Jieba 結巴中文斷詞

Fukuball Lin @ 政大數位足跡計劃

關於我

Fukuball / 林志傑

iNDIEVOX

獨立音樂網

www.indievox.com

Jieba 結巴是什麼？

- 中文斷詞（分詞）程式
- 中文斷詞
  - 自然語言處理
  - 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

例如

- 全台大停電 vs. Power outage all over Taiwan
- Power / outage / all / over / Taiwan
- 全台 / 大 / 停電 or 全 / 台大 / 停電

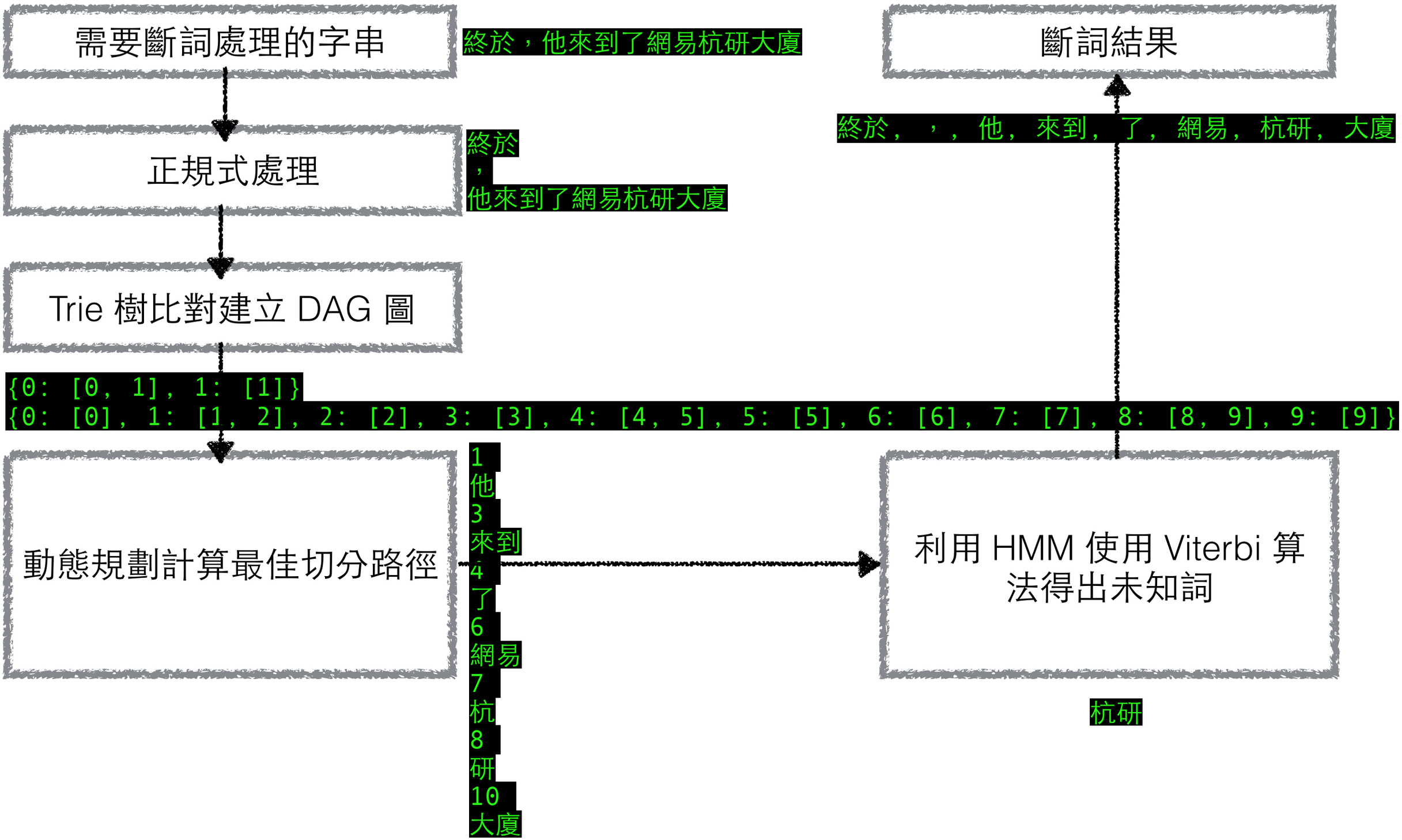
中研院也有斷詞系統啊？

曾經我也使用中研院斷詞系統，
直到我膝蓋中了一箭

open source

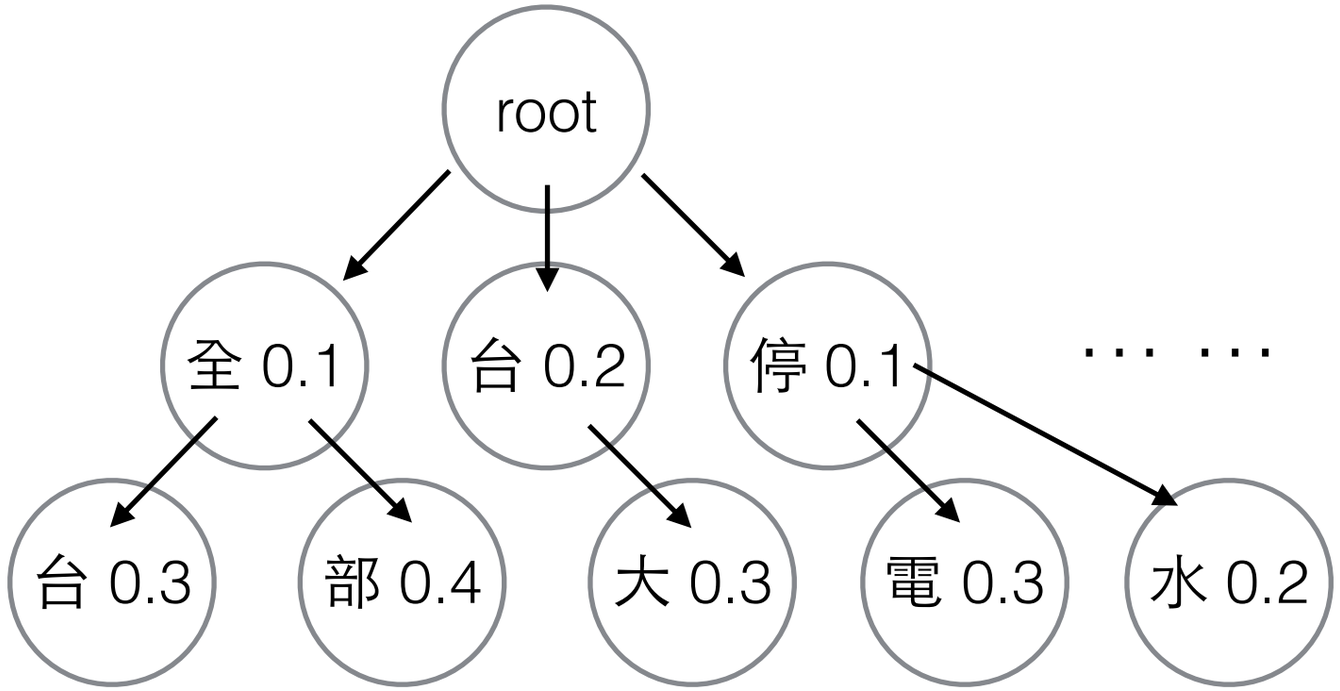
擁抱開源碼

Jieba 結巴所使用的演算法

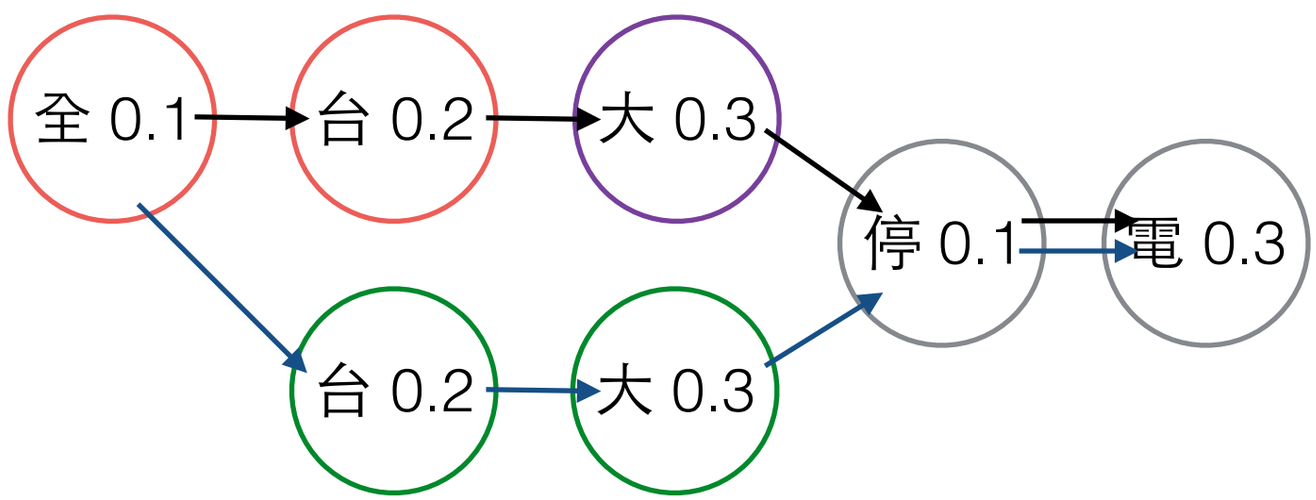


Trie DAG 動態規劃 (1)

- Trie 樹 - 前綴樹、字典樹，增加比對速度

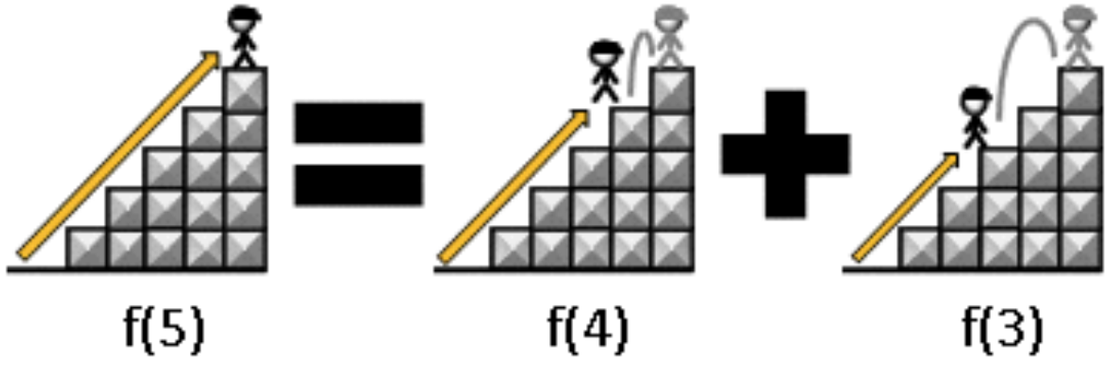


- DAG 有向無環圖

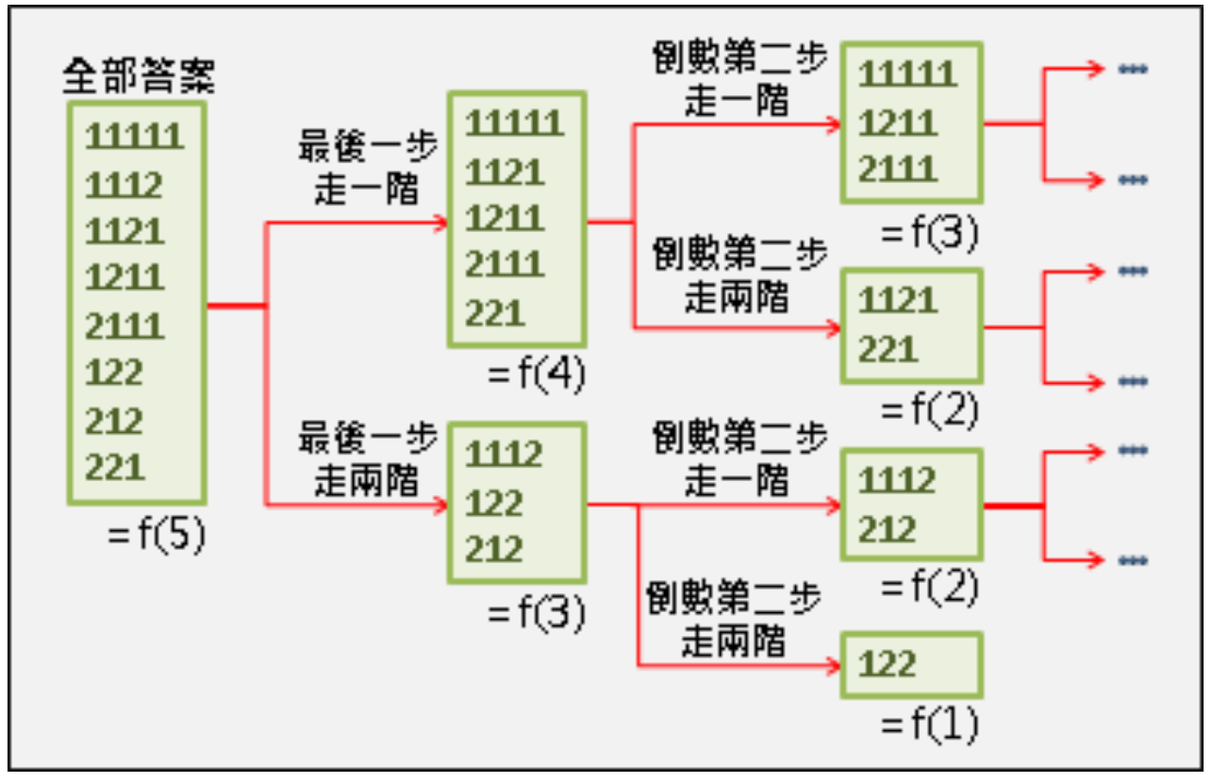


Trie DAG 動態規劃 (2)

- 使用動態規劃計算斷詞的切分組合（加快計算速度）

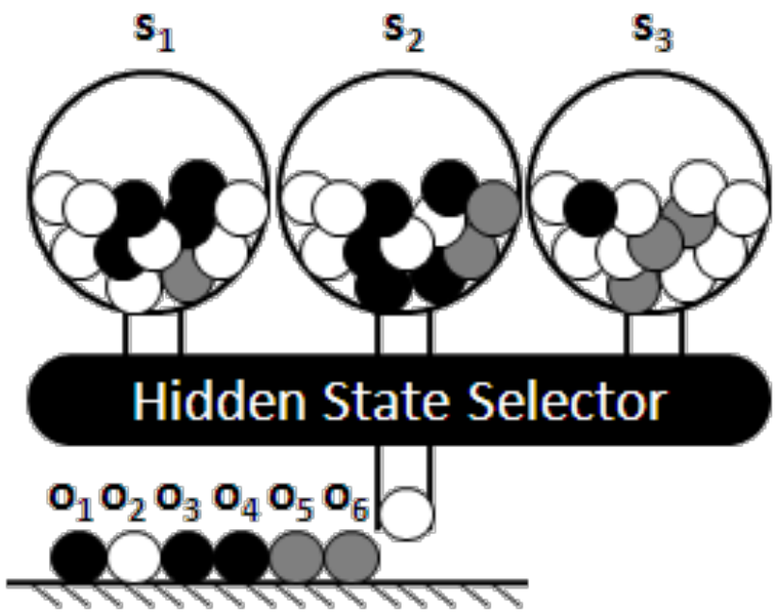


- 舉例：斷詞就像爬樓梯



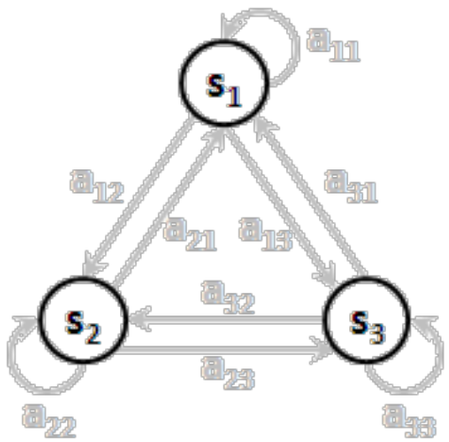
HMM 模型及 Viterbi 動態規劃算法（1）

- 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）



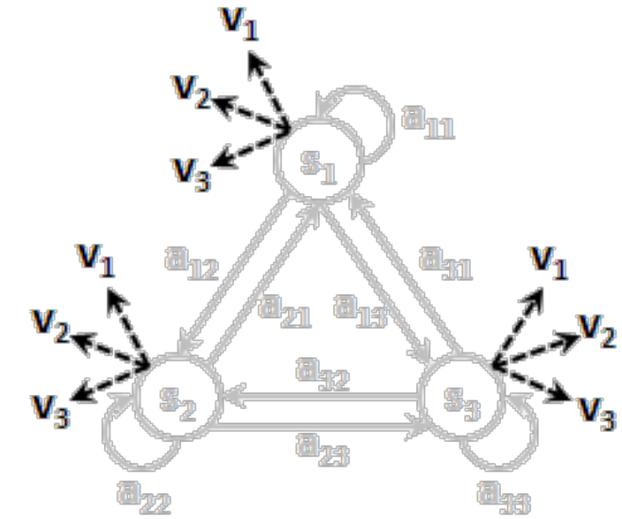
HMM 模型及 Viterbi 動態規劃算法（2）

- 隱藏狀態



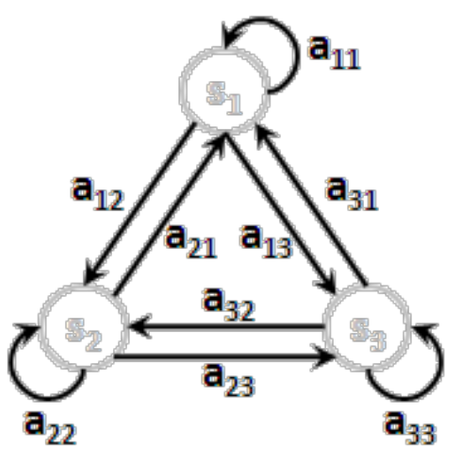
$S = \{s_1 , s_2 , s_3\}$

- 觀察狀態



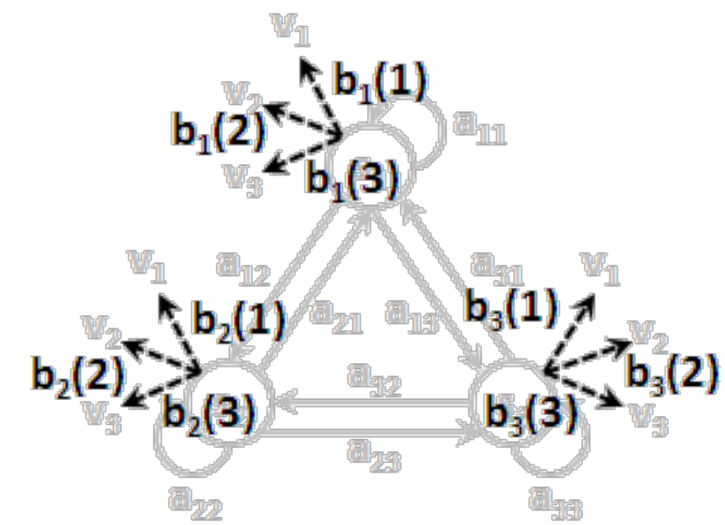
$V = \{v_1 , v_2 , v_3\}$

- 轉移機率



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

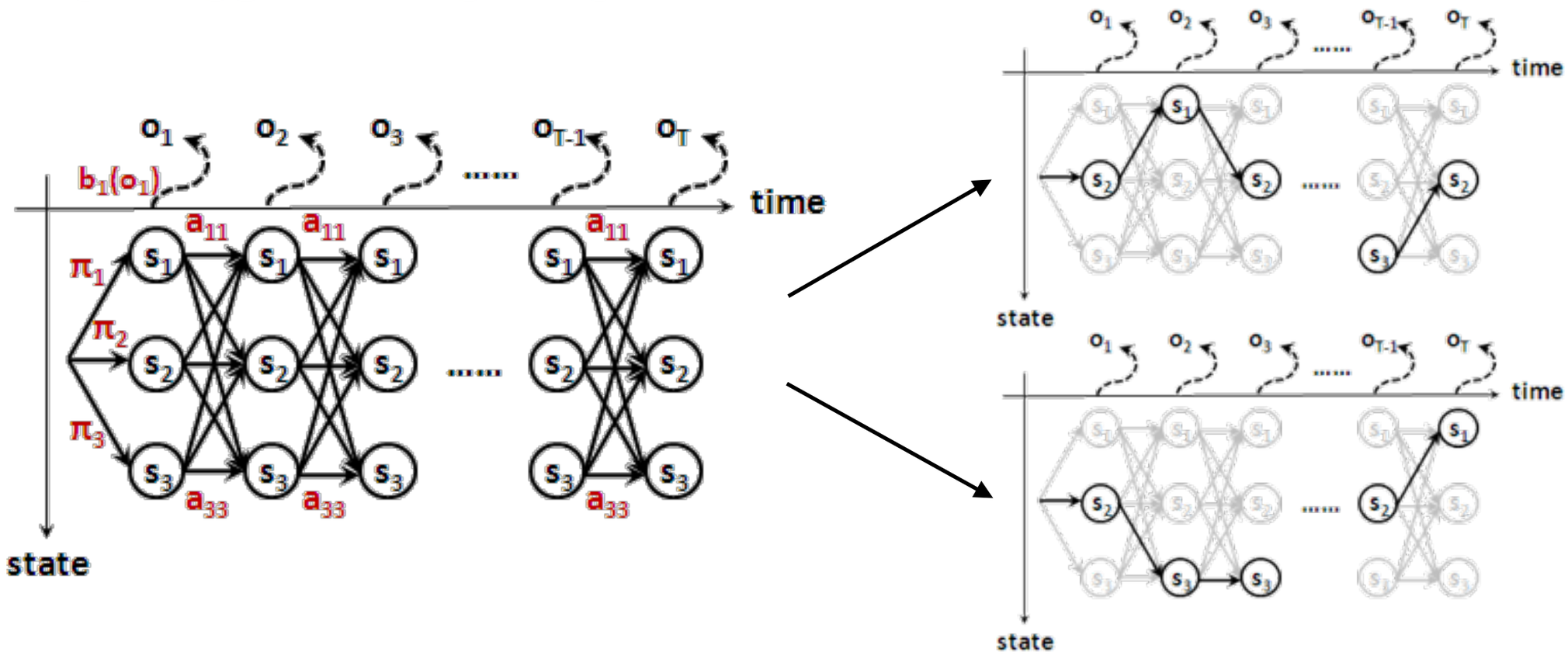
- 觀察狀態機率



HMM 模型及 Viterbi 動態規劃算法（3）

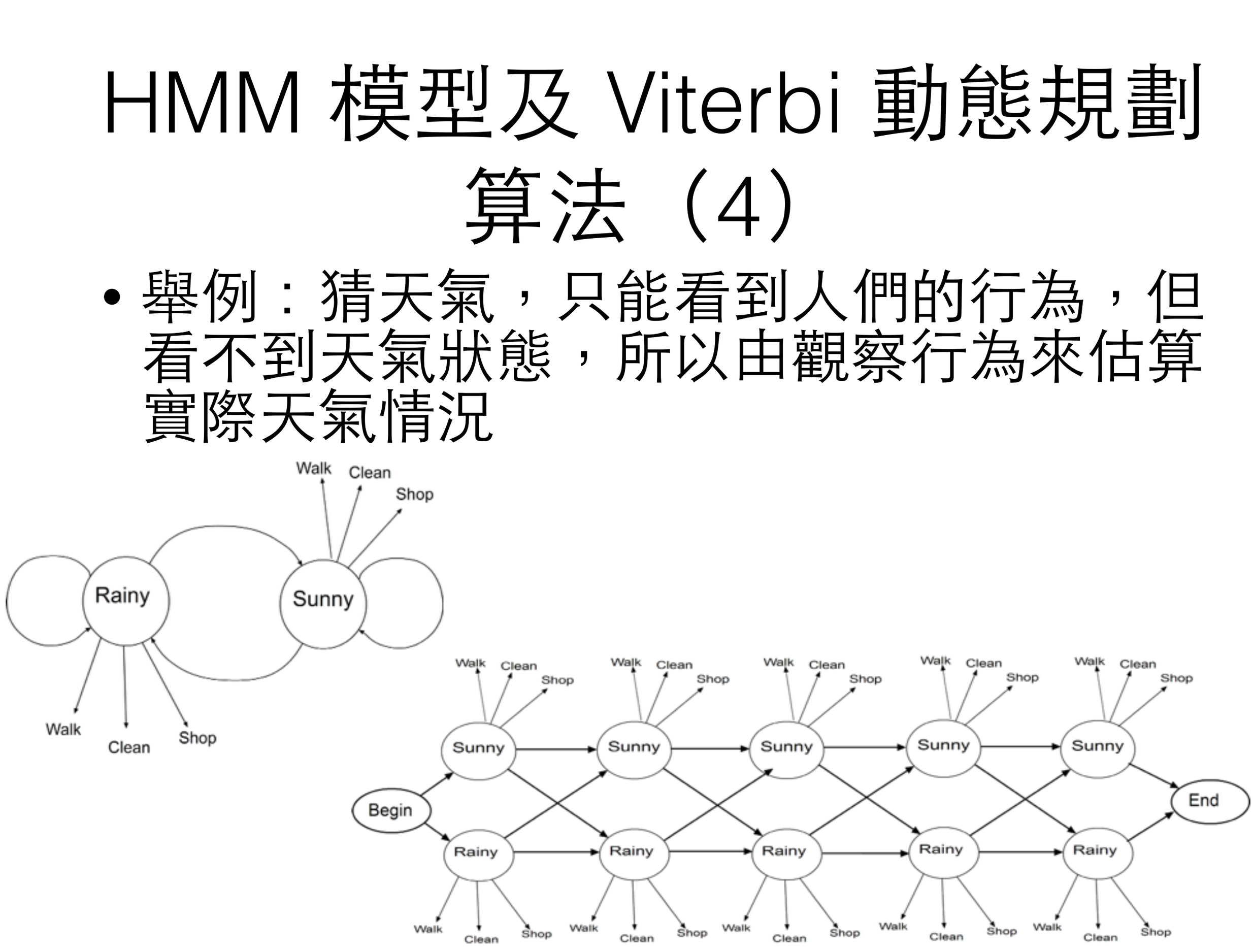
- 其中一條路徑的算法

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



HMM 模型及 Viterbi 動態規劃算法（4）

- 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 動態規劃算法（5）

- 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
- 全台大停電：BESBE

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

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

Jieba 結巴實作

您將在這邊學到：

跳火圈 X
走鋼索 X
如何使用 Jieba O

Python 安裝

- 官方網站：https://www.python.org/downloads/
- Installing Python on Mac OS X（使用 homebrew）
  - http://docs.python-guide.org/en/latest/starting/install/osx/
- Installing Python on Windows
  - http://docs.python-guide.org/en/latest/starting/install/win/

PIP 是什麼

- Python 的套件管理工具
- DRY (Don't Repeat Yourself)

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
```

Jieba 安裝

```
$ [sudo] pip install jieba
```

切換詞庫

```
#encoding=utf-8
import jieba

jieba.set_dictionary("dict.txt.big.txt")
```

斷詞精確模式

```
#encoding=utf-8
import jieba

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

seg_list = jieba.cut("颱風就是要泛舟啊不然要幹嘛")
print(", ".join(seg_list))

seg_list = jieba.cut("先拆坐墊，公道價八萬一，你是在大聲什麼啦")
print(", ".join(seg_list))
```

斷詞精確模式執行結果

颱風，就是，要，泛舟，啊，不然，要，幹，嘛

先，拆，坐墊，，，公道，價，八萬，
一，，，你，是，在，大聲，什麼，啦

斷詞全模式

```
#encoding=utf-8
import jieba

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

seg_list = jieba.cut("颱風就是要泛舟啊不然要幹嘛",
cut_all=True)
print(", ".join(seg_list))

seg_list = jieba.cut("先拆坐墊，公道價八萬一，你是
在大聲什麼啦", cut_all=True)
print(", ".join(seg_list))
```

斷詞全模式執行結果

颱風，就是，要，泛舟，啊，不然，要，幹，嘛

先，拆，坐墊，，，，公道，價，八萬，萬一，，，，你，是，在，大聲，什，麼，

斷詞返回原文的起止位置

```
#encoding=utf-8
import jieba

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

result = jieba.tokenize(u'颱風就是要泛舟啊不然要幹嘛')
for tk in result:
    print("word %s\t\t start: %d \t\t end:%d" % (tk[0],tk[1],tk[2]))
```

斷詞返回原文的起止位置執行結果

```
word 颱風    start: 0     end:2
word 就是    start: 2     end:4
word 要      start: 4     end:5
word 泛舟     start: 5     end:7
word 啊      start: 7     end:8
word 不然     start: 8     end:10
word 要      start: 10     end:11
word 幹      start: 11     end:12
word 嘛      start: 12     end:13
```

詞性標注

```
#encoding=utf-8
import jieba

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

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    v
嘛    y
```

使用實例一

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

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

使用實例：中文歌詞斷詞，使用預設詞庫

```
#encoding=utf-8
import jieba

content = open('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

使用實例：中文歌詞斷詞，使用繁體詞庫

```
#encoding=utf-8
import jieba

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

content = open('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

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」

使用實例：取出文章中的關鍵詞

```
#encoding=utf-8
import jieba

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

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

print "Input：", content

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

使用實例：取出文章中的關鍵詞執行結果

沒有 , 所謂 , 是否 , 一般 , 退縮 , 雖然 , 肌迫 , 矯作 , 來說 , 怯懦

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
- 使用自定義詞典

Jieba 自定義詞典用法

```
#encoding=utf-8
import jieba

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

Jieba 動態新增詞典

```
#encoding=utf-8
import jieba

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

使用實例二

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

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

使用實例：台語歌詞斷詞，使用繁體詞庫

```
#encoding=utf-8
import jieba

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

content = open('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

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」

使用實例：台語歌詞斷詞，使用繁體詞庫加自定義詞庫

```
#encoding=utf-8
import jieba

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

content = open('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

台語歌詞斷詞，使用繁體詞庫加自定義詞庫結果分析

- 完全符合預期結果
- 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
唱著 4 v
每一個 4 m
會使 70 d
```

斷詞運用在音樂

- 歌詞分析
- 情境歌單
- 自動填詞
- 歌詞推薦（創作者或一般使用者）

歌詞情意特徵值

周杰倫 《蝸牛》

情緒

語意

該不該擱下重重的殼　尋找到底哪裡有藍天
隨著輕輕的風輕輕的飄　歷經的傷都不感覺疼

我要一步一步往上爬　等待陽光靜靜看著它的臉
小小的天　有大大的夢想　重重的殼裹著輕輕的仰望

我要一步一步往上爬　在最高點乘著葉片往前飛
任風吹乾　流過的淚和汗　總有一天我有屬於我的天

隱含情緒

隱含語意

Jieba 各種語言版本

- Java https://github.com/huaban/jieba-analysis
- C++ https://github.com/yanyiwu/cppjieba
- Node.JS https://github.com/yanyiwu/nodejieba
- Erlang https://github.com/falood/exjieba
- R https://github.com/qinwf/jiebaR
- iOS https://github.com/yanyiwu/iosjieba
- PHP https://github.com/fukuball/jieba-php -> 歡迎大家加入開發

Q & A

Find Me

Twitter @fukuball
Facebook @fukuball
GitHub @fukuball