小農手作：
語料庫索引與建置

第二屆語料庫程式實務工作坊(HOCOR)
2020年 12月 12日 下午場

講者：洪漢唐、江琼玉

本場次目標(農場標題式)

- 在三個小時內使用Python 3 做出媲美 BNC , COCA 等級、能查詢500萬詞的語料庫搜尋功能。

本場次目標(務實版)

- 用Python 3刻一個語料庫
  - 介紹索引 (Indexing) 概念
  - 手作索引
  - 手作 Concordance 和客製化查詢
  - 手作 Collocation

Coding 環境

- Google Colaboratory 上的 Python 3
- 預設Python知識 : 資料結構(list, dict)、條件判斷、迴圈、函式
- 會使用到的外部套件
    - numpy (計算collocation的association measure)
    - pandas (主要使用DataFrame來呈現搜尋結果)

什麼是Index(ing)?

平常會遇到的問題

- 如何去圖書館找到索緒爾的《普通語言學概論》？
- 如何從一本書找到 ”語法樹”出現的地方？
- 如何從雜亂的電腦桌面找到上次臨時存的筆記檔？
- 你怎麼知道台大圖書館的網站在哪裡？
- 如何找到朋友推薦你的咖啡店？

平常會遇到的問題

Index / Locator

- 如何去圖書館找到索緒爾的《普通語言學概論》？-------------> 圖書編碼
- 如何從一本書找到 ”語法樹”出現的地方？---------------------------> 書末索引
- 如何從雜亂的電腦桌面找到上次臨時存的筆記檔？-------------> 資料夾
- 你怎麼知道台大圖書館的網站在哪裡？-----------------------------> 網址
- 如何找到朋友推薦你的咖啡店？--> 地址

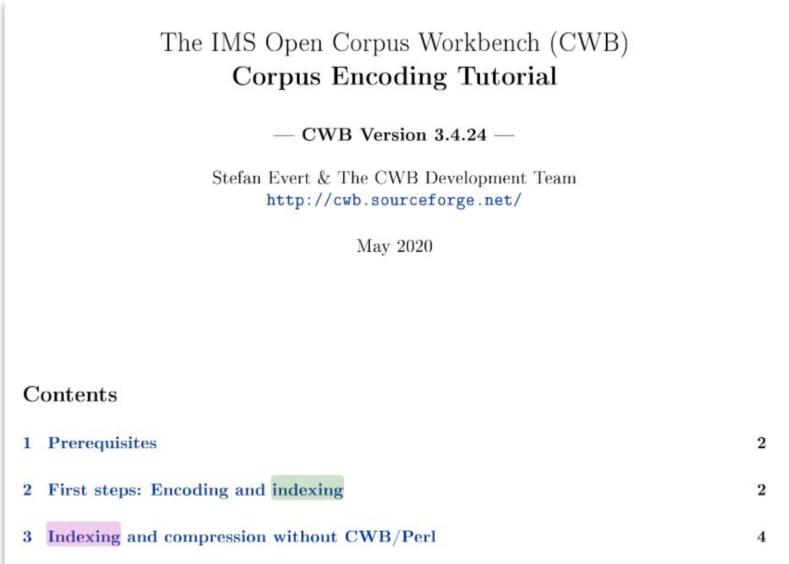
index.html

共通點：
用”空間”和”事前的工”
換取時間

語料庫當然也不例外

系統	Index
搜尋引擎	某一個網站
書籍	某一頁
語料庫	某一文檔的某一句的第幾個詞

CWB和BlackLab官方文件都有Indexing的步驟



The IMS Open Corpus Workbench (CWB)
Corpus Encoding Tutorial

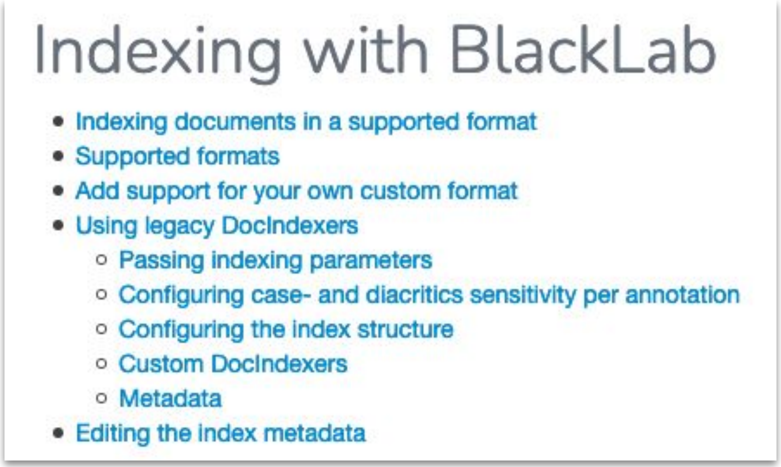
— **CWB Version 3.4.24** —

Stefan Evert & The CWB Development Team
http://cwb.sourceforge.net/

May 2020

Contents

http://cwb.sourceforge.net/files/CWB_Encoding_Tutorial.pdf

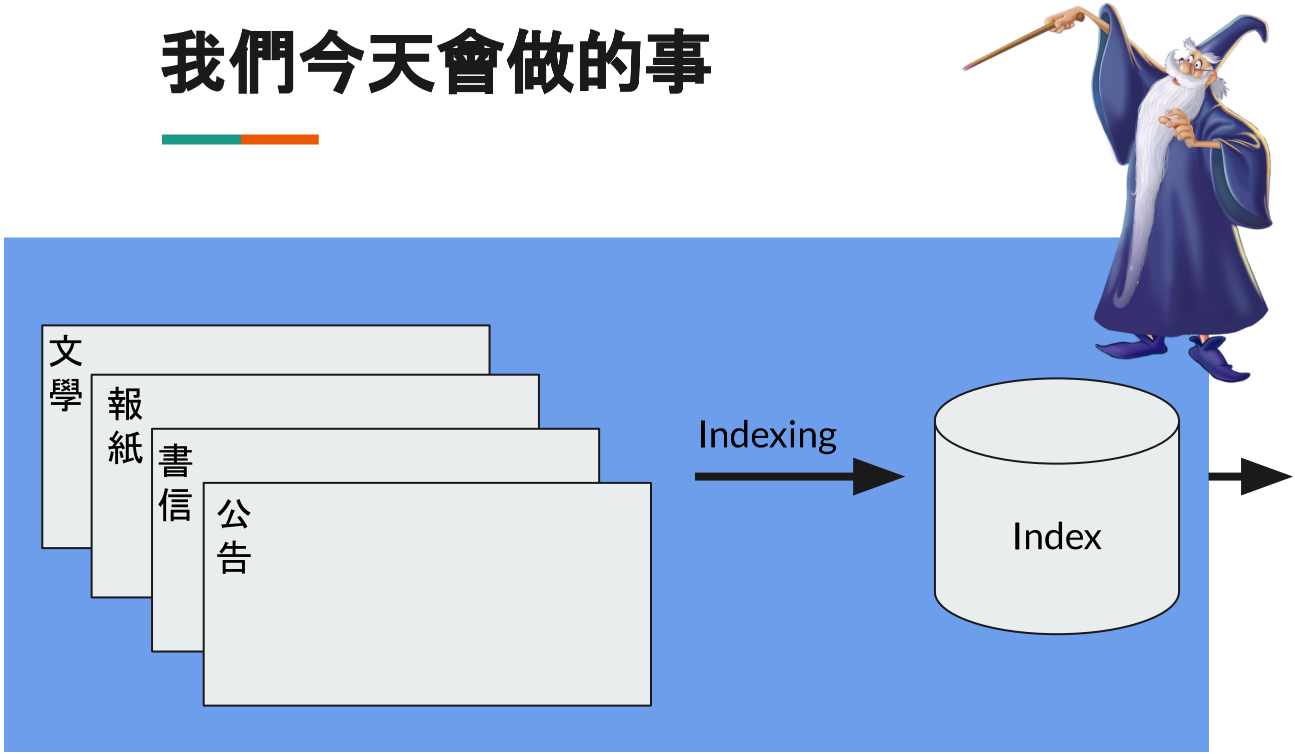


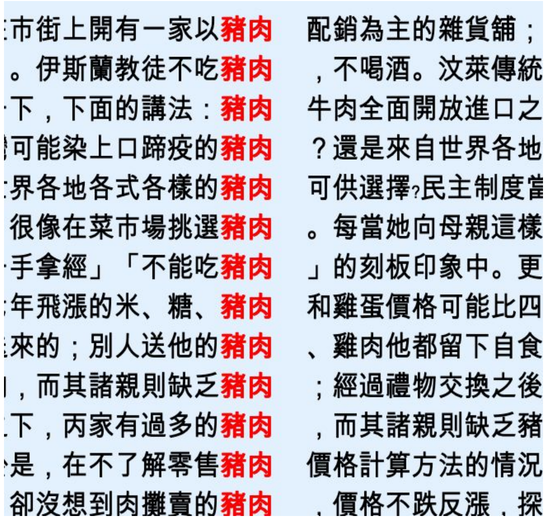
Indexing with BlackLab

- Indexing documents in a supported format
- Supported formats
- Add support for your own custom format
- Using legacy DocIndexers
  - Passing indexing parameters
  - Configuring case- and diacritics sensitivity per annotation
  - Configuring the index structure
  - Custom DocIndexers
  - Metadata
- Editing the index metadata

http://inl.github.io/BlackLab/indexing-with-blacklab.html

我們今天會做的事



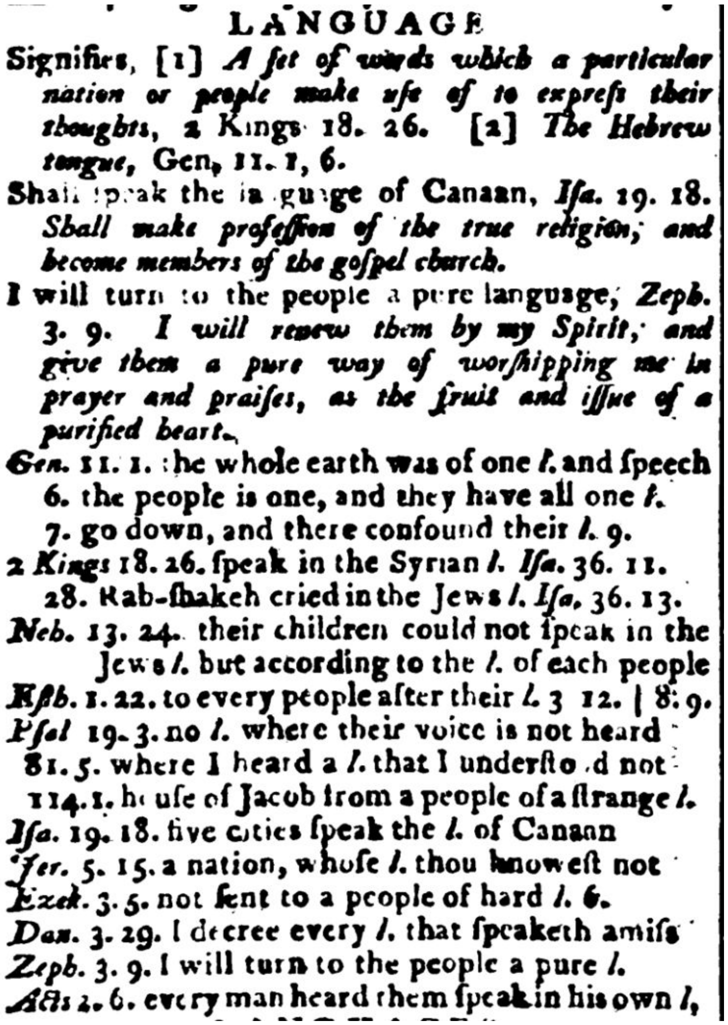


```
市街上開有一家以豬肉　配銷為主的雜貨舖；
。伊斯蘭教徒不吃豬肉　，不喝酒。汶萊傳統
下，下面的講法：豬肉　牛肉全面開放進口之
可能染上口蹄疫的豬肉　？還是來自世界各地
界各地各式各樣的豬肉　可供選擇?民主制度
很像在菜市場挑選豬肉　。每當她向母親這樣
手拿經」「不能吃豬肉　」的刻板印象中。更
年飛漲的米、糖、豬肉　和雞蛋價格可能比四
來的；別人送他的豬肉　、雞肉他都留下自食
，而其諸親則缺乏豬肉　；經過禮物交換之後
下，丙家有過多的豬肉　，而其諸親則缺乏豬
是，在不了解零售豬肉　價格計算方法的情況
卻沒想到肉攤賣的豬肉　，價格不跌反漲，探
```

Concordance line

Concordance
也是一種 index

Cruden, Alexander. (1817). *A Complete Concordance to the Holy Scriptures of the Old and New Testament; Or, A Dictionary and Alphabetical Index to the Bible.* link



LANGUAGE

Signifies, [1] *A ſet of words which a particular nation or people make uſe of to expreſs their thoughts,* 2 Kings 18. 26. [2] *The Hebrew tongue,* Gen, 11. 1, 6.

Shall ſpeak the language of Canaan, *Iſa.* 19. 18. *Shall make profeſſion of the true religion; and become members of the goſpel church.*

I will turn to the people a pure language, *Zeph.* 3. 9. *I will renew them by my Spirit, and give them a pure way of worſhipping me in prayer and praiſes, as the fruit and iſſue of a purified heart.*

Gen. 11. 1. the whole earth was of one *l.* and ſpeech
6. the people is one, and they have all one *l.*
7. go down, and there confound their *l.* 9.

2 *Kings* 18. 26. ſpeak in the Syrian *l.* *Iſa.* 36. 11.
28. Rab-ſhakeh cried in the Jews *l.* *Iſa.* 36. 13.

Neh. 13. 24. their children could not ſpeak in the Jews *l.* but according to the *l.* of each people

Eſth. 1. 22. to every people after their *l.* 3 12. | 8. 9.

Pſal 19. 3. no *l.* where their voice is not heard
81. 5. where I heard a *l.* that I underſtood not
114. 1. houſe of Jacob from a people of a ſtrange *l.*

Iſa. 19. 18. five cities ſpeak the *l.* of Canaan

Jer. 5. 15. a nation, whoſe *l.* thou knoweſt not

Ezek. 3. 5. not ſent to a people of hard *l.* 6.

Dan. 3. 29. I decree every *l.* that ſpeaketh amiſs

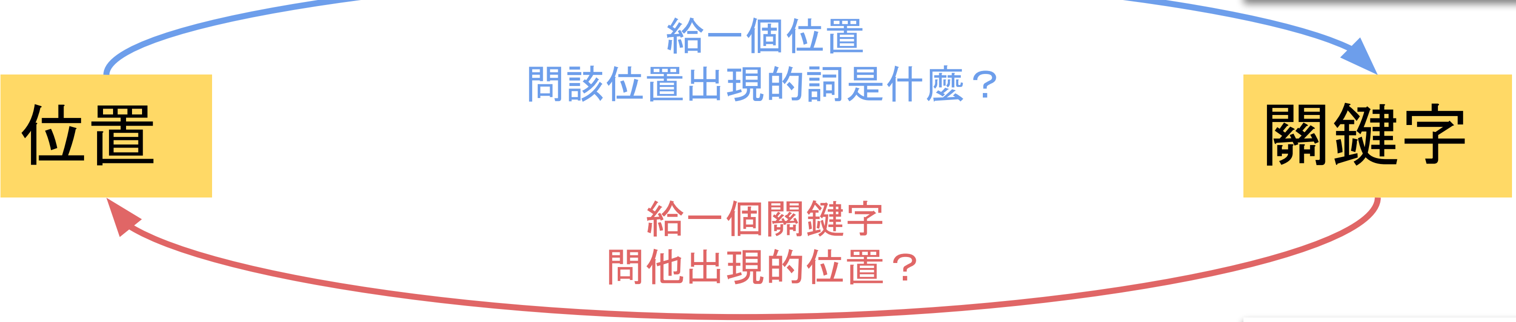
Zeph. 3. 9. I will turn to the people a pure *l.*

Acts 2. 6. every man heard them ſpeak in his own *l,* “The term concordance originally applied just to biblical indexes. Its root, *concord*, means unity, and the term apparently arose out of the school of thought that the unity of the Bible should be reflected in consistency between the Old and New Testaments, which could be demonstrated by a concordance.”

— see Witten et al. (1999) *Managing gigabytes: compressing and indexing documents and images. (p. 1).*

Forward vs. Inverted index 正向 vs. 倒排索引

需要一個
Forward Index



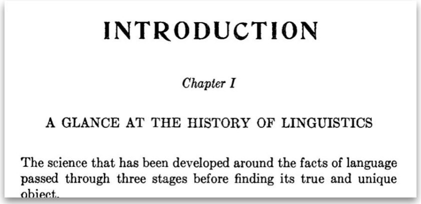
位置

給一個位置
問該位置出現的詞是什麼？

關鍵字

給一個關鍵字
問他出現的位置？

需要一個
Inverted Index

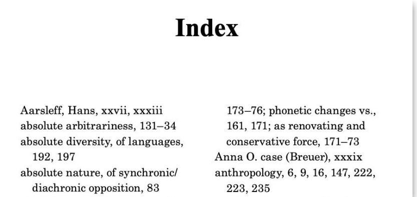


INTRODUCTION

Chapter I

A GLANCE AT THE HISTORY OF LINGUISTICS

The science that has been developed around the facts of language passed through three stages before finding its true and unique object.



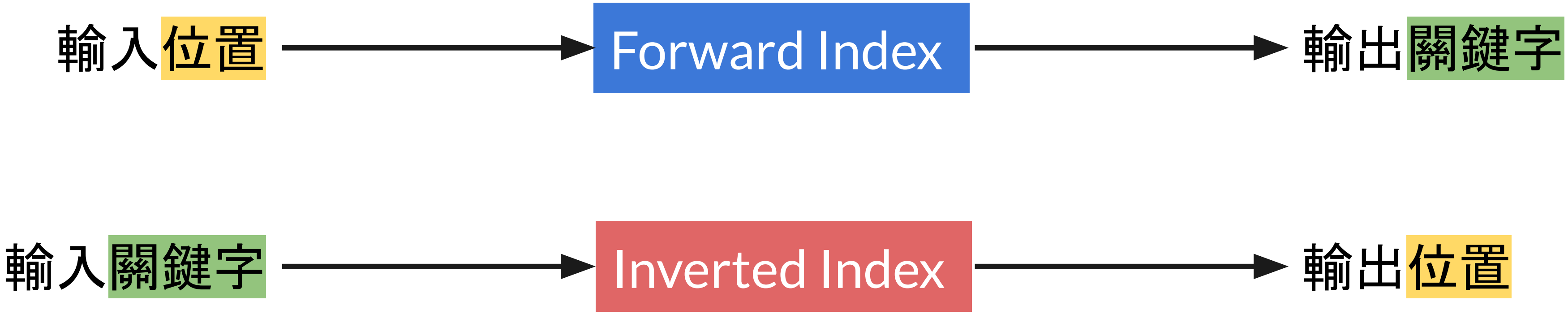
Index

Aarsleff, Hans, xxvii, xxxiii
absolute arbitrariness, 131–34
absolute diversity, of languages, 192, 197
absolute nature, of synchronic/diachronic opposition, 83

173–76; phonetic changes vs., 161, 171; as renovating and conservative force, 171–73
Anna O. case (Breuer), xxxix
anthropology, 6, 9, 16, 147, 222, 223, 235

Difference between Inverted Index and Forward Index. https://www.geeksforgeeks.org/difference-inverted-index-forward-index/

Forward vs. Inverted index 正向 vs. 倒排索引



輸入位置 → Forward Index → 輸出關鍵字

輸入關鍵字 → Inverted Index → 輸出位置

Difference between Inverted Index and Forward Index. https://www.geeksforgeeks.org/difference-inverted-index-forward-index/

如何產生一個反向索引 (Inverted index)？

[Doc 0]
[S0] 天氣 很 好
[S1] 奇怪 的 天氣

[Doc 1]
[S0] 好 的 天氣
[S1] 好 奇怪

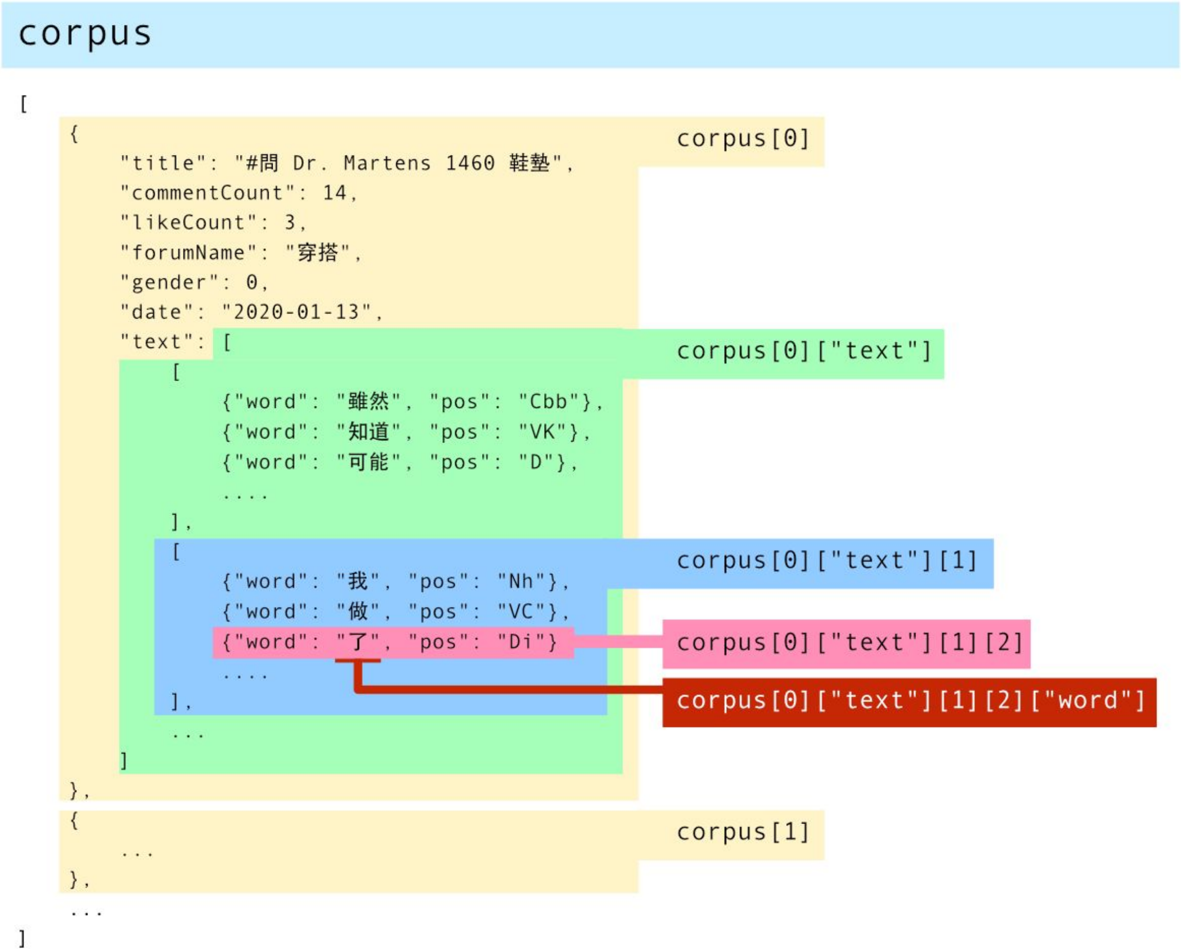
lexicon	postings
天氣	[<D0, S0, T0>, <D0, S1, T2>, <D1, S0, T2>]
很	[<D0, S0, T1>]
好	[<D0, S0, T2>, <D1, S0, T0>, <D1, S1, T0>]
奇怪	[<D0, S1, T0>, <D1, S1, T1>]
的	[<D0, S1, T1>, <D1, S0, T1>]

開始寫程式吧

- 練習版 Colab notebook
  - bit.ly/corpus_indexing_workbook
- 參考解答版 Colab notebook
  - bit.ly/corpus_indexing_solution
- 注意事項
  - 請使用 Firefox 或 Chrome 瀏覽器
  - 開啟後請先執行「檔案 > 在雲端硬碟中儲存副本」，才能儲存自己所做的修改

程式圖解

Colab notebook
Step 2



corpus

```
[
    {
        "title": "#問 Dr. Martens 1460 鞋墊",
        "commentCount": 14,
        "likeCount": 3,
        "forumName": "穿搭",
        "gender": 0,
        "date": "2020-01-13",
        "text": [
            [
                {"word": "雖然", "pos": "Cbb"},
                {"word": "知道", "pos": "VK"},
                {"word": "可能", "pos": "D"},
                ....
            ],
            [
                {"word": "我", "pos": "Nh"},
                {"word": "做", "pos": "VC"},
                {"word": "了", "pos": "Di"}
                ....
            ],
            ...
        ]
    },
    {
        ...
    },
    ...
]
```

corpus[0]

corpus[0]["text"]

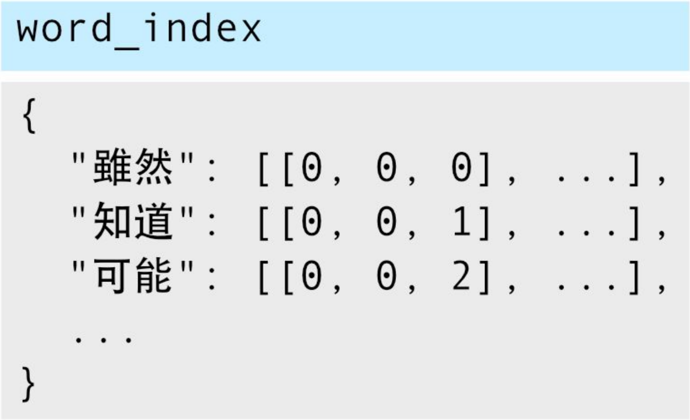
corpus[0]["text"][1]

corpus[0]["text"][1][2]

corpus[0]["text"][1][2]["word"]

corpus[1]

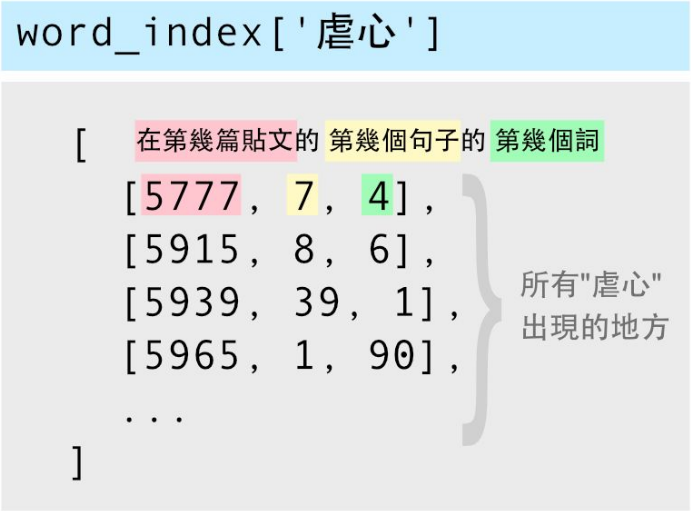
Colab notebook
Step 4-1



```
word_index
```

```
{
   "雖然": [[0, 0, 0], ...],
   "知道": [[0, 0, 1], ...],
   "可能": [[0, 0, 2], ...],
   ...
}
```

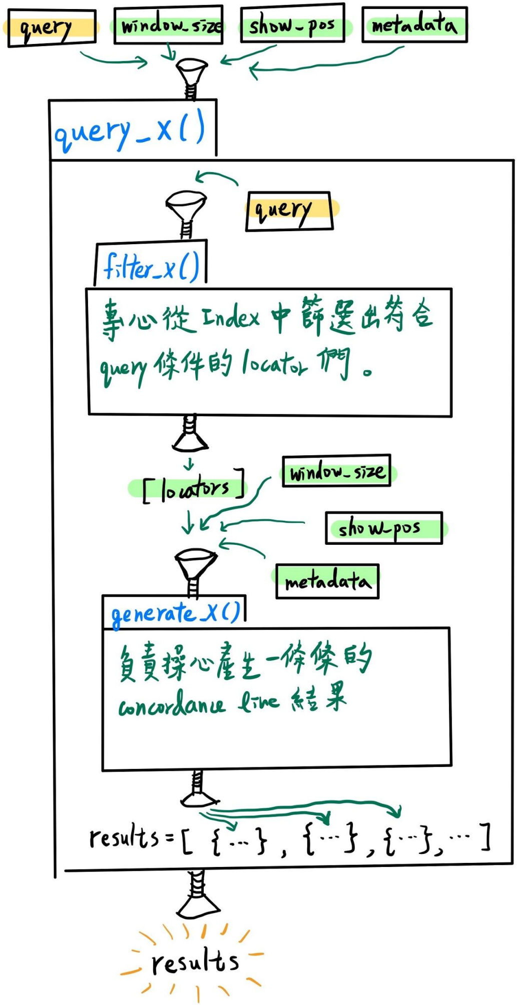
Colab notebook
Step 4-3



```
word_index['虐心']
```

```
[   在第幾篇貼文的 第幾個句子的 第幾個詞
   [5777, 7, 4],
   [5915, 8, 6],
   [5939, 39, 1],
   [5965, 1, 90],
   ...
]
```

Colab notebook
Step 5-3


query
window_size
show_pos
metadata
query_x()
query
filter_x()
專心從 Index 中篩選出符合 query 條件的 locator 們。
[locators]
window_size
show_pos
metadata
generate_x()
負責操心產生一條條的 concordance line 結果
results = [{…}, {…}, {…}, …]
results

之後可以繼續學習的方向

- 文章和書籍
  - 廖永賦, 以 Python 實作 Concordancer
  - Witten, I. H. et al. (1999). *Managing gigabytes: compressing and indexing documents and images*. Morgan Kaufmann.
    - CWB的架構來自本書所介紹的 Indexing方式
  - Manning, C. D. et al. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
  - Zobel, J. & Moffat, A. (2006). *Inverted Files for Text Search Engines*. *ACM Computing Surveys 38(2).*
  - 山田浩之 & 末永匡. (2014). *検索エンジン自作入門: 手を動かしながら見渡す検索の舞台裏* . 技術評論社. [簡中翻譯版: *自製搜尋引擎* (譯者: 胡屹, 人民郵電出版社)]
    - 第一章有很簡要清楚的介紹。
- 閱讀原始碼
  - Python上的indexing套件: whoosh, acora
  - 讀語料庫系統的原始碼: Corpus Workbench, BlackLab