專有名詞產生器

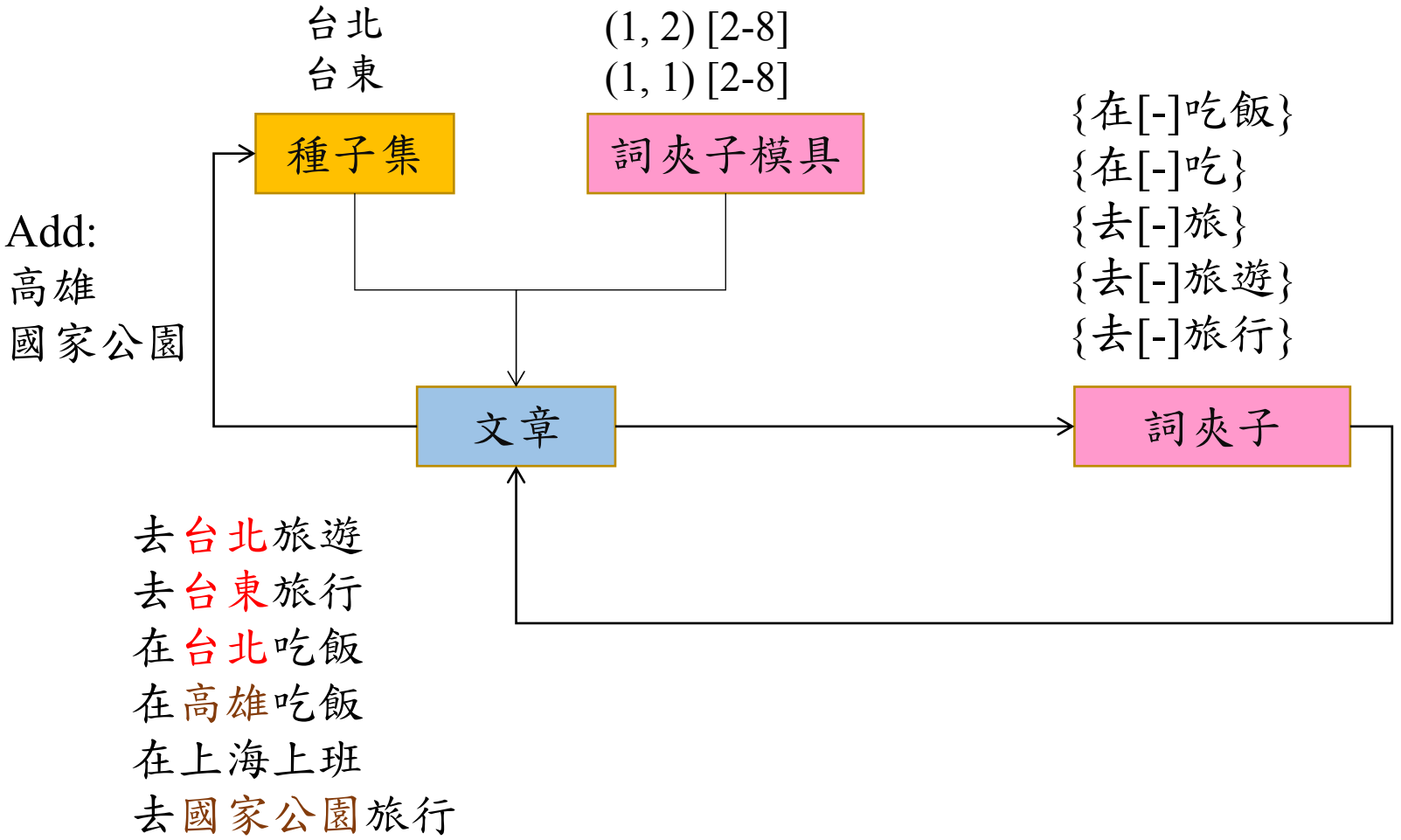
顏秀珍 李御璽

銘傳大學資訊工程學系

詞夾子專有名詞介紹

- 同位性➔一種用來表示文字之間替代性的指標，如果文章之間的詞彙可以替換，而且替代過後的文句仍然非常通順，那我們就可以稱詞彙之間的同位性很高。
  - 例如:台北與高雄就有很高的同位性。
- 同位詞集➔同位性很高的詞彙集合。
- 詞夾子模具➔由 (前文、後文) 組成，主要描述一個詞彙在文件某處的特徵。
  - 例如: 詞夾子模具（1、2）表示前文一個字、後文兩個字，適合用來夾取「地名」。
  - 例如: 上述詞夾子模具從文章中夾取地名可產生詞夾子(在、上班)
  - 可設定所要夾取的專有名詞與詞夾子模具以產生專有名詞的詞夾子

詞夾子演算法


台北
台東
(1, 2) [2-8]
(1, 1) [2-8]
種子集
詞夾子模具
{在[-]吃飯}
{在[-]吃}
{去[-]旅}
{去[-]旅遊}
{去[-]旅行}
Add:
高雄
國家公園
文章
詞夾子
去台北旅遊
去台東旅行
在台北吃飯
在高雄吃飯
在上海上班
去國家公園旅行

詞夾子的缺點

- 需要設置種子集
- 需要設置詞夾子模具
- 需要人為介入
- 不合理(無法解讀)的詞夾子非常多
- 所夾出之專有名詞的雜訊(不正確)依然非常多
- 只考慮單一句子，沒有協同多個句子共同尋找專有名詞
- 人所需花費的時間依然相當的多
- 無法及時找出所有的專有名詞

本系統的優點

- 不需要設置種子集
- 不需要設置詞夾子模具
- 不需要人為介入
- 沒有詞夾子的問題
- 所夾出之專有名詞的雜訊(不正確)較少
- 協同多個句子共同尋找專有名詞
- 人所需花費的時間相當的少
- 可以及時找出專有名詞

原始資料

page id raw	發文粉絲頁ID	發文粉絲頁名稱	主分類	次分類	子分類	分類標籤	文章ID	文章(message)	文章(name)	文章(caption)	文章(descr
'242305665805	2.42306E+14	ETNEWS新聞雲	媒體	網路媒體/社群	網路媒體	新聞	242305665805	#插播	2口罩男站著	ettoday.net	國內金融
'264295946459	2.64296E+11	朱學恒的阿宅萬事通事	人物	網路名人	部落客	0	264295946459	台灣警方這次真的很厲害	破案 一銀3劫	m.appledaily.c	【突發中
'162608724089	1.62609E+14	爆料公社	媒體	網路媒體/社群	社群/論壇	0	162608724089	社員 羅光廷 在 #爆料公社 發文			
'172348966125	1.72349E+14	李開復 Kai-Fu Lee	人物	商管策略	企業家	0	172348966125	我在台灣大學畢業演講影片來了。我在	我在台灣大學畢業演講視頻版		
'353390642311	3.53391E+11	nagee	人物	藝文設計	圖文作家	0	353390642311	【為什麼我們拍不出屍速列車，做不出	動態時報相片		
'101898876511	1.01899E+14	Sway房市觀測站	人物	網路名人	寫作	0	101898876511	投資客毀了台北東區，也毀了台北的市	獨家／享得利	news.housefun	東區店面
'232633627068	2.32634E+11	蘋果日報	媒體	書報雜誌	報紙	0	232633627068	這次真的多虧豬隊友...	【一銀追回6-	appledaily.com	國內金融
'337007633116	3.37008E+11	唐綺陽占星幫	人物	命理占卜	算命專家	占星	337007633116	【唐綺陽03/06-03/12運勢週報】			
'242305665805	2.42306E+14	ETNEWS新聞雲	媒體	網路媒體/社群	網路媒體	新聞	242305665805	我現在去排隊還來得及嗎？（#小陳）	左營麥當勞店	ettoday.net	當前經濟
'118250504903	1.18251E+14	馬英九	人物	政治	首長官員	國民黨	118250504903	兩公約從未廢除死刑，亦未增加執行死	筆震		卜小燈泡
'172348966125	1.72349E+14	李開復 Kai-Fu Lee	人物	商管策略	企業家	0	172348966125	【潸然淚下！41歲英年早逝的北大才子	李開復 Kai-Fu Lee 貼文的照片		
'143944236949	1.43944E+11	王定宇	人物	政治	民代立委	民進黨	143944236949	為台灣治安單位喝采...	破案 一銀3劫	m.appledaily.c	【突發中
'353390642311	3.53391E+11	nagee	人物	藝文設計	圖文作家	0	353390642311	【蔡英文向原住民道歉】	動態時報相片		
'353390642311	3.53391E+11	nagee	人物	藝文設計	圖文作家	0	353390642311	【朱衣班的隊友】	動態時報相片		
'353390642311	3.53391E+11	nagee	人物	藝文設計	圖文作家	0	353390642311	【沒事沒事，不是大便】	動態時報相片		
'272264852807	2.72265E+14	郝明義Rex How	人物	商管策略	企業家	0	272264852807	【對林全院長的回應之一：八億六千多	郝明義Rex How 貼文的照片		
'353390642311	3.53391E+11	nagee	人物	藝文設計	圖文作家	0	353390642311	【賣過期鳳梨酥與賣台的交保金額】	動態時報相片		
'152472278103	1.52472E+14	賴清德	人物	政治	首長官員	民進黨	152472278103	各位親愛的市民朋友，大家早安！	賴清德貼文的照片		
'146846699009	1.46847E+15	小聖蚊的治國日記	人物	網路名人	圖文作家	0	146846699009	中華民國史上第一罪人	【民報】她分	peoplenews.tw	前總統陳
'353390642311	3.53391E+11	nagee	人物	藝文設計	圖文作家	0	353390642311	【憎恨「返校」的布袋爺爺和小華兔】	動態時報相片		
'353390642311	3.53391E+11	nagee	人物	藝文設計	圖文作家	0	353390642311	【徐旭東挖的天坑可以幹麻？】	動態時報相片		
'337007633116	3.37008E+11	唐綺陽占星幫	人物	命理占卜	算命專家	占星	337007633116	把二戰過程說得清清楚楚，道理也明確	黑貓老師		昨天評論
'353390642311	3.53391E+11	nagee	人物	藝文設計	圖文作家	0	353390642311	【我要成為志工王（新版）】	動態時報相片		
'162608724089	1.62609E+14	爆料公社	媒體	網路媒體/社群	社群/論壇	0	162608724089	#天橋下說書爺：#炮神廟第1-2集	動態時報相片		

原始資料

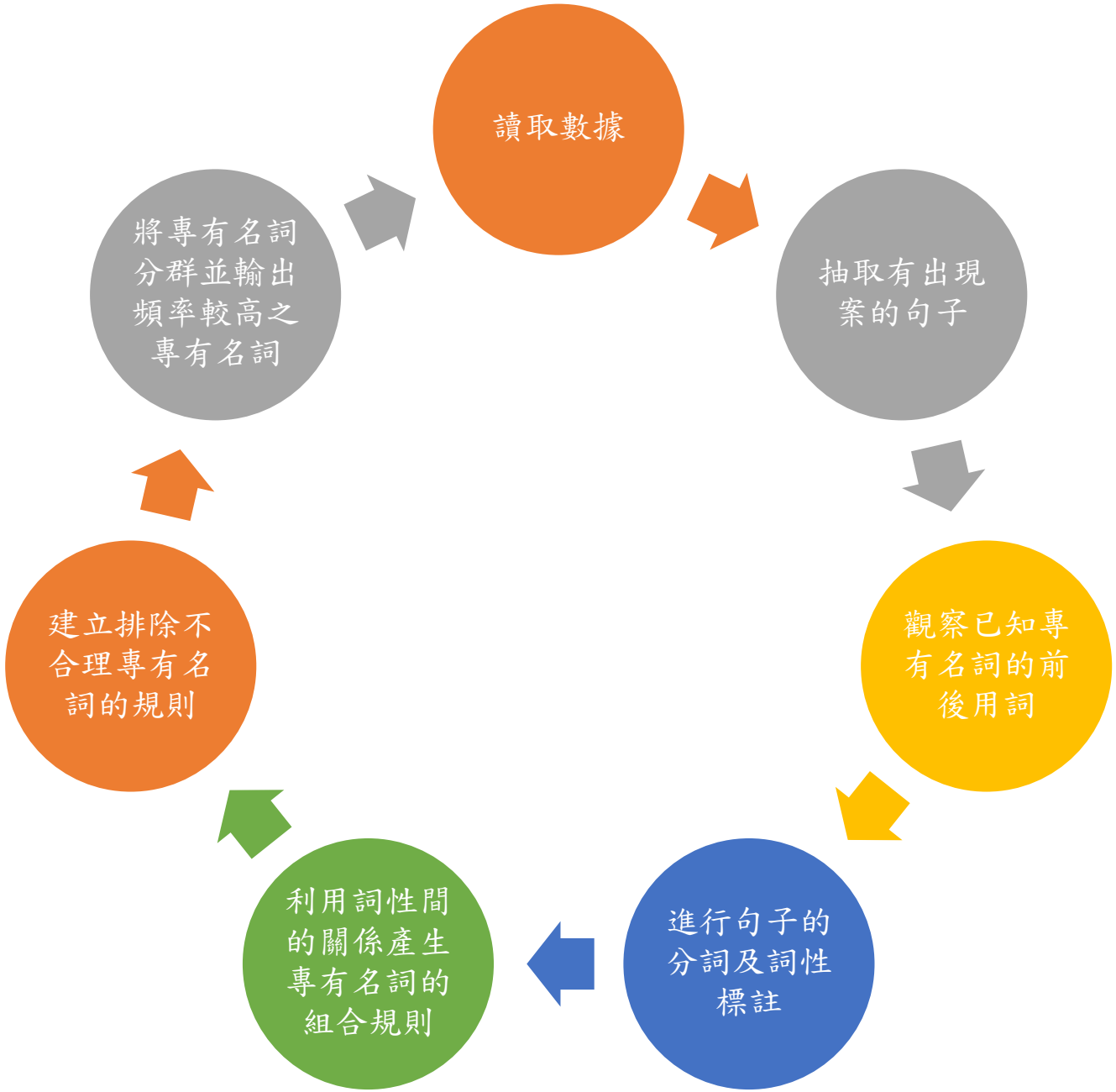
文章(message)
#插播
台灣警方這次真的很厲害
社員 羅光廷 在 #爆料公社 發文
我在台灣大學畢業演講影片來了。我在演講中談到AI時代的三個想像圖：1）你希望自己在AI時代人類工作的金字塔中，佔據哪個位置？2）你想怎麼使用你的魔法棒，引
【為什麼我們拍不出屍速列車，做不出寶可夢】
投資客毀了台北東區，也毀了台北的市容，現在你們要去那兒呢？
這次真的多虧豬隊友...
【唐綺陽03/06-03/12運勢週報】
我現在去排隊還來得及嗎？（#小陳）
兩公約從未廢除死刑，亦未增加執行死刑的條件。日前士林地方法院自稱因受兩公約限制無法判決小燈泡命案兇手死刑，顯然有嚴重誤會。
【潸然淚下！41歲英年早逝的北大才子，給四歲幼子的臨別贈言】
為台灣治安單位喝采...
【蔡英文向原住民道歉】
【朱衣班的隊友】
【沒事沒事，不是大便】
【對林全院長的回應之一：八億六千多萬元蚊子城的例子】
【賣過期鳳梨酥與賣台的交保金額】
各位親愛的市民朋友，大家早安！
中華民國史上第一罪人
【憎恨「返校」的布袋爺爺和小華兔】（續前頁圖文）
【徐旭東挖的天坑可以幹麻？】
把二戰過程說得清清楚楚，道理也明確
【我要成為志工王（新版）】

找到的專有名詞 (以XX案為範例)

Keyword	Frequency	Cluster
cluster=兆豐案	81	*
兆豐銀案	2	1
教室兆豐銀他字案	2	1
兆豐金控弊案	2	1
兆豐金樂陞案	1	1
兆豐弊案	2	1
兆豐案	61	1
從兆豐銀案	1	1
從兆豐案	1	1
兆豐銀美國紐約分行弊案	1	1
兆豐金案	4	1
兆豐金弊案	4	1
cluster=關說案	11	*
一審關說案	4	5
司法關說案	5	5
關說案	1	5
一審案	1	5

Keyword	Frequency	Cluster
cluster=永豐超貸案	28	*
永豐超貸案	1	8
永豐弊案	6	8
超貸案	6	8
豐金超貸案	1	8
永豐金案	6	8
永豐金控超貸案	1	8
永豐案	7	8
cluster=BOT案	36	*
雄大巨蛋BOT案	1	9
OT案	2	9
BOT案	5	9
BOT弊案	2	9
巨蛋案	26	9
cluster=強盜超商店長案	6	*
強盜超商店長案	2	17
強盜案	3	17
超商店長案	1	17

執行步驟


讀取數據
抽取有出現案的句子
觀察已知專有名詞的前後用詞
進行句子的分詞及詞性標註
利用詞性間的關係產生專有名詞的組合規則
建立排除不合理專有名詞的規則
將專有名詞分群並輸出頻率較高之專有名詞

抽取有出現案的句子

台灣警方這次真的很厲害　　其他國家類似	案	件大多一毛錢沒追回來，一個人也抓不到
我2005年面臨有史以來最大的「跳槽訴訟	案	的時候，我對我的律師佩服的五體投地。他呈
值錢！創新工場投資台灣創業者最成功的三個	案	子不是科技公司，而是服務公司，分別是：“
啊，不要亂花錢去研究為何國人不生小孩，答	案	就是高房價。　　高房價讓人縮衣節食只為買
方法院自稱因受兩公約限制無法判決小燈泡命	案	兇手死刑，顯然有嚴重誤會。　　1966年
食屎乎？另外，高金素梅已幫國民黨簽下釋憲	案	，準備推翻千辛萬苦才過的「不當黨產條例」
【朱衣班的隊友】　　他認為A方	案	超讚　你建議C其實更不錯　他聽不懂　腦中
怪這個國家一直停滯不前　　大家專注在個	案	的獵巫、　哀悼、集氣　要進一步邀請你從制
思考的謾罵？原住民議題也是，南海國際仲裁	案	也是，之前中國強擄台灣的跨國詐欺犯也是，
是媒體沒有把焦點放在她身上，她就是廣大興	案	的洪慈綪洪大姐。前年底，洪大姐以廣大興案
案的洪慈綪洪大姐。前年底，洪大姐以廣大興	案	時為民喉舌的民氣，順利當選屏東琉球鄉的議
提起當選無效之訴。　　有稍微了解廣大興	案	的人都知道，洪家是漁民，又剛逢家中經濟支
。　　根據工程會公布的資料，因為多數標	案	都是以最低標得標，結果五千萬到二億元工程
算更高達百分之八十二。　　此外，許多標	案	的時程不合理，五千萬到二億元的延宕比例高
他說公務人員雖怕彈劾，可是彈劾很難成	案	。一般都是糾正案，　但糾正案只是參考而已
雖怕彈劾，可是彈劾很難成案。一般都是糾正	案	，　但糾正案只是參考而已　。除非重大弊案
是彈劾很難成案。一般都是糾正案，　但糾正	案	只是參考而已　。除非重大弊案會走司法途徑
案，　但糾正案只是參考而已　。除非重大弊	案	會走司法途徑，但一拖數年甚至十幾年並不少
還考慮犯嫌有沒有錢？　　[新聞]　共諜	案	爆退役上校辛澎生10年前早被吸收　ht
23D　　後續更新　【共諜動畫】涉共諜	案	少將　轉調空軍司令部委員　https:/
近年（馬英九那8年）簡直是退役將軍共諜	案	的量產期啊　至少包括：　1.　陸軍少將許
將，數次交付中國台灣軍事機密，並洩漏博勝	案	等重要機密。　3.　憲兵司令部前中將副司
太政治太不生活化了　好我們來看看鳳梨酥的	案	例　　[新聞]　維格餅家竟賣過期鳳梨酥
市立美術館的補助」、「大台南會展中心興建	案	」、「中研院南部院區設置案」、「跨曾文溪
南會展中心興建案」、「中研院南部院區設置	案	」、「跨曾文溪大橋興建案」及「新化果菜市

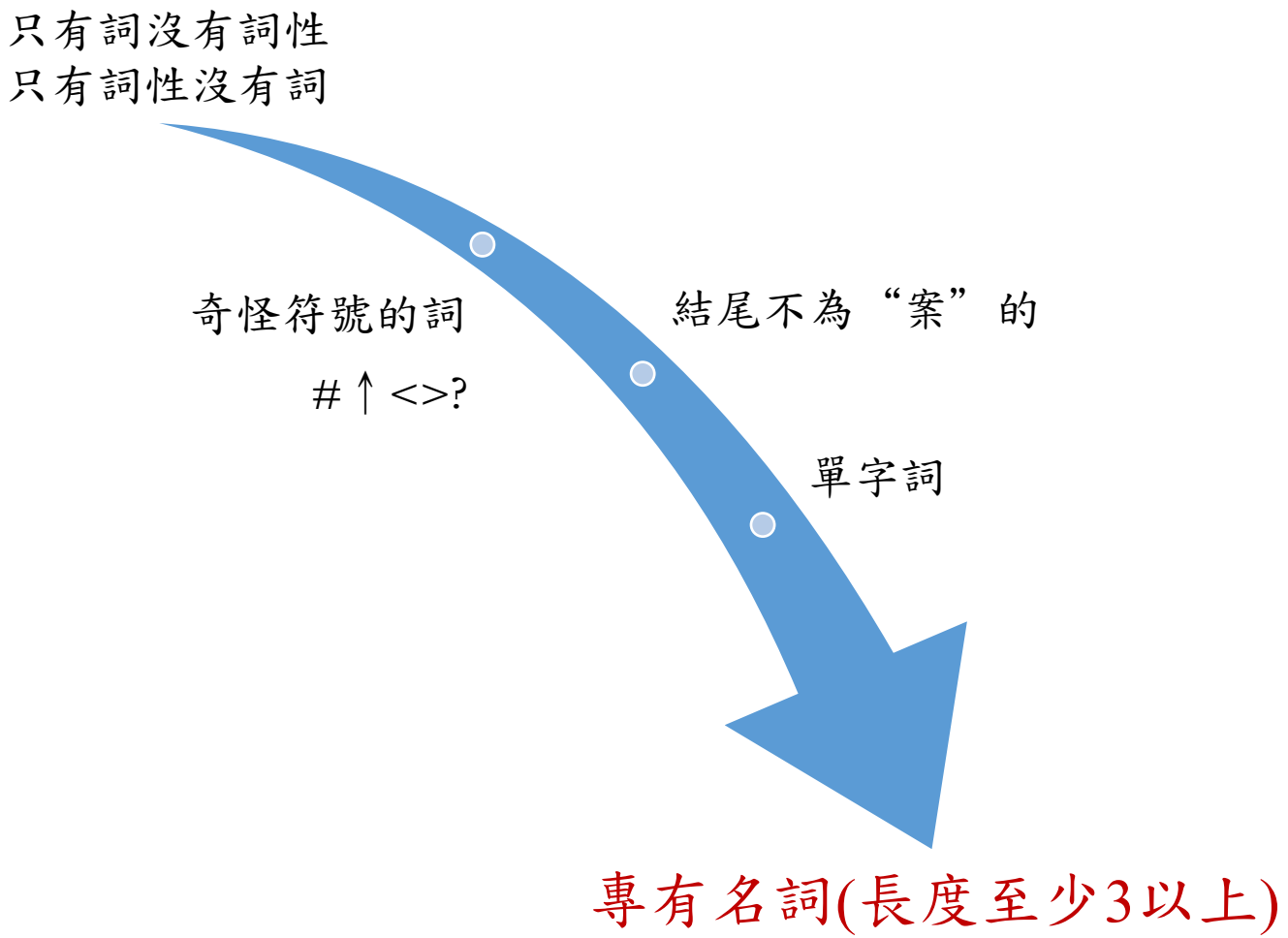
觀察已知專有名詞的前後用詞

面更高的個案關心，像是兆豐金千億洗錢案、	樂陞案	、或者那些動輒百億千億的弊案一堆，卻逍遙
8日，我寫了"打陳文茜是一石三鳥"，直指	樂陞案	綠營媒體操作的目標是甚麼。 9月13日，
月19日(昨天)在各大報刊登半版廣告，對	樂陞案	做了四點聲明。 從一個角度來說，算是回應
立董事: 陳文茜，李永萍和尹啟銘身上。	樂陞案	中，如果陳文茜，李永萍和尹啟銘三個獨立董
毫無過失且盡責的。 那，讓這三個人來背負	樂陞案	破局的罪責，其實真是一石三鳥的連環好計。
美青。明朝的劍既然可以斬清朝的官，那藉著	樂陞案	收購破局，鬥臭鬥垮陳文茜，李永萍和尹啟銘
。 這樣的背後算計，更是驚人。 11	樂陞案	讓我們看到台灣在金融法令和操作上的叢林生
。 讓我們看到無辜者怎樣可以陷入牢籠。	樂陞案	極可能有犯罪者，有瀆職者，有共犯，有合謀
	樂陞案	裡的關鍵代理人--中信商銀 日商百尺竿
樣 ? 如果中信商銀在紐約，那麼，以它在	樂陞案	的作為和角色，對比兆豐金， 它難道不會被
就會被消音 近一兩個月來像是： 兆豐金	樂陞案	中華體協 復興航空解散背後的內線交易炒
上了2002年《時代》雜誌封面。 不論	樂陞案	、黨產，我們希望有人能挺身而出，同樣扮演
人能挺身而出，同樣扮演吹哨者的角色。至於	樂陞案	恐涉入的人物，其中李永萍因楊實秋有共事過
p://bit.ly/2eR9qLv #	樂陞案	#楊瑞仁 #狡猾到測謊都測不出 #壹週
20161018/970585/ #	樂陞案	#超級營業員 #楊瑞仁 鋼琴家都是魔
無能】 台灣發生過的重大金融弊案，像是	兆豐洗錢案	、中信紅火案及購地案等，都不見任何董事或
	樂陞案	應朝經濟犯罪偵辦 樂陞董座與獨董應即約
失，不是沒有道理，從過去一年來，兆豐案、	樂陞案	到永豐案，金管會哪一件事有調查清楚，給民
尹啟銘是樂陞鬥神？ 尹在	樂陞案	發後就神隱了起來。其實三位獨董背離職務，
轄市長，號稱國民黨南霸天。 2000年	中興銀弊案	爆發，總經理王宣仁涉及與中興銀行董事長王
法和洗錢犯行，判處有期徒刑7年確定。另涉	台中商銀超貸案	75億元，93年依違反商業會計法、偽造文
做了個夢，只夢兩分鐘，兩分鐘看完350億	獵雷艦弊案	https://www.youtube
調應即徹查」記者會，針對百尺竿頭公司收購	樂陞案	，創下台灣史上首次「違約」的公開收購案，
達標後，也不會執行收購程序，百尺竿頭收購	樂陞案	必定會違約」後，立刻在8月12日不只以口
不痛不癢。 王麟認為百尺竿頭收購	樂陞案	的疑點部分有五點： 1 在百尺竿頭公

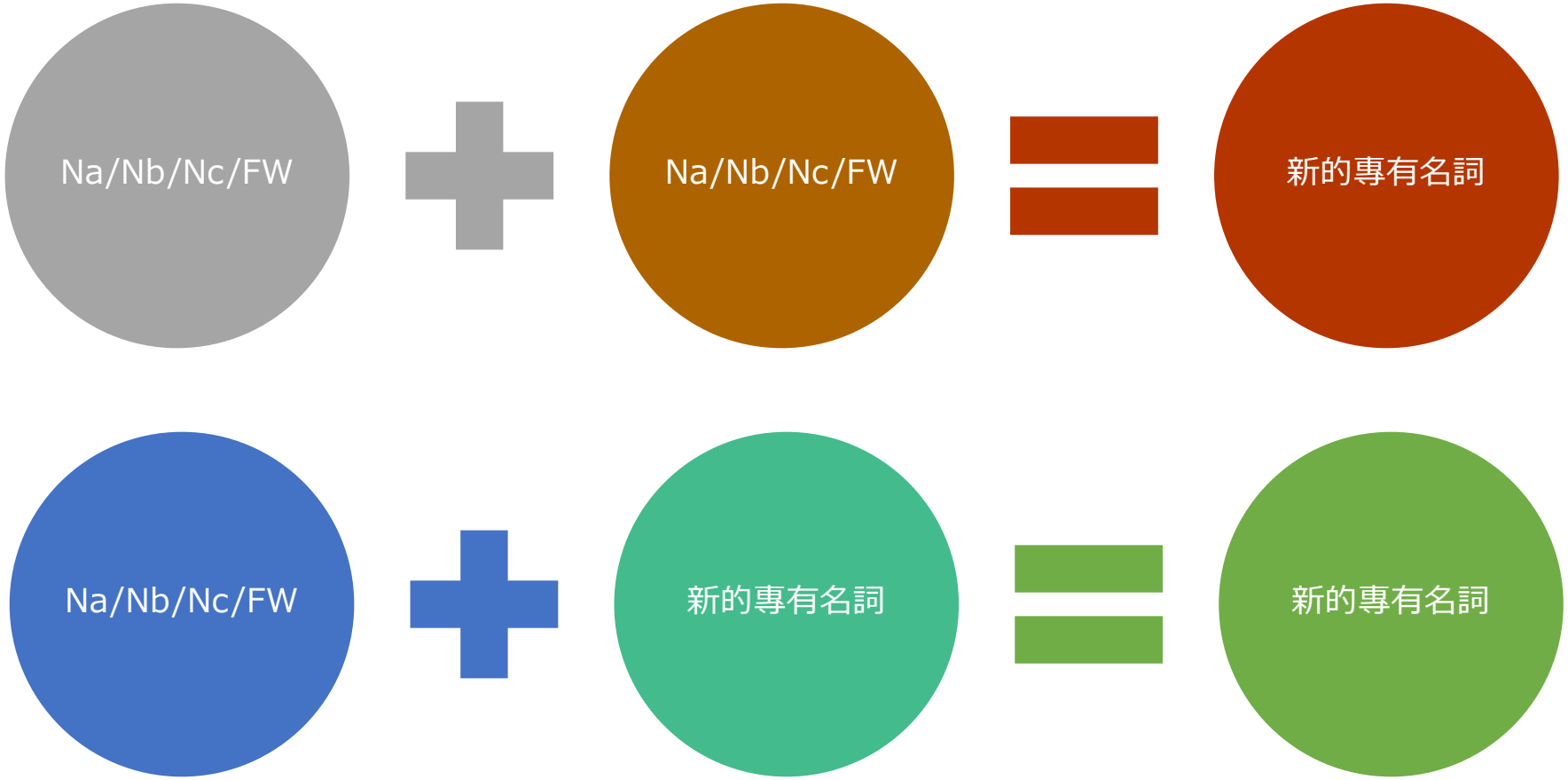
進行句子的分詞及詞性標註

c) 警方(Na) 這(Nep) 次(Nf) 真的(D) 很(Dfa) 厲害(VH) (FW) 其他(Neqa) 國家(Na) 類似(VG)	案件(Na)	大多(Neqa) 一(Neu) 毛(Nf) 錢(Na) 沒(D) 追回來(VB) ，(COMMA
０５年(Nd) 面臨(VK) 有史以來(D) 最(Dfa) 大(VH) 的(DE) 「(PARENTHESISCATEGORY) 跳槽(VA)	訴訟案(Na)	的(DE) 時候(Na) ，(COMMACATEGORY)
創新(VC) 工場(Nc) 投資(VC) 台灣(Nc) 創(VC) 業者(Na) 最(Dfa) 成功(VH) 的(DE) 三(Neu) 個(Nf)	案子(Na)	不(D) 是(SHI) 科技(Na) 公司(Nc) ，(COMMACATEGORY)
	答案(Na)	就是(Cbb) 高(VH) 房價(Na) 。(PERIODCATEGORY)
院(Nc) 自稱(VG) 因(Cbb) 受(VJ) 兩(Neu) 公約(Na) 限制(Na) 無法(D) 判決(VE) 小(VH) 燈泡(Na)	命案(Na)	兇手(Na) 死刑(Na) ，(COMMACATEGORY)
高金素梅(Nb) 已(D) 幫(P) 國民黨(Nb) 簽下(VC)	釋憲案(Na)	，(COMMACATEGORY)
) 朱衣班(Nb) 的(DE) 隊友(Na) 」(PARENTHESISCATEGORY) (FW) 他(Nh) 認為(VE) A(FW)	方案(Na)	超讚(Nb) (FW) 你(Nh) 建議(VE) C(FW) 其實(D) 更(D) 不錯(
南海(Nc) 國際(Nc)	仲裁案(Na)	也(D) 是(SHI) ，(COMMACATEGORY)
她(Nh) 就是(Cbb)	廣大興案(Na)	的(DE) 洪慈(Nb) 綉(FW) 洪大姐(Nb) 。(PERIODCATEGORY)
洪大姐(Nb) 以(P) 廣大興(Nb)	案案(Na)	的(DE) 洪慈(Nb) 綉(FW) 洪大姐(Nb) 。(PERIODCATEGORY)
洪大姐(Nb) 以(P)	廣大興案(Na)	時(Ng) 為民喉舌(VH) 的(DE) 民氣(Na) ，(COMMACATEGORY)
(FW) 有(V_2) 稍微(D) 了解(VK)	廣大興案(Na)	的(DE) 人(Na) 都(D) 知道(VK) ，(COMMACATEGORY)
因為(Cbb) 多數(Neqa)	標案(Na)	都(D) 是以(Cbb) 最(Dfa) 低(VH) 標(VC) 得標(VH) ，(COMMAC
許多(Neqa)	標案(Na)	的(DE) 時程(Na) 不(D) 合理(VH) ，(COMMACATEGORY)
可是(Cbb) 彈劾(VC) 很(Dfa) 難(VH)	成案(Na)	。(PERIODCATEGORY)
一般(D) 都(D) 是(SHI)	糾正案(Na)	，(COMMACATEGORY)
(FW) 但(Cbb)	糾正案(Na)	只是(D) 參考(VC) 而已(T) 雖(Cbb) 怕(VK) 彈劾(VC) ，(COMMA
可是(Cbb) 彈劾(VC) 很(Dfa) 難(VH)	成案(Na)	。(PERIODCATEGORY)
一般(D) 都(D) 是(SHI)	糾正案(Na)	，(COMMACATEGORY)
(FW) 但(Cbb)	糾正案(Na)	只是(D) 參考(VC) 而已(T) (FW) 。(PERIODCATEGORY)
除非(Cbb) 重大(VH)	弊案(Na)	是(SHI) 彈劾(VC) 很(Dfa) 難(VH) 成案(Na) 。(PERIODCATEGOR
一般(D) 都(D) 是(SHI)	糾正案(Na)	，(COMMACATEGORY)
(FW) 但(Cbb)	糾正案(Na)	只是(D) 參考(VC) 而已(T) (FW) 。(PERIODCATEGORY)
除非(Cbb) 重大(VH)	弊案(Na)	會(D) 走(VA) 司法(Na) 途徑案(Na) ，(COMMACATEGORY)

建立排除不合理專有名詞的規則



利用詞性間的關係產生專有名詞的組合規則



將專有名詞分群並輸出頻率較高之專有名詞

專有名詞	次數
兆豐銀案	2
惠心禁藥案	1
林姓男子犯案	1
者保護法草案	1
一審關說案	4
學術弊案	1
死刑個案	3
教室兆豐銀他字案	2
雄大巨蛋BOT案	1
國人答案	1
捷運站共構案	1
中信弊案	1
兆豐金控弊案	2
服貿法案	1
關說案	1
強盜超商店長案	2
燈泡命案	2

Keyword	Frequency	Cluster
cluster=兆豐案	81	*
兆豐銀案	2	1
教室兆豐銀他字案	2	1
兆豐金控弊案	2	1
兆豐金樂陞案	1	1
兆豐弊案	2	1
兆豐案	61	1
從兆豐銀案	1	1
從兆豐案	1	1
兆豐銀美國紐約分行弊案	1	1
兆豐金案	4	1
兆豐金弊案	4	1
cluster=關說案	11	*
一審關說案	4	5
關說案	1	5
一審案	1	5

如何將專有名詞分群

1	兆豐銀案	一審關說案	學術弊案	教室兆豐銀他字案	雄大巨蛋BOT案	兆豐弊案	關說案
兆豐銀案	3	0	0	3	0	2	0
一審關說案	0	4	0	0	0	0	2
學術弊案			3	0	0	1	0
教室兆豐銀他字案				7	0	2	0
雄大巨蛋BOT案					7	0	0
兆豐弊案						3	0
關說案							2

Longest Common Subsequence

4	兆豐銀案	一審關說案	學術弊案	教室兆豐銀他字案	雄大巨蛋BOT案	兆豐弊案	關說案
兆豐銀案	**1**	0	0	**0.65**	0	**0.67**	0
一審關說案		**1**	0	0	0	0	**0.71**
學術弊案			**1**	0	0	0.33	0
教室兆豐銀他字案				**1**	0	0.44	0
雄大巨蛋BOT案					**1**	0	0
兆豐弊案						**1**	0
關說案						0	**1**

Similarity

Longest Common Subsequence:

(兆豐銀案,教室兆豐銀他字案) = 4

(兆豐銀案,兆豐弊案) = 3

Similarity(兆豐銀案,教室兆豐銀他字案) =

$$\frac{LCS(\text{兆豐銀案},\text{教室兆豐銀他字案})-1}{\sqrt{(|(\text{兆豐銀案}|-1)\times(|\text{教室兆豐銀他字案}|-1)}}=\frac{3}{\sqrt{3\times7}}=0.654654$$

專有名詞	次數	群集
兆豐銀案	2	1
一審關說案	4	
學術弊案	1	
教室兆豐銀他字案	2	1
雄大巨蛋BOT案	1	
兆豐弊案	2	1
關說案	1	

專有名詞	次數	群集
兆豐銀案	2	1
一審關說案	4	2
學術弊案	1	
教室兆豐銀他字案	2	1
雄大巨蛋BOT案	1	
兆豐弊案	2	1
關說案	1	2

專有名詞	次數	群集
兆豐銀案	2	1
一審關說案	4	2
學術弊案	1	3
教室兆豐銀他字案	2	1
雄大巨蛋BOT案	1	4
兆豐弊案	2	1
關說案	1	2

如何決定群集名稱

- Cluster Name = 兆豐案
  - 兆豐銀案
  - 教室兆豐銀他字案
  - 兆豐弊案
  - 兆豐案
  - 　從兆豐銀案
  - 從兆豐案
  - 兆豐金案

- 第一步找出最長的案名
  - 教室兆豐銀他字案

- 第二步統計每個字出現在多少案名中
  - 教：1 (1/7=14%)
  - 室：1 (1/7=14%)
  - 兆：7 (7/7=100%)
  - 豐：7 (7/7=100%)
  - 銀：3 (3/7=43%)
  - 他：1 (1/7=14%)
  - 字：1 (1/7=14%)
  - 案：7 (7/7=100%)
- 第三步訂一個門檻值(50%)決定群集名稱
  - 兆豐案

Q & A

顏秀珍 李御璽

銘傳大學資訊工程學系