

對**關鍵詞**的幻想

字典派的遺緒

我們來看看坊間傳說的「真實」文字雲吧！

透過i-Buzz產業語意分析詞庫的文字雲



由i-Buzz文字雲分析結果可發現關鍵字都是與汽車產業相關，可觀察網友對於「原廠」、「業務」與「價格」討論度相當高，也可從中看到消費者對於Toyota車款的「空間」、「引擎」與「操控」表現也是消費者的討論重點。

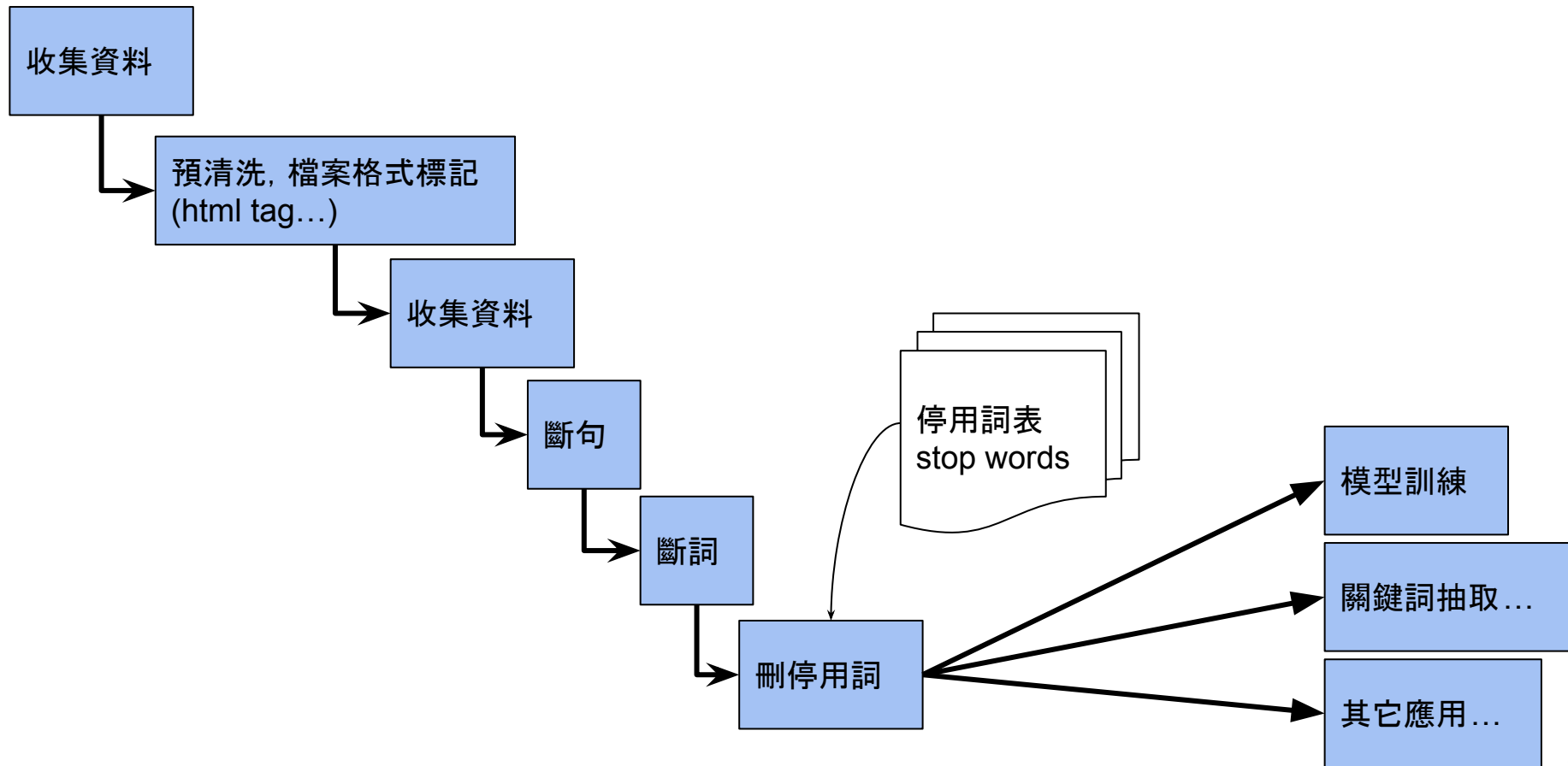
<http://industryresearch.i-buzz.com.tw/文章案例/the-editor-2>

<https://netizen.com.tw/example/>



<https://dailyview.tw/top100/topic/46>

標準的 NLP 流程



TF-IDF

依照人類的習慣，
遇到陌生事物時的第一反應是...

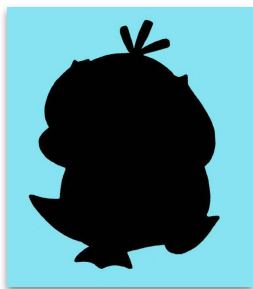


TF-IDF 是兩個東西

TF_(term frequency): 和某個特徵在個體 (一個文本) 中出現的次數有關

IDF_(Inverse Document Frequency): 和這個特徵在群體 (一堆文本) 中出現的次數有關

以神奇寶貝來舉例, 假設我們手邊有這三種神奇寶貝 ...



呱

二足站立



種子種子

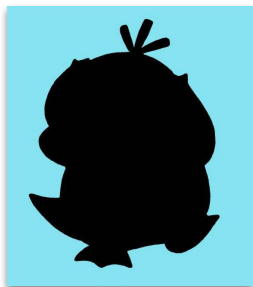
四足站立



皮卡/啾
二足站立
四足站立

個體特徵:

神奇寶貝的 TF-IDF



呱
二足站立



種子種子
四足站立



皮卡/啾
二足站立
四足站立



好討厭的感覺
二足站立

個體特徵:

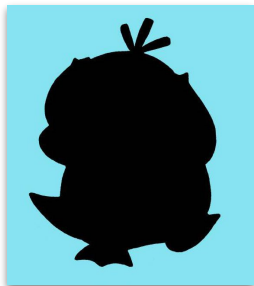
和群體相比的特徵: 呱

種子

皮卡/啾

好討厭的感覺

神奇寶貝 (文本) TF-IDF 特徵:



(0.23104906018664842, '呱'),
(0.0, '腳'),
(0.0, '聲音'),
(0.0, '站'),
(0.0, '發出'),
(0.0, '用')],



(0.39608410317711157, '種子'),
(0.39608410317711157, '種子'),
(0.0, '腳'),
(0.0, '聲音'),
(0.0, '站'),
(0.0, '發出'),
(0.0, '用')

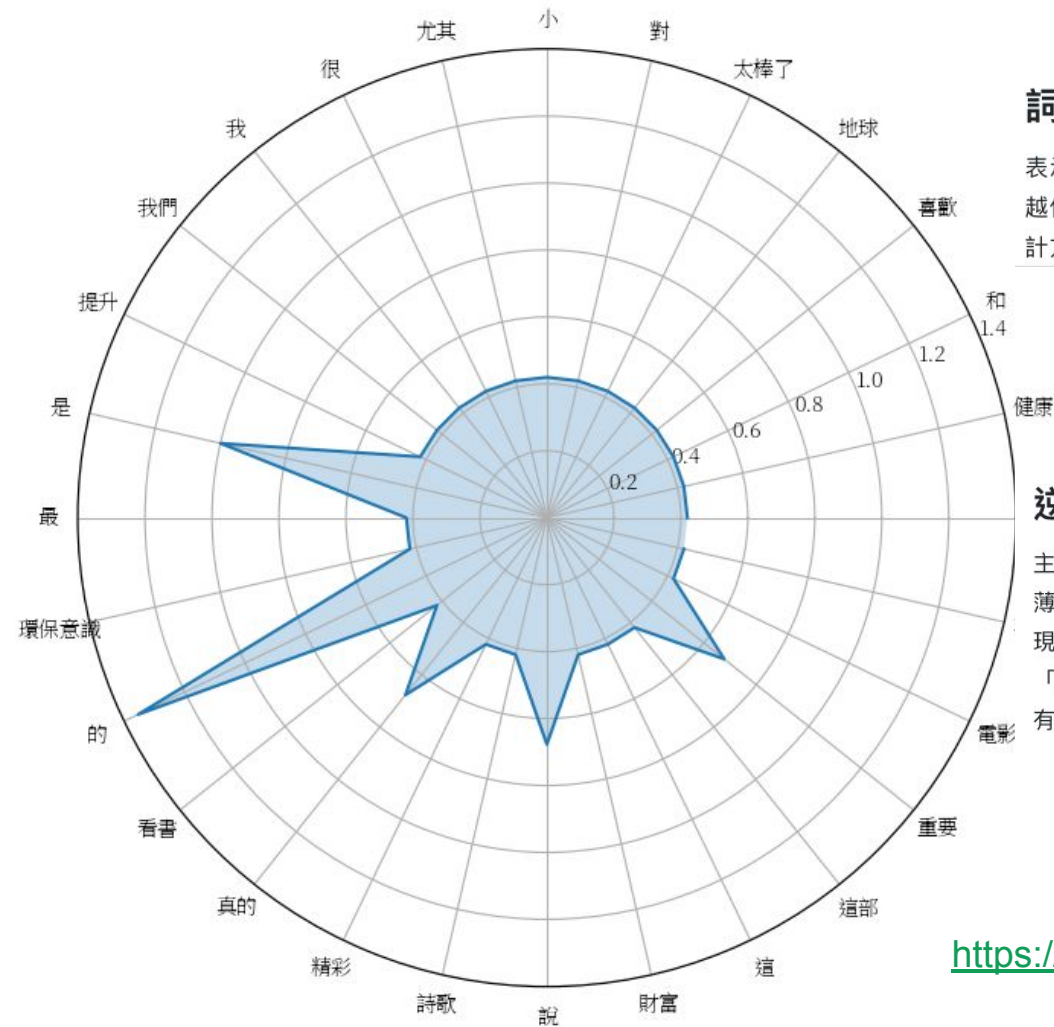


(0.15403270679109896, '皮卡啾'),
(0.15403270679109896, '皮卡'),
(0.0, '腳'),
(0.0, '腳'),
(0.0, '聲音'),
(0.0, '站'),
(0.0, '站'),
(0.0, '發出'),
(0.0, '用')



(0.19804205158855578, '討厭'),
(0.19804205158855578, '感覺'),
(0.0, '腳'),
(0.0, '聲音'),
(0.0, '站'),
(0.0, '發出'),
(0.0, '用')

TF-IDF Scores for Words



詞頻: TF

表示詞在文檔中出現的頻率, 就統計學而言, 只要這個詞在文本中出現越多次代表越值得關注, 因此它會具有一個重要的統計評估指標之一, 但並不是完全相信此統計方式, 看完底下的IDF就會知道為什麼。

逆文檔頻率: IDF

主要目標在於「衡量一個詞語對整個文檔集合的重要性」, 簡單來講就是補足TF薄弱的評估依據, 因為單憑TF並不足以評斷詞語的重要性, 例如一段文章中常常出現「是」這個詞, 以TF的角度來說可能出來的數據是非常重要的, 但對於我們來說「是」這個詞可能只是肯定、接受到了, 並不具備有太重要的資訊, 因此單憑TF會有失真的狀況出現。

<https://vocus.cc/article/647dcbf7fd89780001afdb56>

TF-IDF 和 「基礎群體」 的文本集有關

- **Jieba**: 以「人民日報」的內容做為基礎群體
- **scikit-learn**: 以輸入的文本集為基礎群體
- **Articut**: 只算「實詞 (有意義的詞)」！
 - 可指定預設的台灣報紙為基礎群體，
 - 或是自行撰寫以文本集為基礎群體。



請利用下圖的文字雲進行「文本分析」

設有一期末報告題目：請依以下的文字雲，從人、事、時、地、物 ...等角度寫出 A4 兩頁的分析報告。



請利用下圖的文字雲進行「文本分析」

設有一期末報告題目：請依以下的文字雲，從人、事、時、地、物 ...等角度寫出 A4 兩頁的分析報告。

補充背景一：這是高雄市長選舉結束後，候選人發表的演講 (但可能是不同屆哦！)



請利用下圖的文字雲進行「文本分析」

背景一：這是高雄市長選舉結束後，候選人發表的演講（但可能是不同屆哦！）

背景二：這兩個講稿，一個是勝選人講的，一個是敗選人講的



請利用下圖的文字雲進行「文本分析」

有沒有發現，每多一點背景的補充，你的分析就愈具體，愈容易下筆？



文字雲？腦補 + 暗示的「文本分析」

其實剛剛的文字雲是從這兩篇我自己亂寫的東西生出來的。

所以剛才都是你自己受引導以後的腦補！

277萬一個一天一座一樣不夠不好意思之前了人口今天他鄉任
但是偉大充滿內再次到剛剛勇敢包容向哪裡困難國瑜團結報告好
如子弟孕育對就工作已經市民朋友希望幫幫助很很多很大忘恭喜
意思愛跟感受應該打拼拜託挑戰提到時間更好替朋友期永遠熱情
特別當市長疼惜相信禁結束給綠繼續美好而已能致電藍表示許多
誠請變得變故鄉責任走起跟著路辜負這這塊這座這段選舉還有都
長大陪伴陳其邁雖然難過韓市長願望高雄市他他作為作為其邁其
邁努力努力四年四年家園家園從從或或所有所有故鄉故鄉明天明
天更更最後最後生命生命相依相依這裡這裡開始開始一起一起
一起也也以以共同共同共同加油加油加油地方地方地方市民市
民市民為榮為榮為榮管管管能夠能夠能夠要要要還是還是還是城
市城市城市城市感謝感謝感謝感謝未來未來未來未來為為為人
人人人人市長市長市長市長市長跟跟跟跟跟在在在在支持支
持支持支持支持支持支持不不不不不不不我們我們我們我們我
們我們我們我們大家大家大家大家大家大家大家大家是是是
是是是是是我我我我我我我我我我高雄高雄高雄高雄高雄高
雄高雄高雄高雄高雄高雄高雄高雄高雄高雄的的的的的的的的

2018年280萬——一件一個一個一個一個月一定一定一定一幕一瓶一絲一起
一起一起一起一遍一點上下一心上陣下一代下來不不不不不停不用不
用不用世界世界中中中央也也事人人人人民今天份佛教徒你來保佑做做做
做做做價值先內內內內人全全全全全力全力以赴全部兩年再出來出現分分分
分族群到前前前前剛剛才力量動員包容千斤可台灣台灣台灣各位各位向喜
悅因為國民黨團結團隊團隊團隊在在在在在城市報告報告報告外多少大家
大家大家大家大家大家大家太多太多好不好好好朋友如果姊妹委員
委員宣言將對手對手對手小內閣就是就是就是工作已經市民市民市民市
民市民市長市長市長希望幫忙府廉潔廣大建立影往往往很清楚後後後得心必
須志工朋友愛感謝感謝感謝感謝感謝感謝感謝感謝感謝感謝感謝成長我
我我我我我我我我我我我我我我我我我我我我我我們我們我們我們
我們我們所以所有所有所有所有所有所有所有所有有打拚打拼打造批判批判過
找到找到把投票拚拚挑戰挽起提名擔心擔心支持改變政府政府政治效率敗選
教育敬新明天是晚安更好更好最困難最神奇最重要最高會會會會會會有有
朝有陰朋友朋友們未來未來未來未來未來機會機會民主政治史氣水沉重沒有
沒有沒有沒有海深深清廉為為為無比父特別王金平現場生活的的的的
的的的的的的的的的的的的的的的的的的的的的的的的的的的的的的
的盡速直接相信相信相信相信盾盾看看到真眼礦泉祝福票數秉持立刻等待組織
給給經濟經濟老兄弟肩扛肩膀能自己與菩薩華人萬斤藍綠藍綠藍綠衝表情
表情袖子要要要要要要要親愛訊號訊號請請謝謝謝謝謝謝謝謝謝謝讓讓走
走走起來跟跟跟跟跟跟跳出來輕裝逐步這個這次這裡造勢過程過程選舉選舉
選舉選舉那那都都都鄉親鄉親鄉親鄉親鄉親朋友釋放重新重組重複開始關係
關心院長陳其邁陳其邁陳其邁集合非常非常非常巨大非常強烈鞠躬願意願意
高市府高雄高雄高雄高雄高雄高雄高雄高雄高雄高雄高雄高雄高雄高雄
高雄高雄高雄高雄高雄高雄市高雄市高雄市高雄市點黨部

解讀數據：文本分析的洞見？

- a. 永遠對資料保持懷疑
 - i. 資料來源？
 - ii. 誰寫的？
 - iii. 誰做的資料清洗？
 - iv. 清洗怎麼做的？

- b. 永遠要讀到 [數字化/視覺化] 背後的意義
 - i. 數字怎麼算的？(背後代表的意義為何？)
 - ii. 分類怎麼分的？
 - iii. 圖表怎麼做出來的？

More about TF-IDF

<https://zh.wikipedia.org/zh-tw/Tf-idf#tf-idf的理论依据及不足>

<https://lopentu.github.io/cllt2023/weeks/week9.html#解法二 tf-idf>

坊間最愛用的 TF-IDF 所找出的 "keywords" 並不是語意的關鍵，甚至不該被譯為「關鍵詞」。

TF-IDF 的意義一直是「本文有很多，而其它文本很少的」詞彙！

如果「整批文本，本來就有一個重要資訊是貫穿每一篇，每一篇都有的」，那麼它也會被 TF-IDF 錯誤地濾除，不被視為是「關鍵詞」

關鍵字真的關鍵嗎？

「關鍵字」為什麼叫「關鍵字」？
一個美麗 (而愚蠢) 的翻譯錯誤...

它來自 Dictionary 的操作原理：

```
{key01: value_01,  
key02: value_02,  
key03: value_03,  
key04: value_04,  
}
```

```
{[tf-idf1, tf-idf2, tf-idf3]: document_01,  
[tf-idf4, tf-idf5, tf-idf6]: document_02,  
[tf-idf7, tf-idf8, tf-idf9]: document_03,  
[tf-idf11, tf-idf, tf-idf0]: document_04,  
}
```

**是時候展現
真正的技術了**

NLP 的流程再思考

IT邦幫忙 技術問答 技術文章 IT 徵才 Tag 聊天室 鐵人賽 鐵人館

語文的特性

不像影像相對於點陣圖(Bitmap)的單純，人類的語言具高度曖昧性，一句話可能有多重的意思或隱喻，且隨著時代演進不斷變化，人們不斷創造新詞，語言學家越來越難建立規則(文法及語意學)來規範語文。另外，由於語文的不規則性，在放入模型解析之前，我們通常會先對輸入的語文做前置處理(Preprocess)，包括：

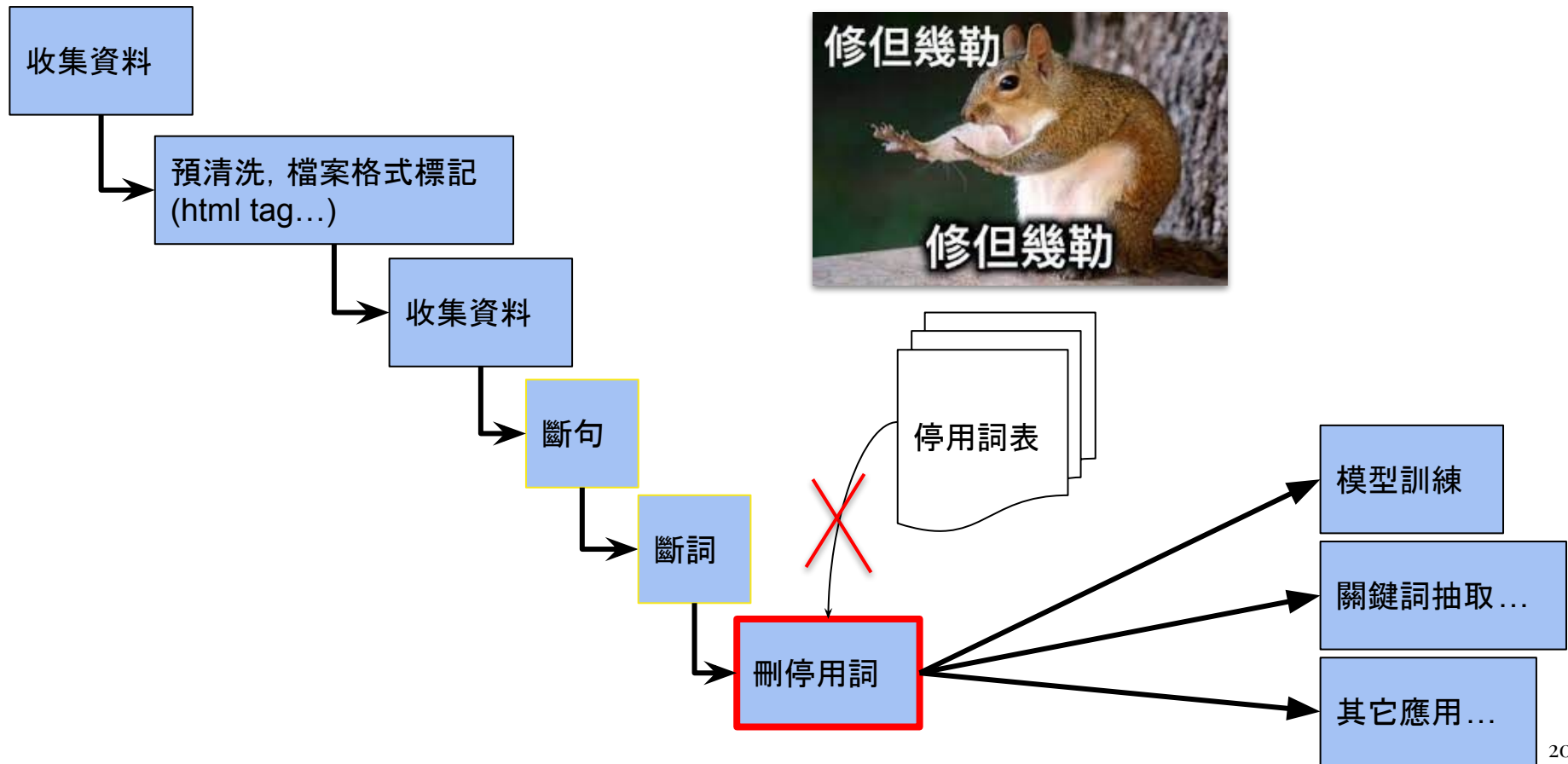
1. 資料清理(Data Cleaning)：例如我們抓網頁資料，必須先清除 HTML 標籤(Tag)，取出乾淨的本文。
2. 標點符號(Punctuation)：對語意沒有影響，通常會忽略他們。
3. 分詞(Tokenize)：英文較容易，一般以空白即可，但中文就比較困難。
4. Stop Words：例如英文的『the』、『as』、『to』、『from』...等介係詞或助詞，他們對語文大意的瞭解可能沒有太大的幫助，通常會忽略他們。
5. 『詞嵌入』(Word Embedding)：要交給電腦計算，將單字轉數值會比較容易處理，『詞嵌入』技術通常會將單字轉為實數，以形成連續的向量空間，比較有名的模型包括Word2Vec 及 GloVe。
6. 相似詞整理：意義相近的單字或片語，在解析字句及分類時必須能呈現出來，Google Tomas Mikolov 創造 Word2Vec 模型，以向量(Vector)空間來定義單字(Word)，就如之前介紹的『[照片比對](#)』計算Cosine，如接近1，就表示兩單字的意義相似。

這句是錯的。語言學家「不」建立文法或語意規則。語言學也「不」規範語言該怎麼講。就像生物學家「不」規範生物該長成什麼模樣。

<https://ithelp.ithome.com.tw/articles/10193224>

<https://www.google.comsearch?q=中文停用詞下載>

Linguistics-Reasoned 的 NLP 流程

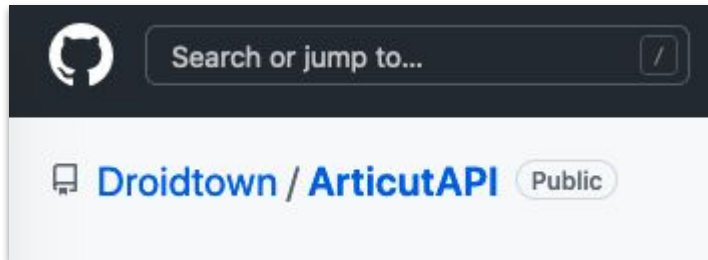


Articut 的基本操作

ps. 如果你手邊沒有 Python 環境
請試試 <https://replit.com/>



<https://api.droidtown.co/>



<https://github.com/Droidtown/ArticutAPI>

材料.. <https://bit.ly/CCU2022NLP>

遊記文本: <https://www.dropbox.com/s/6s5wnutr07eqxll/%E9%81%8A%E8%A8%98.txt?dl=0>

Question: 這是關於哪裡的遊記？

新聞文本: <https://www.dropbox.com/s/9z5ck2d85qlvsu0/localNews.txt?dl=0>

Question: 這是關於哪裡的新聞？涉及人物有誰？這是哪一天的新聞？發生了什麼事？

食記文本: <https://www.dropbox.com/s/s50yc8k8j96ugja/%E9%A3%9F%E8%A8%98.txt?dl=0>

Question: 這是哪裡的食記？在吃什麼？

延伸問題：

1. 三篇文本的 TF-IDF 呈現出什麼？
2. 三篇文本的 ContentWord 呈現出什麼？
3. 如果知道文本來源，是否能透過選擇適當的 NER 工具，呈現出比 TF-IDF 更好的特徵詞彙？

Articut 的進階操作 (Loki): 什麼是語意？



semantics

/sɪˈmantɪks/

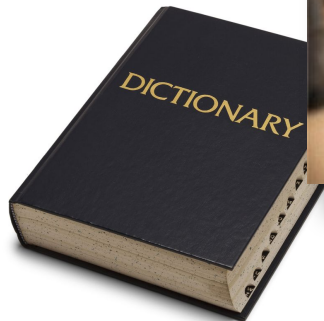
noun

the branch of linguistics concerned with meaning is called *semantics*, concerned

- the meaning of a plural noun: **seminar** "such quibbling o



are *logical semantics*, *implication*, and *lexical* them.



sur- *n.* 1 < OFr *sur-*, above (see *sur-*)
sur- *adj.* forming a surplus
sur- *vt.* **'prised'**, **'pris'ing** [*< Fr prendre, to take*] 1 to come up
 expectedly; take unawares 2 to attempt
 3 to amaze; astonish 2 to attempt
 2 something that surprises
real, se-, rel') **adj.** 1 surrealistic
 2 surrealist
sur- (*-iz'em*) *n.* [see *SUR-* & *REAL*] a movement
 of the unconscious mind trying to depict the work
 of the unconscious mind — **surre'al-ist** *adj.*
sur- *adj.* **'surre'al-ist'**
sur- (*sa ren'dar*) *vt.* [*< Fr sur-, up + rendre,*
 on compulsion 2 to give up possession of; yield to another
 oneself up, esp. as a price or abandon
derine]

語意學具備「可實作的工程定義」來計算什麼叫做「有意義」

自然語言的「語意」工程定義如下：

1. 一個可被理解的句子，必然具有特定語境下的最小句法完備性
2. 一個具有句法最小完備性的句子，必然具有可計算真 值的條件
3. 真值 (truth value) 為「真 (true)」和「假 (false)」兩種可能的集合

試答：

1. [[*今天一早我從出發]] 為什麼沒有語義？
2. [[塑膠工廠]] 的語義為？
3. [[動手術的是我爸爸]] 的語意為？

Team Assignment: (Due: 10/22 on Discord)

1. 請找一篇 500 字以上的遊記文本
2. 同學請設計一支程式, 完成以下功能:
 - a. 利用 jieba 的 TF-IDF 功能, 找出 (1) 中遊記的關鍵字 (ref.
<https://ithelp.ithome.com.tw/articles/10204535>)
 - b. 利用 Articut 的 [實義詞] 功能, 找出 (1) 遊記中的實義特徵詞 (ref.
[https://github.com/Droidtown/ArticutAPI/wiki/03_列出所有的實義詞-\(Content-Word\)](https://github.com/Droidtown/ArticutAPI/wiki/03_列出所有的實義詞-(Content-Word)))
 - c. 利用 Articut 的 NER 位置、地點描述功能, 找出 (1) 中的地點相關詞彙 (ref.
https://github.com/Droidtown/ArticutAPI/wiki/06_列出所有的地方、位置描述)
3. 將你的程式和遊記文本壓縮成一個 zip 檔, 上傳到 Discord 頻道裡, 並用 (2) a, b, c 的結果, 在頻道裡回答下列問題:
 - a. 你使用的遊記, 出遊地點是哪裡? 你怎麼知道的?
 - b. 你使用的遊記, 作者的重點是「遊玩」還是「美食」? 你怎麼知道的?