

A decorative graphic consisting of various colored circles (blue, green, yellow, orange, red, pink) and dashed lines arranged in a circular pattern around the central text.

# 上市揀材： PTT語料庫介紹

台大語言所 | 廖聿璫 許家誠

## 還記得Session 2嗎？

- ⚠ 若是我們的corpus太大(幾百MB, 甚至到達GB), 進行一次query需耗費大量CPU資源
- ⚠ 想將自己的corpus提供給外界使用

有沒有什麼工具可以幫助我們解決以上問題？

當然有！那就是：



可以借助**搜尋引擎**(如：**BlackLab**)大幅加速query流程  
Ptt Corpus即是建置在BlackLab之上！



1

BlackLab引擎



## 什麼是BlackLab?



BlackLab是一個**免費**、開源的語料庫搜尋引擎

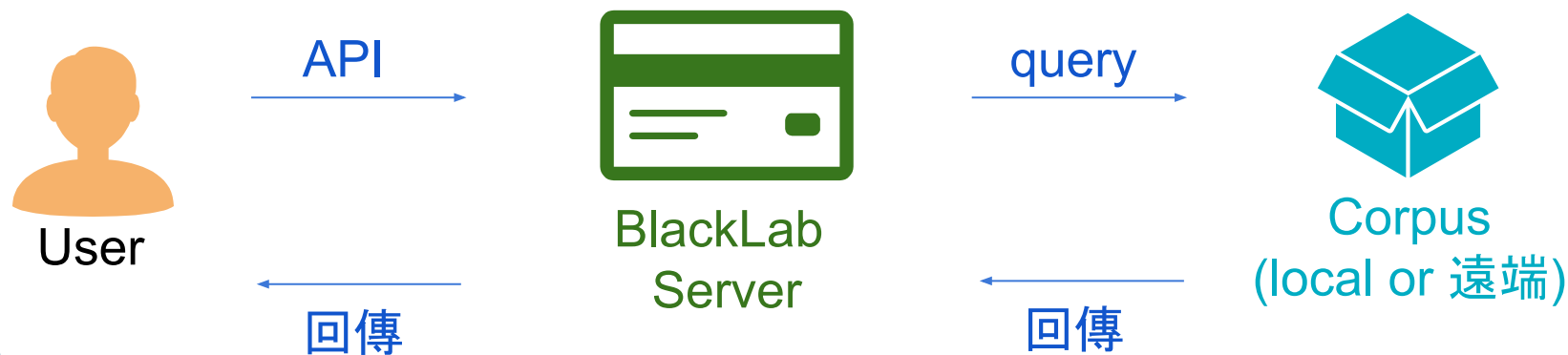


如同ElasticSearch建立在Apache Lucene之上



主要以**Java**寫成

## 透過BlackLab 的query 流程



# BlackLab建置流程

1. 下載BlackLab & Java

3. 產出介面、開始query

2. 準備語料及index

The background is white and decorated with various colorful circles and dashed lines. In the top left, there is a large orange circle with a dashed red outline, overlapping a yellow circle. Below them is a small pink circle. In the top right, there is a green circle with a white center, a small orange circle, and a yellow circle with a dashed green outline. In the bottom left, there is a green circle with a dashed green outline, a large yellow circle, and a small cyan circle. In the bottom right, there is a large cyan circle with a white center, a small cyan circle with a dashed blue outline, and a small cyan circle. In the center, there is a large dashed blue circle containing the number 2.

2

什麼是ptt corpus?



Ptt Corpus 是可供使用者進行**語料庫搜索**的網路UI



語境共現(Concordance)

- ✓ 指定看板
- ✓ 指定年份
- ✓ 詞性顯示
- ✓ CQL查詢

# Concordance

Query

台灣

---

可使用CQL查詢或一般查詢

☐ 使用CQL模式

找哪個版

BabyMother

---

搜尋對象

☐ 顯示詞性

我會去店面買好一點的如果是正版	台灣	代理商進口的,可以請賣家拍商標
,百貨公司買吧!80肯定假的啊	台灣	買貝親超簡單型的固齒器都要120了麵包
方面疑問的人。因為這個疾病在	台灣	能找到的資訊真的太少了。如果未來
的圍欄要6xxx(含運)	台灣	蝦皮只要4xxx(免運)想請問
網路上都說g市買會比較便宜	台灣	蝦皮G市感覺不錯,我再來找找
4x2加上12片圍欄含運送到	台灣	是7900NTD最近g市有活動,可以
沒有看過外國媽媽養小孩之後就覺得	台灣	媽媽也太辛苦了XD D機洗,奶瓶
做家務一個月10萬我覺得阿姨回來	台灣	跑foodpanda還比較好就像夫妻兩
!都在美國工作生活了,還把	台灣	那種請人的模樣帶過去~還問
早晚班可能更合適?(但如果都從	台灣	請,還得處理住宿問題就是人家還要

## 一般查詢

Query

旋轉

---

Query

旋轉 我

---

## CQL查詢

Query

[word='旋轉']

---

Query

[word='旋轉' & pos='N.\*']

---

## 我可以用Ptt Corpus做到哪些事情？



找出特定詞彙在哪些**情境(句法、語意、語用)**中常出現



可搭配**regex**, **CQL**等語法綜合分析



將語境**量化**後深入分析



產出TF-IDF, 詞向量, ...[Session 5]



透過**classification**、**clustering**等方法找出不同語境間相似度／包含哪些主題...



## 量化語意分析

退休 財富 自由 賠錢 的 也 一 堆 ， 不 懂 玩 ， 不 懂 自 律 ， 乾 脆 不 要 玩



[0.5, 0.3, -0.2, 1.1, 0.7, ...]



[0.9, -0.4, 0.9, -1.4, 0.6, ...]

Cosine similarity, clustering...



[0.6, 0.7, 2.5, -0.2, 0.8, ...]

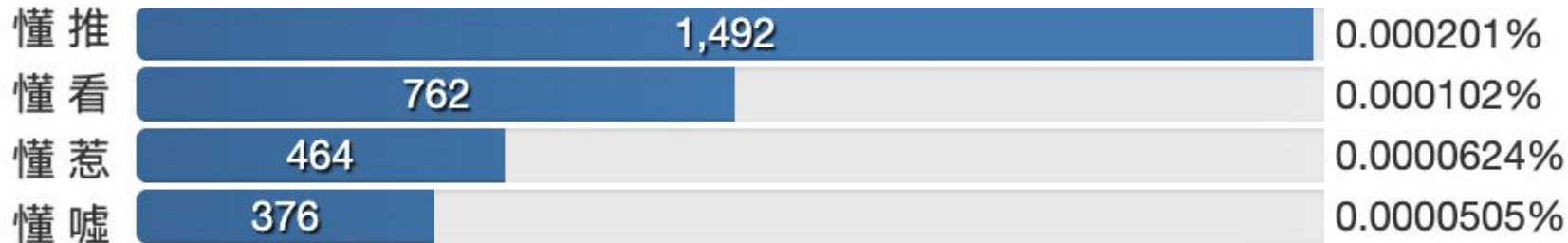


[0.2, -0.8, -1.9, 0.3, 0.2, ...]

雲林 小吃 也 很 強 ， 真 的 一 堆 不 懂 玩 桃園 啊 公 園 都 可 以 當 景 點

## 有辦法透過UI就得到統計嗎？

例如：



輸入[懂V.], 統計所有[懂V.]數量 & 出現頻率



## 還記得BlackLab嗎？

BlackLab所提供的前端(Autosearch)提供了各種功能：

- ✓ 基本 & 進階搜索(word, pos, CQL)
- ✓ N-gram搜索
- ✓ 詞頻、構式統計
- ✓ 語境統計

詳細建置方法請參照：<https://github.com/INL/corpus-frontend/>

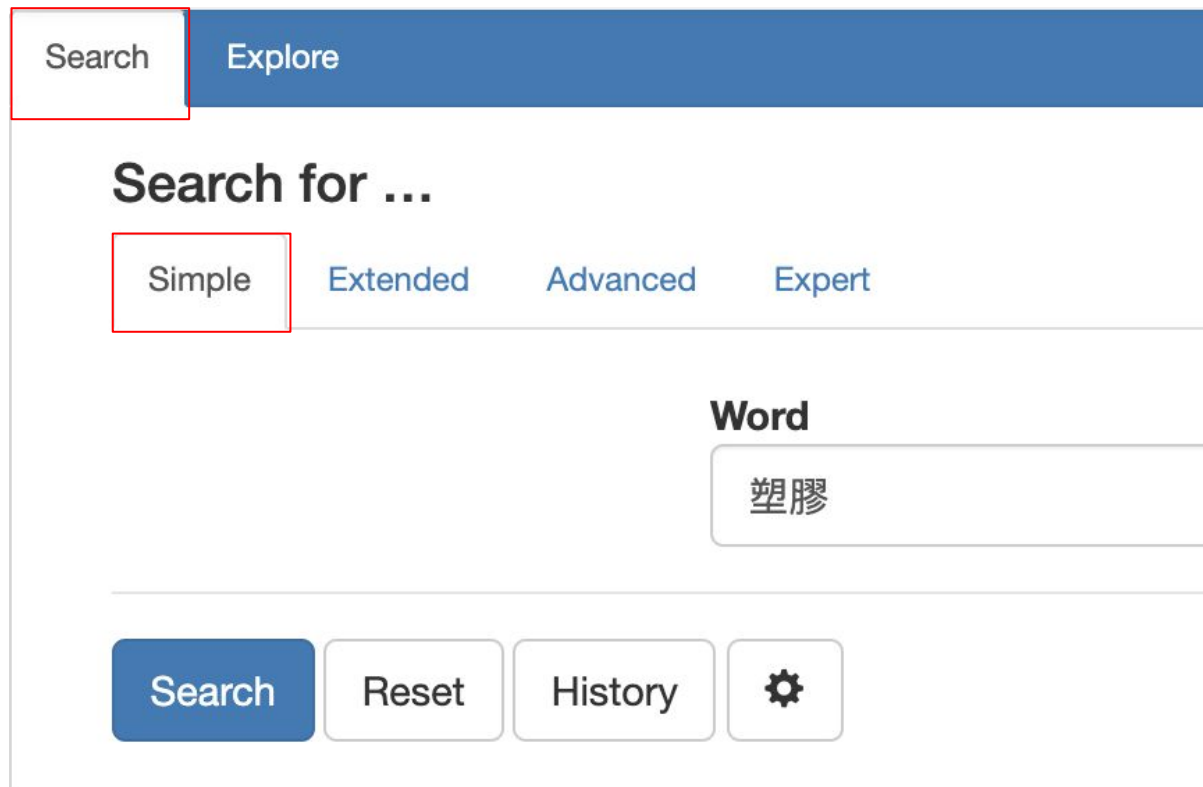




## Ptt Corpus開發版

: <http://140.112.147.132:8999/corpus-frontend>

你應該會先看到：



The image shows a search interface with a blue header bar containing 'Search' and 'Explore' tabs. Below the header, the text 'Search for ...' is displayed. Underneath, there are four tabs: 'Simple', 'Extended', 'Advanced', and 'Expert'. The 'Simple' tab is highlighted with a red box. Below the tabs is a text input field labeled 'Word' containing the Chinese characters '塑膠'. At the bottom, there are four buttons: 'Search' (blue), 'Reset', 'History', and a settings gear icon. The 'Search' button is also highlighted with a red box.

Search Explore

Search for ...

Simple Extended Advanced Expert

Word

塑膠

Search Reset History ⚙️

## 執行結果：



Per Hit

Per Document

Hits / Grouped by hit:word

Group by Word



Case sensitive

這是什麼？

Hit

Group by Word

Group by PoS

Before hit

Group by Word before

Group by PoS before

After hit

Group by Word after

Group by PoS after

Metadata

Group by Author (Metadata)

Group by Board (Metadata)

Group by Doc id (Metadata)

Group by From input file (Metadata)

Group by Year (Metadata)

Before hit

Group by Word before

「你 當 北檢 是

3

0.00000975%

Before	Hit	After
... 「你 當 北檢 是	塑膠	嗎？」 「你 ...
... 「你 當 北檢 是	塑膠	嗎？」 「你 ...
... 「你 當 北檢 是	塑膠	嗎？」 霸氣一點 ...

。關於 胡采蘋 的 財務  
講 不 出來 當 人民  
投 第三 勢力 的人  
就 很 真的 當 台灣人  
想 利用 完 後 當  
會 倒」 這種  
人 翻白眼的 海量 財務  
不 起來 " 當 賴清德

2

0.0000065%

2

0.0000065%

2

0.0000065%

2

0.0000065%

2

0.0000065%

2

0.0000065%

2

0.0000065%

2

0.0000065%

# Extended可指定pos

## Search for ...

Simple

Extended

Advanced

Expert

Word

旋轉

☐ Case and diacritics sensitive

PoS

N.\*

## Filter search by ...

Author

Author

Board

Board

Doc id

Doc id

From input file

From input file

Year

From

To

# 很好用的Advanced:

Simple Extended **Advanced** Expert

token

token

[ word = "" ]

search options

Word =

☐ Case and diacritics sensitive

+

+

AND

PoS =

# 「被消失」

↔ [ word = "被" ] ×

search options

× Word ▾ = 被 ⬆ ⬇ +

☐ Case and diacritics sensitive

+

↔ [ word = "消失" ] ×

search options

× Word ▾ = 消失 ⬆ ⬇ +

☐ Case and diacritics sensitive

+



練習:[被「？」消失]、[被「不及物V.」]

小提示:名詞類= N.\*, 動詞類= V.\*, 形容詞類= A.\*

更詳細pos tag請參考 <https://ckip.iis.sinica.edu.tw/CKIP/paper/poslist.pdf>



被「？」消失

↔ [ word = "被" ] ×

search options

× Word ▾ = 被 ⬆ ⬆ +

☐ Case and diacritics sensitive

+

↔ [ ] ×

search options

+

也可指定pos

↔ [ word = "消失" ] ×

search options

× Word ▾ = 消失 ⬆ ⬆ +

☐ Case and diacritics sensitive

+

被「不及物V.」

↔ [ word = "被" ] ×

search options

× Word ▾ = 被 ⬆ ⬆ +

☐ Case and diacritics sensitive

+

↔ [ pos = "VA" ] ×

search options

× PoS ▾ = VA ⬆ ⬆ +

+



## 挑戰：找出 Middle voice (關身語態)

例如：油加好了/ 澱粉吃多了會胖/ 問題解決了  
水滴下來/ 煙火爆炸了/ 電視壞掉了

小提示：名詞類= N.\*，動詞類= V.\*，形容詞類= A.\*

## 問題解決了/ 小孩生了

↔ [ pos = "Na" ] ×

search options

× PoS ▾ = Na ⬆ ⬆ +

+ +

↔ [ pos = "VC" ] ×

search options

× PoS ▾ = VC ⬆ ⬆ +

+ + **VA?**

↔ [ word = "了" ] ×

search options

× Word ▾ = 了 ⬆ ⬆ +

☐ Case and diacritics sensitive

+ +

## 酒喝多了/ 電影看太多了

↔ [ pos = "Na" ] ×

search options

× PoS ▾ = Na ⬆ ⬆ +

+ +

↔ [ pos = "VC" ] ×

search options

× PoS ▾ = VC ⬆ ⬆ +

+ +

↔ [ ] ×

search options

+ +

↔ [ word = "了" ] ×

search options

× Word ▾ = 了 ⬆ ⬆ +

☐ Case and diacritics sensitive

+ +

Search for ...

Simple

Extended

Advanced

Expert

Corpus Query Language:

這是什麼？



3

CQL

## CQL → Corpus Query Language



專門用來檢索語料庫的語法

The IMS Open Corpus Workbench



源自於語料庫工具網站(CWB)



很多語料庫都支援這種檢索語法

e.g. Sketch Engine、**BlackLab** (PTT語料庫)、  
國教院語料庫索引典系統

## CQL能做什麼？

word lemma part of speech



可以根據詞(條)、詞性等條件來檢索



一次精準檢索出想要的語料

## CQL語法——基本型 → [word = " " ]



以BlackLab介面所支援的語法為主



[word = "豬油"]、[pos = "Na"]



一個中括弧表示一個詞



[word = "豬油"] ≡ "豬油"



## 寫在" "裡——Regular Expression Supported

 .

→ 任一字元

e.g. "中.人" → 中國人、中間人、中年人、.....

 [ ]

→ character set 中擇一字元

e.g. "[英菸韓柯]粉" → 英粉、菸粉、韓粉、柯粉

"[a-c5-9]" → a、b、c、5、6、7、8、9

"[^米]粉" → 綠粉、憨粉、昌粉、冬粉、.....

## 寫在" "裡——Regular Expression Supported

Quantifier 數量標記 → 作用於前一個物件



?

→ 0或1

e.g. "老年?人" → 老人、老年人



\*

→ 0個以上

e.g. "1[0-9]\*" → 1、1450、165、1111、.....



+

→ 1個以上

e.g. "[東南西北]+方" → 東方、西南方、東北東方、.....

## 寫在" "裡——Regular Expression Supported

Quantifier 數量標記 → 作用於前一個物件



**{n}**

→ 剛好n個

e.g. "G{5}" → GGGGG



**{n,m}**

→ n到m個

e.g. "K.{2,4}" → KKC、KTV、KOBЕ、KMTER、.....

## 寫在" "裡——Regular Expression Supported



練習：請找出「台北」、「台北人」、「台中」、「台中人」、「台南」、「台南人」

## 寫在" "裡——Regular Expression Supported



()

→ grouping

e.g. "(XD){3}" → XD<sup>3</sup>XD

cf. "XD{3}" → XD<sup>3</sup>DD



|

→ 或; 作用於整個括弧／引號

e.g. "(台北|高雄)人" → 台北人、高雄人

cf. "(台[北高]雄)人" → 台北人、台<sup>3</sup>雄人

## 寫在" "裡——Regular Expression Supported



→ 跳脫字元作用於後一個字元

e.g. "\+" → +

cf. "+" → + ( ? ? ? ? )

基本上還是加上跳脫字元比較保險

需要加上跳脫字元的符號: . ? \* + | ( ) [ ] { } ^ \$ ' "

# 寫在" "裡——~~Regular Expression Supported~~



(?-i)

→ 大小寫鎖定；作用於整個引號

e.g. "cd" → CD、Cd、cD、cd

"(?-i)cd" → cd

## 寫在[ ]裡——邏輯符號

**&**

→ and 且; 用於連接前一個及後一個物件

e.g. [word="反應" & pos="V.\*"]

**|**

→ or 或; 用於連接前一個及後一個物件

e.g. [word="反應" | word="反映"]

**!**

→ not 非; 作用於後一個物件

e.g. [word="反應" & !pos="V.\*"]




## 寫在[ ]裡——邏輯符號



練習：請找出所有有「婚」字的名詞

## 寫在[ ]外

 Quantifier 記量符號 `?` `*` `+` `{ }`

 集合符號 `&` `|`

e.g. "台灣" "安全" `|` "人民" "有錢"

"綠色" "執政" `&` `[pos="Na"]` `[pos="Va"]`

$\equiv$  `[word="綠色" & pos="Na"]` `[word="執政" & pos="Va"]`

## 寫在[ ]外



[ ]

→ 任何token

e.g. [word="扛"] [ ] [word="住"] → 扛 得 住、扛 不 住、  
扛 的 住、.....



A:

→ anchor

e.g. WHO:[pos="Nh"] "GG" "了"



::

→ Global constraints

e.g. A:[ ] "東" B:[ ] "西" :: A.word = B.word

## 寫在[ ]外——XML語法



`<s>`

→ 句子開頭

`</s>`

→ 句子結尾

`<s/>`

→ 整個句子



`containing` → 前項條件中需包含後項

e.g. `<s/> containing "政治" "巨嬰"`



`within`

→ 前項條件需包含在後項之中

e.g. `"政治" "巨嬰" within <s/>`

## 寫在[ ]外——XML語法



練習：請找出含有「台女」的 NN- 複合詞

The background is white and decorated with various colorful circles and dashed lines. In the top left, there is a large orange circle with a dashed red outline, overlapping a solid yellow circle. Below them is a small pink circle. In the bottom left, there is a large green circle with a dashed green outline, a small cyan circle, and a large yellow circle. In the top right, there is a large green circle with a white dot in the center, a small orange circle, and a yellow circle with a dashed yellow outline. In the bottom right, there is a large cyan circle with a white dot in the center, and a cyan circle with a dashed cyan outline. A large, faint dashed blue circle is centered in the background.

# Q&A