

# 字詞頻率語料分析與 語意分析

台師大通識教育課程

## 文本分析與程式設計

授課：卓騰語言科技 \_ PeterWolf

# 什麼是詞頻？

## 詞條名稱：詞頻

國家教育研究院辭書

基本資料

英文：

term frequency

作者：

曾元顯

日期：

2012年10月

出處：

圖書館學與資訊科學大辭典

<https://pedia.cloud.edu.tw/Entry/Detail/?title=%E8%A9%9E%E9%A0%BB>

辭書內容

名詞解釋：

詞頻（term frequency，簡稱TF）是指文件中詞彙出現的頻率或是次數，是衡量一個詞彙重要性的一種指標。在各種資訊檢索模型中，如向量空間模型、機率模型、語言模型，都會用到詞頻的概念。一般而言，一個詞彙在某一篇文章中出現的次數越高，即詞頻越高，則其在該篇文件中的代表性越重要。例外的情況，有虛詞、連接詞、代名詞等功能詞（function words），這些詞彙，經常有高詞頻，卻不帶有任何內容意義，在文件的詞彙處理過程，常被特意地停用、過濾掉，因而被稱為停用詞（stop words）。此外，詞彙的重要性，也會考慮到這個詞彙出現在所有文件中的篇數，亦即其文件篇數（document frequency，簡稱DF）。若其出現在越多篇文件中，即DF越高，則表示該詞彙可能為常用字，故而重要性越低，相對於其反向文件篇數

（inverse document frequency，簡稱IDF）也低。例如，在一批有關電腦的文件中，搜尋電腦這個詞彙，幾乎所有的文件都會被找回，則不管其詞頻（TF），電腦在這批文獻中，對檢索沒有幫助，其重要性要降低，剛好對應到其IDF也低。因此，詞彙的重要性，經常以詞頻（TF）以及反向文件篇數（IDF）這兩個概念，一起考慮。真正在應用詞頻的概念時，常會對詞頻做轉換而獲得一個重要性數值。例如，詞彙A與詞彙B在某一篇文章中各出現 $TF(A)=8$ 次與 $TF(B)=2$ 次，但不代表詞彙A的重要性是詞彙B的 $8/2=4$ 倍。常用的轉換作法，有取對數函數，如 $\log_2(TF(A)) = \log_2(8) = 3$ ， $\log_2(TF(B)) = \log_2(2) = 1$ ，如此詞彙A的重要性只比詞彙B多3倍。此外，還有對文件內所有的詞彙頻率做正規化處理，例如 $TF(A)$ 除以該文件所有詞彙的總詞頻（等於詞彙A的出現機率），詞彙的重要性，從頻率的次數分布，變成詞彙的出現機率分布，如此可以在機率的理論中探討詞彙與文件的各種特性。

# 使用「頻率」是迫不得已的

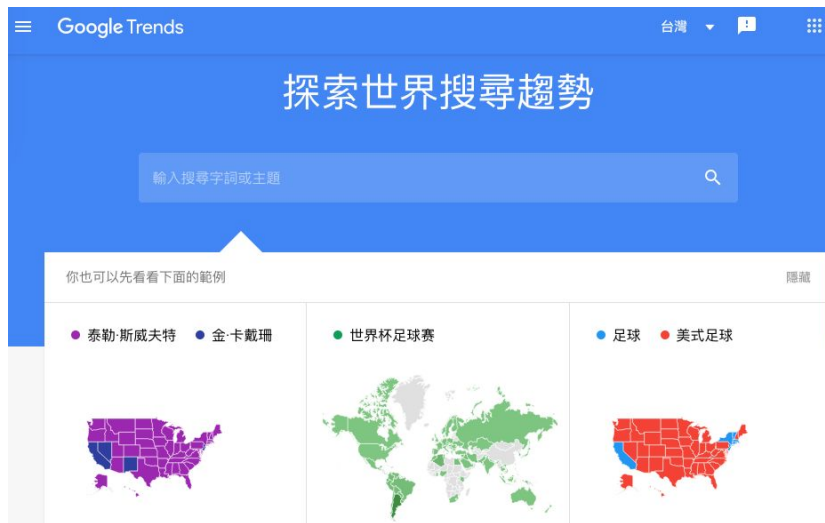
- NLP 尚未出現「**直接聽得懂/讀得懂**」語言的工具
  - 主因是「懂」的工程定義尚不明確，最接近的是「數學應用問題」的解題挑戰



1. 使用「出現頻率」來讓電腦理解語言，我們的電腦將和「狗對人類語言的理解」能力差不多。
2. 事實上，狗做得更好！因為狗對情緒、對語調以及牠所身處的語境，和人類之間的互動關係是更敏感的。
3. 對大多數 NLP 任務而已，電腦只拿到了「文字」(或是加上「語音」)

# 但...頻率還是可以做很多事啦！(畢竟我們(以前)也沒有別的辦法)

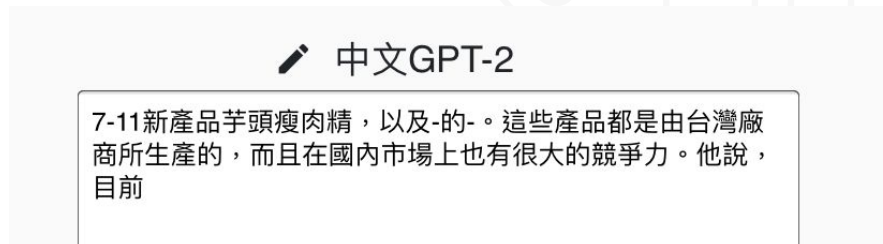
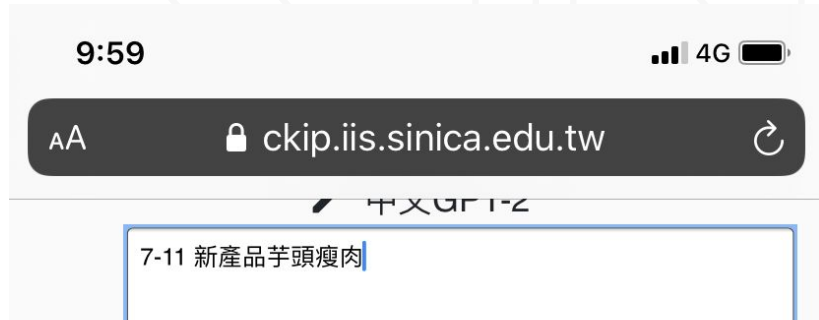
- 輸入法自動選字/矯正/矯誤：<https://mcbopomofo.openvanilla.org/textpool.html>
- 文本生成(加上 ML/DL)：<https://ckip.iis.sinica.edu.tw/service/gpt2/>
- 語言學習教材設計：<https://elearning.ling.sinica.edu.tw/CWordfreq.html>
- 文字雲、文本主題分類、輿情分析、網路聲量以及各種過度腦補的解釋...



利用「詞頻」技術計算的 NLP 應用結果，往往需要領域經驗豐富的人 (或是幻想能力豐富的人) 才能做出解釋。

# 輸入法 (ASR/Text-Gen)自動選字/矯正/矯誤

10大新注音不選字出糗事件	
1	遊戲打到一半說有隻豬在我電腦上〈蜘蛛〉
2	遇到色狼別擔心，馬上抱緊就好〈報警〉
3	在即時通上吹噓：我弟李添財呢〈地理天才〉
4	抱怨找不到老婆，去取悅男人算了〈越南人〉
5	在線上問朋友：你會癢是嗎？〈你會仰式嗎〉
6	朋友說要幫老鼠用一下墓穴〈木屑〉
7	老師教學生電腦，問大家觀音笑了嗎〈關音效〉
8	我爺爺適合男人〈是河南人〉
9	在外面吃東西，要盡量吃氫彈一點〈清淡〉
10	同學去世真的超好笑〈趣事…〉



丁口々く为：<https://tw.appledaily.com/life/20180601/FE5T5FFJMD6ILZZTIV3YVQKO44/>

# 語言學習教材設計

[https://www.huayuworld.org/material-download.php?utm\\_source=%E8%8F%AF%E8%AA%9E%E6%96%87%E6%95%99%E5%AD%B8%E8%B3%87%E6%BA%90%E5%B0%88%E5%8D%80&utm\\_medium=%E5%AD%B8%E8%8F%AF%E8%AA%9E%E5%90%91%E5%89%8D%E8%B5%B0&utm\\_campaign=2018&volume=12](https://www.huayuworld.org/material-download.php?utm_source=%E8%8F%AF%E8%AA%9E%E6%96%87%E6%95%99%E5%AD%B8%E8%B3%87%E6%BA%90%E5%B0%88%E5%8D%80&utm_medium=%E5%AD%B8%E8%8F%AF%E8%AA%9E%E5%90%91%E5%89%8D%E8%B5%B0&utm_campaign=2018&volume=12)



第二課

dì èr kè

家人

jiā rén

心美：你家有幾個人？

xīn měi

nǐ jiā yǒu jǐ ge rén

東明：我家有三個人，爸爸、

dōng míng

wǒ jiā yǒu sān ge rén

bà ba

媽媽和我。

mā ma hàn wǒ

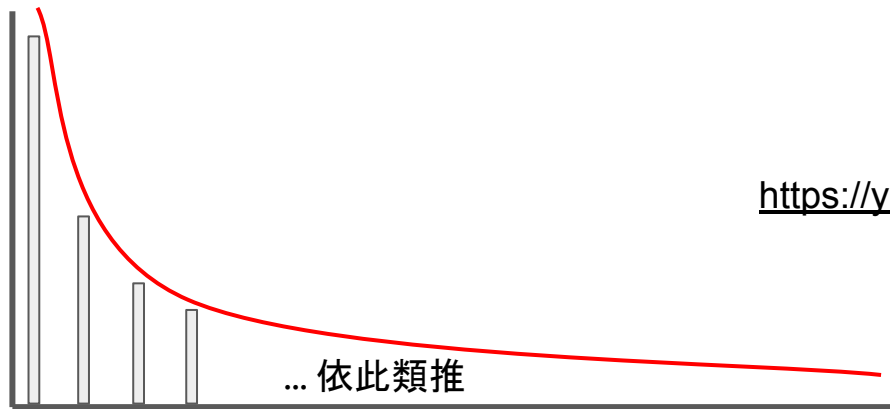




# 「詞頻」的神秘分佈: Zipf's Law

**齊夫定律** (英語: Zipf's law, IPA: /ˈzɪf/) 是由哈佛大學的語言學家喬治·金斯利·齊夫 (George Kingsley Zipf) 於1949年發表的實驗定律。它可以表述為: 在 自然語言 的語料庫裡, 一個單詞出現的頻率與它在頻率表里的排名成 反比。所以, 頻率最高的單詞出現的頻率大約是出現頻率第二位的單詞的 2倍, 而出現頻率第二位的單詞則是出現頻率第四位的單詞的 2倍。這個定律被作為任何與 幂定律機率分布 有關的事物的參考。

<https://zh.wikipedia.org/wiki/%E9%BD%8A%E5%A4%AB%E5%AE%9A%E5%BE%8B>



<https://youtu.be/fCn8zs912OE?t=49>

<https://medium.com/@devalshah1619/a-mysterious-law-so-simple-and-yet-so-universal-aa9f1c8903d1>



# 文本主題分析：台灣關係法

(', ', 126), ('.', 122), ('。', 79), ('臺', 71), ('國', 67), ('我', 58), ('們', 55), ('灣', 55), ('在', 36), ('關', 35), ('美', 34), ('一', 32), ('係', 32), ('人', 26), ('民', 26), ('為', 26), ('也', 24), ('是', 24), ('法', 24), ('力', 23), ('這', 23), ('與', 22), ('不', 21), ('《', 20), ('》', 20), ('主', 20), ('以', 19), ('個', 19), ('會', 19), ('有', 19), ('中', 18), ('和', 18), ('時', 18), ('「', 17), ('」', 17), ('全', 17), ('、', 16), ('共', 16), ('自', 16), ('作', 15), ('及', 15), ('年', 15), ('成', 15), ('持', 15), ('由', 15), ('際', 14), ('合', 13), ('同', 13), ('對', 13), ('能', 13), ('過', 13), ('了', 12), ('來', 12), ('加', 12), ('區', 12), ('定', 12), ('世', 11), ('保', 11), ('兩', 11), ('到', 11), ('大', 11), ('家', 11), ('展', 11), ('平', 11), ('界', 11), ('要', 11), ('o', 10), ('上', 10), ('安', 10), ('心', 10), ('發', 10), ('續', 10), ('起', 10), ('進', 10), ('e', 9), ('並', 9), ('之', 9)

沒有斷詞處理

(', ', 126), ('.', 122), ('。', 79), ('臺灣', 53), ('我們', 50), ('在', 34), ('也', 24), ('美國', 24), ('是', 22), ('《', 20), ('關係法', 20), ('》', 20), ('「', 17), ('」', 17), ('、', 16), ('與', 16), ('不', 15), ('國際', 12), ('和', 11), ('自由', 11), ('合作', 11), ('及', 10), ('世界', 10), ('民主', 10), ('人民', 10), ('關係', 9), ('全球', 9), ('我', 8), ('上', 8), ('共享', 8), ('並', 8), ('成為', 8), ('支持', 8), ('有', 8), ('國家', 8), ('之', 8), ('要', 7), ('中國', 7)

有斷詞處理  
(以 Articut 處理)

# 文本主題分析：台灣關係法

以下斷詞工具採用 Articut NLP 系統：  
<https://api.droidtown.co>

('<FUNC\_inner>的</FUNC\_inner>', 122), ('<LOCATION>臺灣</LOCATION>', 53),  
('<ENTITY\_pronoun>我們</ENTITY\_pronoun>', 50), ('<FUNC\_inner>在</FUNC\_inner>',  
31), ('<MODIFIER>也</MODIFIER>', 24), ('<LOCATION>美國</LOCATION>', 24),  
('<AUX>是</AUX>', 22), ('<ENTITY\_noun>關係法</ENTITY\_noun>', 20),  
('<FUNC\_conjunction>與</FUNC\_conjunction>', 16), ('<FUNC\_negation>不  
</FUNC\_negation>', 15), ('<MODIFIER>國際</MODIFIER>', 12), ('<FUNC\_conjunction>  
和</FUNC\_conjunction>', 11), ('<MODIFIER>自由</MODIFIER>', 11), ('<ACTION\_verb>  
合作</ACTION\_verb>', 11), ('<FUNC\_conjunction>及</FUNC\_conjunction>', 10),  
('<LOCATION>世界</LOCATION>', 10), ('<ENTITY\_noun>人民</ENTITY\_noun>', 10),  
('<ENTITY\_noun>關係</ENTITY\_noun>', 9), ('<LOCATION>全球</LOCATION>', 9),  
('<ENTITY\_pronoun>我</ENTITY\_pronoun>', 8), ('<AUX>為</AUX>', 8),  
('<RANGE\_locality>上</RANGE\_locality>', 8), ('<ACTION\_verb>共享</ACTION\_verb>',  
8), ('<FUNC\_inner>並</FUNC\_inner>', 8), ('<ACTION\_verb>成為</ACTION\_verb>', 8),  
('<ACTION\_verb>支持</ACTION\_verb>', 8), ('<ACTION\_verb>有</ACTION\_verb>', 8)

有斷詞處理  
+  
有詞性標記

# 文本主題分析：台灣關係法

以下步驟採用 Articut NLP 系統的

.getContentWordLIST() 工具處理

<https://api.droidtown.co>

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

('也', 24), ('關係法', 20), ('國際', 12), ('自由', 11), ('合作', 11), ('民主', 10), ('人民', 10), ('關係', 9), ('共享', 8), ('成為', 8), ('支持', 8), ('有', 8), ('國家', 8), ('要', 7), ('價值', 7), ('確保', 7), ('共同', 7), ('和平', 7), ('承諾', 7), ('一起', 7), ('區域', 7), ('人', 6), ('發展', 6), ('夥伴', 6), ('看到', 5), ('時刻', 5), ('臺美', 5), ('良善', 5), ('安全', 5), ('能力', 5), ('參與', 5), ('計畫', 4), ('力量', 4), ('過去', 4), ('政府', 4), ('挑戰', 4), ('架構', 4), ('經濟', 4), ('持續', 4), ('做', 4), ('利益', 4), ('理念', 4), ('戰略', 3), ('研究中心', 3), ('主任', 3), ('會議', 3), ('事務', 3), ('通過', 3), ('臺海', 3), ('很少', 3), ('到', 3), ('基礎', 3), ('國會', 3), ('社會', 3), ('地區', 3), ('脅迫', 3), ('歷史', 3), ('決心', 3), ('努力', 3), ('來', 3), ('進展', 3), ('不只', 3), ('國', 3), ('精神', 3), ('其他', 3), ('願', 3), ('預算', 3)

有斷詞處理  
+  
有詞性標記  
+  
去除功能詞  
(保留有意義的詞)

藉由 Articut 的前處理，很快地就可以讓原來不明顯的「關係法」變成排名超前的關鍵字，以顯示為本篇文本的重點。但...詞頻真的是擷取關鍵字的好方法嗎？



# 文本主題分析：館長直播逐字稿

(', ', 442), ('.', 254), ('我', 248), ('的', 247), ('是', 167), ('你', 166), ('不', 138), ('一', 136), ('們', 116), ('國', 94), ('這', 93), ('個', 92), ('在', 83), ('有', 81), ('台', 80), ('人', 80), ('了', 74), ('要', 73), ('大', 73), ('灣', 66), ('民', 63), ('說', 62), ('他', 62), ('以', 61), ('?', 59), ('黨', 58), ('就', 53), ('來', 52), ('家', 52), ('統', 49), ('會', 48), ('麼', 48), ('天', 48), ('沒', 44), ('都', 43), ('陸', 41), ('講', 40), ('也', 39), ('可', 39), ('好', 38), ('為', 36), ('跟', 36), ('多', 36), ('看', 35), ('所', 34), ('阿', 34), ('現', 33), ('中', 32), ('那', 31), ('今', 31), ('到', 29), ('後', 29), ('意', 27), ('樣', 27), ('然', 26), ('經', 26), ('自', 25), ('共', 25), ('下', 25), ('總', 24), ('去', 24), ('給', 23), ('當', 23), ('再', 22), ('主', 22), ('呢', 21), ('濟', 21), ('很', 21), ('被', 20), ('生', 20), ('但', 20)

沒有斷詞處理

(', ', 441), ('.', 254), ('的', 247), ('我', 182), ('你', 140), ('是', 135), ('不', 73), ('我們', 64), ('台灣', 64), ('?', 59), ('說', 58), ('人', 58), ('在', 52), ('一個', 46), ('要', 45), ('他', 40), ('大陸', 39), ('跟', 36), ('都', 35), ('也', 34), ('沒有', 33), ('就', 32), ('阿', 32), ('會', 31), ('今天', 31), ('有', 30), ('講', 30), ('現在', 29), ('可以', 28), ('這個', 27), ('國民黨', 26), ('了', 25), ('好', 25), ('你們', 25), ('那', 24), ('所以', 24), ('總統', 22), ('他們', 22), ('呢', 21), ('被', 20)

有斷詞處理  
(以 Articut 處理)

# 文本主題分析：館長直播逐字稿

以下斷詞工具採用 Articut NLP 系統：  
<https://api.droidtown.co>

('<FUNC\_inner>的</FUNC\_inner>', 247), ('<ENTITY\_pronoun>我</ENTITY\_pronoun>', 182), ('<ENTITY\_pronoun>你</ENTITY\_pronoun>', 135), ('<AUX>是</AUX>', 135), ('<FUNC\_negation>不</FUNC\_negation>', 73), ('<ENTITY\_pronoun>我們</ENTITY\_pronoun>', 64), ('<LOCATION>台灣</LOCATION>', 64), ('<ACTION\_verb>說</ACTION\_verb>', 58), ('<ENTITY\_noun>人</ENTITY\_noun>', 58), ('<ENTITY\_classifier>一個</ENTITY\_classifier>', 46), ('<ACTION\_verb>要</ACTION\_verb>', 45), ('<FUNC\_inner>在</FUNC\_inner>', 41), ('<LOCATION>大陸</LOCATION>', 39), ('<ENTITY\_pronoun>他</ENTITY\_pronoun>', 36), ('<FUNC\_conjunction>跟</FUNC\_conjunction>', 36), ('<QUANTIFIER>都</QUANTIFIER>', 35), ('<MODIFIER>也</MODIFIER>', 34), ('<FUNC\_inner>就</FUNC\_inner>', 32), ('<FUNC\_negation>沒有</FUNC\_negation>', 32), ('<CLAUSE\_particle>阿</CLAUSE\_particle>', 32), ('<MODAL>會</MODAL>', 31), ('<TIME\_day>今天</TIME\_day>', 31), ('<ACTION\_verb>有</ACTION\_verb>', 31), ('<ACTION\_verb>講</ACTION\_verb>', 30), ('<TIME\_justtime>現在</TIME\_justtime>', 29), ('<MODAL>可以</MODAL>', 28), ('<ENTITY\_DetPhrase>這個</ENTITY\_DetPhrase>', 27), ('<ASPECT>了</ASPECT>', 26), ('<ENTITY\_noun>國民黨</ENTITY\_noun>', 26)

有斷詞處理  
+  
有詞性標記

# 文本主題分析：館長直播逐字稿

以下步驟採用 Articut NLP 系統的

.getContentWordLIST() 工具處理

<https://api.droidtown.co>

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

('說', 58), ('人', 58), ('要', 46), ('也', 34), ('講', 32), ('有', 31), ('國民黨', 26), ('總統', 22), ('國家', 20), ('經濟', 20), ('來', 19), ('再', 18), ('大家', 18), ('統一', 16), ('給', 15), ('去', 15), ('覺得', 14), ('看', 14), ('幫', 13), ('希望', 13), ('新聞', 12), ('真', 12), ('主權', 12), ('支持', 10), ('拉', 10), ('這樣', 10), ('自由', 10), ('只', 10), ('維護', 10), ('到', 9), ('當', 9), ('話', 9), ('拚', 9), ('好', 9), ('選', 8), ('辦法', 8), ('共產', 8), ('拿', 8), ('可能', 8), ('民進黨', 8), ('民主', 8), ('買', 7), ('認為', 7), ('一樣', 7), ('看到', 7), ('共產黨', 7), ('中華民國', 7), ('叫', 7), ('做', 7), ('想要', 7), ('和平', 7), ('很多', 7), ('多', 7), ('卡', 7), ('出來', 6), ('武器', 6), ('需要', 6), ('富', 6), ('這麼多', 6), ('選舉', 6), ('解放', 6), ('員工', 5), ('士康', 5), ('東西', 5), ('罵', 5), ('不見', 5), ('立委', 5), ('住', 5), ('生意', 5), ('生意人', 4)

找不到哪一個是重點嗎？  
那麼...用詞頻來呈現重點關鍵字，  
是否為一個好方法呢？



有斷詞處理  
+  
有詞性標記  
+  
去除功能詞  
(保留有意義的詞)



# 小結：用「詞頻」做語意分析

<http://nccur.lib.nccu.edu.tw/handle/140.119/55328>

- 「關鍵詞」的「詞頻比較高」是一個常見假命題；「詞頻較高」則有可能為「關鍵詞」。
- 若「關鍵詞的詞頻比較高」此描述為真，則我們得到「台灣關係法四十週年談話」的重點是「關係法」的一系列方法，應該同樣可以得到「館長直播逐字稿」的重點才對。
- 換言之，若計算詞頻可以得到X 文本的重點，而無法得到Y 文本的重點，則這個方法就無法保證抓到的是重點。
- 如果一個方法無法保證抓到重點，而你卻覺得「抓到了」，可見這個重點，很有可能是你的「腦補」產生的。
- 如果你「一定要」用詞頻方法，請「正確、仔細、依需求」調整你的前處理步驟，而不要粗暴地跟它拼了！
- 最後，詞頻本身沒有什麼用，但它啟發了後續許多諸如詞句向量和機器翻譯的方法。
  - （至少在西方語系的NLP 任務上，變得比較有用一點點。）

# Quiz:

課堂中已經演示了文本處理中，利用「詞頻」進行文本分析的幾個任務，包括輸入法自動選字、語言學習教材設計、文字雲、語意分析...等。試思考以下問題：

1. 請討論前述的幾個「NLP tasks」，如果不用詞頻，是否有其它方法？
2. 課堂裡一下說「詞頻有時候有用」，一下說「詞頻沒什麼用」。那麼究竟哪一句才是正確的？
  - a. 沒有別招的時候，不得不用。
  - b. 非用不可的時候，也請先做前處理 (CWS -> POS -> NER/Filter\_Content...)
  - c. 詞頻是一種統計方法，統計方法來處理 NLP 最好在「別無它法」時使用。因為 NLP 裡的 "L" 有其內在結構，你應該先嚐試利用其內在結構。

# Assignment: 小組作業，每組繳一份至你們的「組名目錄」即可

1. 從課程 github repo 中把課程中提供的 week11 的目錄 git pull 下來。
2. 把 week11.py 改名為 **week11\_分組隊名.py**
3. 在 **week11\_分組隊名.py** 中，設計你的程式，利用 Jieba/CKIPTagger/Articut 任選一種 NLP 工具完成以下指定規格：
  - a. 把 "**dbp.txt**" 和 "**pbd.txt**" 的內容取出進行詞頻計算。
  - b. 計算兩文本的「字符」出現次數 (如同本簡報 p8 上半頁)，並存成 charCount\_dbp 和 charCount\_pbd
  - c. 計算兩文本的「字詞」出現次數 (如同本簡報 p8 下半頁)，並存成 wordCount\_dbp 和 wordChount\_pbd
  - d. 計算兩文本含有詞性標記的字詞出現次數 (如本簡報 p9)，並存成 posWordCount\_dbp 和 posWordCount\_pbd
  - e. 計算兩文本「去除功能詞」(如本簡報 p10)，並存成 contentWord\_dbp 和 contentWord\_pbd
  - f. 將以上所有的 \_dbp 都存入 count\_result.json 裡
  - g. 請仔細觀察前述兩兩一組的 \_dbp vs. \_pbd 的結果，並小組討論「是否能從詞頻裡分辨究竟哪一篇是人咬狗，哪一篇是狗咬人？。若能，是為什麼，若不能，又是為什麼。」討論結果請另存成 discussion\_分隊隊名.txt 一併上傳到課程 week11 的 repo 裡。

# Assignment Reference:

是否有辦法可以判斷出哪一篇是人咬狗？哪一篇是狗咬人？

如果有辦法，這個辦法能不能被「程式化」自動執行？

('戴', 4), ('工寮', 3), ('飼養', 3), ('多', 3), ('福南步道', 2), ('養', 2), ('土狗', 2), ('飼主', 2), ('婦人', 2), ('咬傷', 2), ('依', 2), ('傷害', 2), ('檢方', 2), ('動物', 2), ('歲', 1), ('男子', 1), ('有', 1), ('盡', 1), ('負', 1), ('看管', 1), ('責', 1), ('導致', 1), ('往步道', 1), ('運動', 1), ('遊客', 1), ('洪姓', 1), ('土狗圍', 1), ('攻', 1), ('雙腳', 1), ('手臂', 1), ('地檢署', 1), ('過失', 1), ('罪嫌', 1), ('起訴', 1), ('調查', 1), ('許', 1), ('洪姓婦人', 1), ('到', 1), ('步道', 1), ('健行', 1), ('走到', 1), ('男工', 1), ('寮處', 1), ('其', 1), ('犬隻群起', 1), ('攻擊', 1), ('致', 1), ('受有', 1), ('左', 1), ('小腿', 1), ('撕裂', 1), ('傷', 1), ('右臂', 1), ('雙下', 1), ('肢', 1), ('口', 1), ('右膝', 1), ('右', 1), ('背', 1), ('左手', 1), ('擦', 1), ('挫傷', 1), ('等', 1), ('指出', 1), ('保護法', 1), ('規定', 1), ('負有', 1), ('防止', 1), ('故', 1), ('侵害', 1), ('生命', 1), ('身體', 1), ('自由', 1), ('財產', 1), ('法律', 1), ('作為義務', 1)

('戴', 4), ('工寮', 3), ('飼養', 3), ('多', 3), ('福南步道', 2), ('養', 2), ('婦人', 2), ('飼主', 2), ('洪姓土', 2), ('狗', 2), ('咬傷', 2), ('依', 2), ('傷害', 2), ('檢方', 2), ('動物', 2), ('歲', 1), ('男子', 1), ('有', 1), ('盡', 1), ('負', 1), ('看管', 1), ('責', 1), ('導致', 1), ('往步道', 1), ('運動', 1), ('遊客', 1), ('婦人圍', 1), ('攻', 1), ('雙腳', 1), ('手臂', 1), ('地檢署', 1), ('過失', 1), ('罪嫌', 1), ('起訴', 1), ('調查', 1), ('許', 1), ('到', 1), ('步道', 1), ('健行', 1), ('走到', 1), ('男工', 1), ('寮處', 1), ('其', 1), ('人', 1), ('隻群起', 1), ('攻擊', 1), ('致', 1), ('土狗', 1), ('受有', 1), ('左', 1), ('小腿', 1), ('撕裂', 1), ('傷', 1), ('右臂', 1), ('雙下', 1), ('肢', 1), ('口', 1), ('右膝', 1), ('右', 1), ('背', 1), ('左手', 1), ('擦', 1), ('挫傷', 1), ('等', 1), ('指出', 1), ('保護法', 1), ('規定', 1), ('負有', 1), ('防止', 1), ('故', 1), ('侵害', 1), ('生命', 1), ('身體', 1), ('自由', 1), ('財產', 1), ('法律', 1), ('作為義務', 1)