛 / 的 / 媽媽 / 請 / 你 / 毋通 / 煩惱 / 我 / 原諒 / 我 / 行袂 / 開跤 / 我 / 欲 / 去 / 對抗 / 袂 / 當 / 原諒 / 的 / 人 / 歹勢 / 啦 / 愛人 / 啊 / 袂 / 當 / 陪你去 / 看 / 電影 / 原諒 / 我 / 行袂 / 開跤 / 我 / 欲 / 去 / 對抗 / 欺負 / 咱 / 的 / 人 / 天色 / 漸漸 / 光 / 遮有 / 一陣 / 人 / 為 / 了 / 守護 / 咱 / 的 / 夢 / 成 / 做 / 更加 / 勇敢的人 / 天色 / 漸漸 / 光 / 已經 / 不再 / 驚惶 / 現在 / 就是 / 彼一工 / 換阮 / 做 / 守護 / 恁 / 的 / 人 / 已經 / 袂 / 記 / 是 / 第幾 / 工 / 請 / 毋通 / 煩惱 / 我 / 因為 / 阮 / 知道 / 無行過 / 寒冬 / 袂 / 有 / 花開 / 的 / 一工 / 天色 / 漸漸 / 光 / 天色 / 漸漸 / 光 / 已經 / 是 / 更加 / 勇敢的人 / 天色 / 漸漸 / 光 / 咱 / 就 / 大聲 / 來 / 唱 / 著歌 / 一直 / 到 / 希望 / 的 / 光線 / 照光 / 島嶼 / 每 / 一個 / 人 / 天色 / 漸漸 / 光 / 咱 / 就 / 大聲 / 來 / 唱 / 著歌 / 日頭 / 一爬 / 上山 / 就 / 會 / 使 / 轉去 / 啦 / 現在 / 是 / 彼 / 一工 / 勇敢 / 的 / 台灣 / 人

結巴實作

台語歌詞斷詞，使用繁體詞庫結果分析

- 「袂當」斷成了「袂」「當」
- 「袂記」斷成了「袂」「記」
- 「袂有」斷成了「袂」「有」

結巴實作

使用實例：台語歌詞斷詞，使用繁體詞庫加自定義詞庫 (demo11)

```
#encoding=utf-8
import jieba

jieba.set_dictionary("data/dict.txt.big")
jieba.load_userdict("data/userdict.txt")

content = open('data/lyric2.txt', 'rb').read()

print "Input：", content

words = jieba.cut(content)
print(" / ".join(words))
```

使用實例：台語歌詞斷詞，使用繁體詞庫加自定義詞庫執行結果

親愛 / 的 / 媽媽 / 請 / 你 / 毋通 / 煩惱 / 我 / 原諒 / 我 / 行袂開跤 / 我 / 欲 / 去 / 對抗 / 袂當 / 原諒 / 的 / 人 / 歹勢 / 啦 / 愛人 / 啊 / 袂當 / 陪你去 / 看 / 電影 / 原諒 / 我 / 行袂開跤 / 我 / 欲 / 去 / 對抗 / 欺負 / 咱 / 的 / 人 / 天色 / 漸漸 / 光 / 遮有 / 一陣 / 人 / 為 / 了 / 守護 / 咱 / 的 / 夢 / 成 / 做 / 更加 / 勇敢的人 / 天色 / 漸漸 / 光 / 已經 / 不再 / 驚惶 / 現在 / 就是 / 彼一工 / 換阮 / 做 / 守護 / 恁 / 的 / 人 / 已經 / 袂記 / 是 / 第幾 / 工 / 請 / 毋通 / 煩惱 / 我 / 因為 / 阮 / 知道 / 無行過 / 寒冬 / 袂有 / 花開 / 的 / 一工 / 天色 / 漸漸 / 光 / 天色 / 漸漸 / 光 / 已經 / 是 / 更加 / 勇敢的人 / 天色 / 漸漸 / 光 / 咱 / 就 / 大聲 / 來 / 唱著 / 歌 / 一直 / 到 / 希望 / 的 / 光線 / 照光 / 島嶼 / 每 / 一個 / 人 / 天色 / 漸漸 / 光 / 咱 / 就 / 大聲 / 來 / 唱著 / 歌 / 日頭 / 一爬 / 上山 / 就 / 會使 / 轉去 / 啦 / 現在 / 是 / 彼 / 一工 / 勇敢 / 的 / 台灣 / 人

結巴實作

台語歌詞斷詞，使用繁體詞庫加自定義詞庫結果分析

- 符合預期結果
- 自定義詞庫格式：

```
行袂開跤 2 v
袂當 4 d
袂記 4 v
袂有 4 d
會使 70 d
```

中文斷詞實際應用

中文斷詞應用在音樂

- 歌詞分析
- 情境歌單
- 自動填詞
- 相似歌詞推薦

實際應用

中文歌詞相似推薦系統 (1)

- Step 1：中文斷詞，集合成資料集
- Step 2：去掉停用字
- Step 3：將每首歌詞轉成向量表示(doc2vec)
- Step 4：LSA 算法降維
- Step 5：使用降維後的向量計算 cosin similarity
- 這邊使用了 gensim 套件處理步驟 2-5，但全部自幹也不算太難

實際應用

中文歌詞相似推薦系統 (2)

Demo

中文歌詞相似推薦系統 (3)

- 輸入：楊培安 我的驕傲 歌詞
- 輸出：
  - 楊培安 我的驕傲
  - 五月天 倔強
  - 張雨生 我的未來不是夢
  - 五月天 憨人
  - 五月天 一顆蘋果

實際應用

中文歌詞相似推薦系統 (4)

《 我的驕傲》節錄

沒有山不能跨越 沒有海不能冒險
讓歷史記得這一天 當我用心立下諾言
沒有事不能改變 沒有夢不能實現
我站在未來最前線 抬頭迎接每個考驗

海闊天空是我的地圖 想寫下全新紀錄
放眼天下在等我去征服 用熱血燃燒黑夜
等待最燦爛的日出

看陽光與我賽跑 風雨和我狂飆 我的驕傲自己打造
每個夢 永遠比天高 一顆心 為希望在跳躍
讓世界為我歡呼 大地為我炫耀
我的驕傲你會看到 汗和淚痛苦的煎熬
在這一刻都是我光榮的記號

《 倔強》節錄

最美的願望 一定最瘋狂
我就是我自己的神 在我活的地方
我和我最後的倔強 握緊雙手絕對不放
下一站是不是天堂 就算失望不能絕望
我和我驕傲的倔強 我在風中大聲的唱
這一次為自己瘋狂 就這一次 我和我的倔強

逆風的方向 更適合飛翔
我不怕千萬人阻擋 只怕自己投降
我和我最後的倔強 握緊雙手絕對不放
下一站是不是天堂 就算失望不能絕望
我和我驕傲的倔強 我在風中大聲的唱
這一次為自己瘋狂 就這一次 我和我的倔強

補充說明

LSA 潛在語意分析(1)

歌詞
c1: 你是我的 巧克力，你是這 世界 最美的 風景
c2: 愛情 的 甜美，添加了你我的 理想，如 巧克力 般的 快樂 滋味
c3: 你我的 世界 中，愛情 的 咖啡 那麼 甜美
c4: 愛情 如 風景 般美，愛情 如 咖啡 般濃
c5: 愛情 的 滋味 很 甜美
m1: 我要達成 夢想
m2: 抓一把 陽光，夢想 開始
m3: 等待 陽光，往前 飛翔，向 夢想
m4: 擁抱 陽光，滿懷 理想，到處 飛翔

補充說明

LSA 潛在語意分析(2)

	風景	世界	巧克力	甜美	愛情	快樂	滋味	咖啡	理想	夢想	陽光	飛翔
c1:	1	1	1	0	0	0	0	0	0	0	0	0
c2:	0	0	1	1	1	1	1	0	1	0	0	0
c3:	0	1	0	1	1	0	0	1	0	0	0	0
c4:	1	0	0	0	2	0	0	1	0	0	0	0
c5:	0	0	0	1	0	1	1	0	0	0	0	0
m1:	0	0	0	0	0	0	0	0	0	1	0	0
m2:	0	0	0	0	0	0	0	0	0	1	1	0
m3:	0	0	0	0	0	0	0	0	0	1	1	1
m4:	0	0	0	0	0	0	0	0	1	0	1	1

SVD 分解 ↓ LSA 降維　無法完整呈現歌詞情意

X≈ {D}

0.20	-0.06
0.61	0.17
0.46	-0.13
0.54	-0.23
0.28	0.11
0.00	0.19
0.01	0.44
0.02	0.62
0.08	0.53

{S}

3.34	0
0	2.54

$\{T\}^T$

0.22	0.20	0.24	0.40	0.64	0.27	0.27	0.30	0.21	0.01	0.04	0.03
-0.11	-0.07	0.04	0.06	-0.17	0.11	0.11	-0.14	0.27	0.49	0.62	0.45

row 1 有關愛情的字詞值都較高，所以Topic1應代表愛情
row 2 有關夢想的字詞值都較高，所以Topic2應代表夢想

原本要用12個詞來描述歌詞，現在只要用兩個Topic就可以描述

表示歌詞如何用兩種Topic來呈現，可明顯看出有兩類歌詞

補充說明

LSA 潛在語意分析(3)

{X'}	風景	世界	巧克力	甜美	愛情	快樂	滋味	咖哩	理想	夢想	陽光	飛翔
c1	0.16	0.14	0.15	0.26	0.45	0.16	0.16	0.22	0.10	-0.06	-0.06	-0.04
c2	0.40	0.37	0.51	0.84	1.23	0.58	0.58	0.55	0.53	0.23	0.34	0.25
c3	0.38	0.33	0.36	0.61	1.05	0.38	0.38	0.51	0.23	-0.14	-0.15	-0.10
c4	0.47	0.40	0.41	0.70	1.27	0.42	0.42	0.63	0.21	-0.27	-0.30	-0.21
c5	0.18	0.16	0.24	0.39	0.56	0.28	0.28	0.24	0.27	0.14	0.20	0.15
m1	-0.05	-0.03	0.02	0.03	-0.07	0.06	0.06	-0.07	0.14	0.24	0.31	0.22
m2	-0.12	-0.07	0.06	0.08	-0.15	0.13	0.13	-0.14	0.31	0.55	0.69	0.50
m3	-0.16	-0.10	0.09	0.12	-0.21	0.19	0.19	-0.20	0.44	0.77	0.98	0.71
m4	-0.09	-0.04	0.12	0.19	-0.05	0.22	0.22	-0.11	0.42	0.66	0.85	0.62

能夠完整呈現歌詞情意

中文歌詞相似推薦系統 (5)

- 輸入：周杰倫 安靜 歌詞
- 輸出：
  - 周杰倫 安靜
  - 黃品源 那麼愛你為什麼
  - 孫燕姿 我不難過
  - 陳奕迅 婚禮的祝福
  - 周杰倫 斷了的弦


其他議題

其他議題(1) (demo12)

- 網路上有人問：
  - 下雨天留客天留我不留
  - 海水朝朝朝朝朝朝朝朝落；浮雲長長長長長長長消
- 有各種斷法，基本上不太算是斷詞的問題

其他議題(2)

- 結巴對於新詞辨識表現還不錯，但對於歧異詞辨識則有待加強
- 歧異詞辨識目前最有效的解法是使用 Deep Learning LSTM 模型來斷詞

Q & A

Find Me

Twitter @fukuball

Github @fukuball

Facebook @fukuball