# 文本擷詞工具 2020 版

操作手册

撰稿人:黃韋菱、洪一梅 撰稿時間:2020年11月

# 目錄

壹	·工具網址	3
貢	文本與種子詞彙	3
	- 、文本選擇	3
	二、種子詞彙的使用	4
參	· 文本與詞彙檢核	11
	一、詞彙增修與筆記	11
	二、文件编號與查找詞彙	12
肆	功能設定	13
	-、一般參數設定	13
	二、詞夾參數設定	13
	三、進度存取	13
	四、詞彙輸出	14
伍	· 進階運用	14
	、工作排程建議	14
	二、左右夾的策略	14
	三、巨集的策略	16
	四、筆記的策略	19
陸	· 練習用文本範例	20
	- > DocuXML	20
	二、UTF-8 純文字詞彙表	20

## 壹、工具網址

<u>https://docusky.org.tw/DocuSky/docuTools/TermClipper/TermClipper2020.html</u> 工具頁面如圖1所示:

文本欄詞 2020 版 (v0.22) + 詳量 詳量 → 詳史 調夾機具: 3(2-5)2 依据决:0, 领路阴囊:0, 已编取阴囊

## 請點我輸入文本與種子詞彙

#### 圖1 文本 詞首頁

## 貳、文本與種子詞彙

本工具目的是提供使用者利用已知的種子詞彙在文本中自動探勘並擷取更多相 關詞彙,以利後續的研究使用,例如進行文本詞彙標記。

#### 一、文本選擇

首先點取圖1反藍字體(點我),即可進入圖2文本選擇與載入的頁面。文本設定主要有四種檔案載入方式,如下:

- 由 DocuSky 取得文本:由雲端取得所需文本,可由 DocuSky 的個人資料庫取 得文本,或可自系統公開庫清單中載入文本。
- 2. 本地端 DocuXml 檔: 選取電腦中的 Xml 檔文本。
- 3. 本地端 UTF-8 純文字檔
- 4. 直接將文本貼入文字框

編詞 2020 版 (v0.22)		每人文本科核子 拆洗→拆置 封筆→拆井
	設定文本、種子詞曲與現亮參數	S
可 Mat 位 DocuSky 影得文本 · · · · · · · · · · · · · · · · · · ·	DoouXmi 欄・ CCC 購入未培園 UTF-B 文字幅・ 姫奈	直接的文本面记员以下文字程
文本内容		
· · · · · · · · · · · · · · · · · · ·		
· · · · · · · · · · · · · · · · · · ·	· 》 · 》 · · 》 · · · · · · · · · · · · ·	
景高調算系統中的有許表與約集: ○第 高		
MEANING WATCHING CONTRACTOR	■2 (第入文本→10月1日日)	

圖 2 文本選擇與載入

### 二、種子詞彙的使用

在文本中欲探勘並擷取相關詞彙,需先提供可用以學習的種子詞彙,方法如下:

(一)直接輸入法

於圖 3 紅框處,直接輸入已知的種子詞彙,一次可輸入多個詞彙,中間須以半形分號(;)隔開。例如:大雨;地震。

1000 HE 1.0.	約31文字。第2指集團部中會制
可 NIL 使 DecuSe	- 取得文本・ 国民議入本地場 DesuXmi 幅: 「「購入本地鏈 UIF
	stol saar Hennes- Jah Garmer Arestal, he has a0007-2 all Que Merkalles tri, name, p0001-taans-cito-days
文本內容 118—	マー第「後・安 スカ」、学校などの時期、単立工具集合、第2年、2015年、第2年、2015年、第2年、2015年、第2日は高校市場の開催・第2年後、第2日英国国家集合・第人会市・村市業業・第日2年 
<b>《文王记》从多祖君的</b>	anna
種子詞章 (片味入)(量	HARRING INT MARKARING - BAZMANNE - BAT / J.R. U.B
是古清除系统中州有	A先期回義: ○月 首 第7 (8) (5 × 119)(5 × 119

圖3直接輸入種子詞彙

點選確定(載入文本→計算詞夾),結果頁面,如圖4。



#### 圖 4 直接輸入種子詞彙後的結果頁面

1.候選夾

藉由所輸入的種子詞彙,工具於候選夾中列出夾住種子詞彙的左右詞夾,使 用者可於候選夾中進行篩選,以決定採用的詞夾。候選夾列表可依字母排序或依 系統排序,如圖 5。



#### 圖5候選詞夾

可點選候選詞夾中查看該詞夾在文本中的內文片段,以確認是否夾有相關詞 彙,橘色標示字體即為利用詞夾「\_...水五」所夾出之新詞彙,如圖 6。依此決定 該詞夾要加入種子詞夾或丟棄為無效詞夾,如圖 7。此外也可直接於詞夾檢視畫 面,新增需求詞彙。該檢視畫面可顯示該詞夾所夾詞彙,可直接新增於擷取詞彙 中,即使非該詞夾所夾詞彙,於檢視畫面中看到,亦可直接新增至擷取詞彙中, 如圖 8。

#10 5050 #i h-h-650	<b>股定备</b> 第4	RR · 台以建用的用于农业的资金		
ARRA D	四日月日日	10人的排子的声		
• A	00			
* A				
*1	08			
1.25	08			100
1	文本中日	含此調量或與來的片間(2)		
• 1 (doc#1804), /span-	-	and the second se		
・ + ト六月 - 庚寅 + 儲江)	11 日本に 林田日 - 三軍民衆法	三千根・シン		
2.1000年4月2日 副5		旧江者一丈・同多道死・トト		
*1.8t	08			
* 4	08			
● 4	88			
*1\$N	00			
*1	00			
a L. R.M.	05			
41.5月間	00			
ALLESS .				
			The second s	-
Matthe T	R: REAR UNAN	23 1 30 - 10, <del>2</del> 0	N-SIIR (	- 11



文本	AND DOOR HE LOD DOOL	RA	古湖洞夫,呈以重用的横	<b>7 天理</b> 取於 <b>第</b>	Personal Statements	
文件編 5 Zai/Y *NO_1 1 Zai/Y 十一年 大水」 書・			• 1#X	#78,*	五東町地道洪央	802
	新聞員內 1 在3	N HARR II	(名前)(第):20 名法 (名文 (1)前)(2)	A140	Rink (	

圖 7 加入種子詞夾

-	I <b>II</b> 2020	M (v0.22)				新大文本創版? 同次————————————————————————————————————	
-			<b>拉克</b> 贝萨	<b>胡夫 - 菲以書用的種子共取調</b> 量	-	Contraction of the second	
		G GART		2 3/108-88		2 5年の無い対応	(***)
1.54			おもう	二期上,上下文统权(围垦:505			
1.207	+	1. (doc#24) p002438	11. 11. 11. 11. 11. 11. 11. 11. 11. 11.		21-2月代・-		
**.	* # 31	2. (doc#57) 1/_p00578	運業五行出地学会元5年~和常 2011年夏季に日本に得るい2月		*847++		
2.31	* EE	4 (由10年1月) 55日後漢慶	日付吉賀希永和3年~三年、開幕	1、大如陽子,清陽	DEGENA * 8. 1		
2.ZaN		% (doc#91) 37_p00598	建富五位石炭带元的4年。光初	SZYATS	LENS 1	63	
#S3	* <b>A</b> E	6. (doc#61) 51 # 2 # 2 # 2 7. (doc#61) _00062 # 3	15日間中国第4年1日内国際の 副五日市地帯電費7年1七天五				
	* E 11	8 (doc#63) VI p00638	#東五行古篇句譜巻4日 - 四王	ER · RE E CERTE			
A 14	# 文判	<ol> <li>(doc#145) 後端書五?</li> <li>西ボー 換算尺</li> </ol>	2回農物連帯2年に営物課ディン	ERASE - RE COURSE	RSBN+BELEAR	「波石着位の展江」まで出た	
0.51	• 234	10. (doc#267) . ZaPA_p0	267届春三月百余月春前4年十日	法七代 國際 化清极一素	·····································	· ·推祥·已至你居所算·封	
**-	*+=	11. (doc#267) 封张国政	- 非非利用公子证明任用意问题:	1、運動管理 ( 使用用用		195月前:日第一20日	
4.12	(*=/\	12 (doc#273) 0273	5月世武帝东 <b>李</b> 5年。五平五月7	x-48-38 - 02-3	不安くした毎年に		
4.Zalv	1	13. (doc#374) - Zaévi_p0	274發展五行地就常居事5在世界		防衛軍・		
八庄田		14. (mo#275) 着着五行5	11年9日第5年,六月是古《田市	- 第十・第四・世界 王) 四第二	1		

圖 8 詞夾檢視畫面,可新增發現的詞彙

2.候選詞彙

選擇好種子詞夾,於候選夾中按確定(計算詞彙),即可自動擷取相關詞彙。 點選候選詞彙,即可看到已擷取之詞彙列表,一樣可透過檢視內文片段,決定是 否將該詞彙加入選用詞彙或廢棄詞彙,並可再依選用詞彙計算新詞夾,如圖9。 於詞彙檢視畫面中,若有發現新詞彙,也可在該畫面就加入擷取詞彙中,如圖10。



圖9候選詞彙



圖 10 詞彙檢視畫面,可新增詞彙

3.已擷取詞彙

為擷取到並經使用者選用的詞彙結果列表。使用者仍可在此區進行結果檢視 與增刪。

(二)巨集使用法

所謂的巨集,即是前人建置且公開分享的詞庫,或利用演算法獲得的詞庫,以及

使用者個人自行整理的詞彙表。現行工具上包含有二十六種巨集,包括時間、數量、 人名、地名、醫病藥等相關詞庫。

使用巨集時,只要點選**巨集**,打開支援的巨集列表,如圖 11,將欲使用的巨集(縮 寫)名稱複製於種子詞彙輸入欄中即可,如:#數字#。如欲使用多個巨集,亦以半形分 號區隔。

若無合適的巨集,可選取自訂的巨集,只要在種子詞彙輸入欄中填入「#UDEF#」, 即可選擇個人自行整理的詞彙檔案(UTF-8 純文字檔,一詞彙一行),如圖 12。

_		支援的巨集編寫 (macro) 編要說朝		
Nd.	巨魚 (細丸)	10 40 50 FM	完整表 建式	
1	#中期#	通用於部設備代以他的中間日期。例如「描太宗元年十二月廿日」、「勉強三十五年參月」、 「光緒七年」等。	an I	
2	#年(月日)#	某某年,後方可加上月日(可比對出千支紀日,例如甲午年,廣涨年等)。	1520	土用用,均能已用,用
3	#干支#	同比對出六十千支	植植	
4	a Brijka	可比對出中文數目字,例如「九千靈百二十三」,「七十六」等	建設	
ð	#數量或含儲#	可比對出數量或全額。例如「董平會百醇排伍兩伍幾伍分」。「豐軟二分」等	植植	
6	<b>#蘇州國#</b>	可比對蘇州碼(含單位)的字串、倒如「川++   兰兰 十石」、「三十六石 六斗 一升 八合 八勺」、「36.8188石」等	1273	通信公共建立制 - 刊口
7	#UDEF#	可提取本地硬碳的 .cov 或 .txt 詞彙檔來進行比對(檔案編碼格式要為 UTF-8 · 每一行重但 含一個詞彙)	19.01	
8	#人名_cbdb#	可比對出 cbdb (中國歷代人物傳記資料庫) 中的人名。由於 cbdb 包含有數十萬個人名。比 對可能其時報久。	28.25	WER . LAN
9	#人名_dia#	可比對出法該山人名福威權所包含的人名。由於該權威權所包含上萬個人名,比對可能耗詳較 久 +	19.21	
10	#人名_thd權威權#	可比對出 THOL (臺灣歷史數位圖書數) 人名福威福 (來源包含故至清代人名福威福 · 圖圖明 清人物小傳 · 文建會臺灣歷史辭典) 約人名	1921	
11	#地名_thdl演新檔案#	可比對出 THDL(臺灣歷史數位圖書藝)清新檔案所使用別的地名	機構	
	五年十八十年	Endblind Stream Constants on the South Long United Streams (Streams) Constants (Streams)) 王朝教城	and the local data	
Ŧ	國 印刷人间象 不由开始分	# (F) 約玉晴長隆式集 <u>二集</u> ・多式2度以分類: 陽間( 大売, 竹葺		
1	除赤肌中所有药夹肉间量	- 02 B		
		## 1#//X+ ** 11#//S/		

圖 11 系統支援之巨集(縮寫)列表

HILL SOOR HILLS	201. Million Control Announces Announces	-
	<b>設定文本、唯子調量與調英參設</b>	-
村 2月前 提 DocuSk	y 教授文本 · []]]建入本地理 Doculani 權 · []]]讓入本地理 UTF-II 文字權 · 國產產種將文本與點於以下文字框 ·	г.
	7月里,黑手,那些小熊生,这个个钢笔,和笔符开放银脚桨,都将开放嘴脸,就是有五子	Ŀ
	aana Alarbi sayan filmame Tahi Quantikefulmen (ri Jaarius, 2000) (Dahi Quantikefulmen (u Jaarius, 2000) (Jaarius (du 100)	Г
1	文学报三百年,大观,曹操不说三十五日,就从山市这一世九五期前。武,道民里八千桥时,称三百胜人,朱度,战人胜臣学口波指帝臣,兵上立理除五帝臣,世法用道,以武臣代,臣 先上帝,最后常期,政治这一智能,要服,来二致,是曰,兰帝者公之臣御子,指遣官僚,何以宫侯,何故之道,则称王帝有敌人,其诸臣举行过过国。	L
	eans cp-6, coper feesame+ Tal*D,uetMertulines.bt, markus_p000F17a/*D-antimetulines.bt; markus_p000F10paro-cb-doco	ŀ
7500	力型 人生生化量法化、人力、我们、法用、规范基、使用用化用、法介介的人、九油、本有可用用用器用、自用化用处法、肥大上等、甲基甲基用、脱制用等、脱脂以为造工机、化固 工具用量。	L
178	nera 1927 - sigar Nasara-Zetti Quaddenhilawini, mahar, 2000 ti (Jah Quadwelalman bit mahar, p022 ti igano cito deri 48 28 - 28 - 3.5 - 28	l
	##2_1##2_###173#LevelABD/##A_#E%#2_##178#_1#2_##0_##_1##2##2###2##2##2##2##2##2##2##2##2###2####	L
	29 29 - АДХИ, ИМЕЛЛЕ, НАТАБАЛЕ, ХАКИ, АТЕН-КОНИ.	L
	anna dir B. regan Renamer 7a R Cantiller Martin, anna, 2000 - 247 Cantiller Adment Martin, 2000 - 1440 - 1610 - 1610	L
-	Sarr A - Michigan	
種子脱棄 印刷入30里	TANKOR (J) NERABRATE SKERLIGE BED PLOY LERV	L
展开通信来经中所有	「「「「「「」」」「「」」「「」」」「「」」」「「」」」」「「」」」」」	L
	#〒 (単人文本→1)〒22本)	

圖 12 自訂巨集輸入格式

1.展開巨集

輸入巨集並點選**確定**後,在候選詞彙中,可於選用詞彙欄看到藉由巨集所擷取到 的詞彙,點選展開即可全部列出,如圖 13、14。

2020 🎬 (vn 22)	200			MARKING STREET	
C sans		A LOUP SUR C		anus -	
					MIRS
	11078 (RB	196104 ; 100-100 (		-	
	2020 मि (M. 22)	909) के (40.29) No (10) (10) (10) (10) (10) (10) (10) (10)	2020 SE (40.22) MER ARAR + BOURRAUMERAN F SAME		

圖 13 巨集展開前

文本編集	1 2020 10 (vn 22)				CONTRACTOR CONTRACTOR	C STREET
		観測	5個記彙・言以凱用記彙作為描子・3	解新的装置夹		×
219.001	C HARR		States B			
\$ 289 *N0_1 1. ZaN +			<ul> <li>東風</li> <li>東風</li> <li>東風</li> <li>東風</li> <li>東風</li> <li>東町</li> <li>第四</li> <li>第四</li> <li>第四</li> <li>第四</li> </ul>			<b>NH</b> 42
			MERK 2 BEAMER 1		-	
			WE LEWIS WE WAS	SEMIRAL S		

圖 14 巨集展開後

2. 不展開巨集

如想區隔那些是來自巨集的詞彙,那些是新擷取到的詞彙,可選擇不展開選用 詞彙中的巨集結果。以在文本中擷取到的新詞彙「河竭」為例,在未展開巨集的情 況下,可以清楚看出為新擷取的詞彙,如圖 15。

文本開設	2 9020 BE (v0.22)					
	eonomicon e asse	観定	B期职业·世以通用职业作品相子。	计算新的保護表	ARMAN AND A CONTRACT	×
文作編 5 Zally *NO_1 1. Zal 用版王 二年・ 二年・ 二年・ 3	C waxe		* ALDEF, SLAA * ALDEF, SLAA * RM			Ellig R2
		R-HIGH GR	111년 : 1811년)		EA.	

圖 15 選用詞彙中新擷取詞彙

所有擷取到的詞彙,都會在文本中標示,從巨集擷取到的標示為黃色,新擷取 到的詞彙則標示為綠色,如圖 16。

文本編詞 2020 版 (v0.22)	MAXERIAN DIRIGHT DIRIGHT
文件编辑: 1 -10 7 8 9 10 10 10 14 16 16 40 8770 211	1月1日時期 2日-8日 (朝鮮一日)(三日) (明治(11日日) - 11日日日) (明治(11日日) - 11日日) - 11日日)
\$ ZahnGuanWeerfulnew.txt_merkus_p0011 § Dimministra: 11	查找回来: #N (##)
"NO_TITLE"	
11. ZalYiGuariWenfulinew.txt_markus_p0011	
用業王	
二年 - 周三川県 - 新向以為会木 - 2 2 章士者也 - 伯陽甫曰 : 「周時亡矣 ! 天地之氣不成其序 : 老嶋	R序・民亂之也・陽伏茂不亂出・陽原而不振升・於是有

#### 圖 16 文本標示已擷取詞彙

3.新增種子詞彙或巨集

在擷取詞彙的過程中,可利用候選詞彙中的新增詞彙功能新增詞彙或是巨集,多個詞彙或多個巨集以半形分號隔開,如圖17。

文本相同	2020 🛱 (vn 22)				AND THE REAL PROPERTY.	WARRANTING INCOME	
	- Storing and a story	1212	8-100-100 · 10-0-00.0	1月累作為種子 - 計	算新的候選夫		
文作編 \$ ZallY *NO_1 1. ZalY +3 2.巻・	• • • • • • • • • • • • • • • • • • •		₩ ALGEF_ELR	C Andre		A REAR	M.Rec.
			MELA ; BLAND	2) M2 (MAR	1804		-

圖 17 新增種子詞彙或巨集

## **參、文本與詞彙檢核**

本工具提供直觀的文本與詞彙檢核功能,使用者可於工具頁面就載入的文本,細緻檢核詞彙。可增修已擷取詞彙、可為已擷取詞彙加注筆記,也可查找 文本中的任意詞彙。

## 一、詞彙增修與筆記

使用者可透過工具頁面在文本中直接檢視已擷取的詞彙,如發現有未擷取 之新詞彙,如:山崩,將其選取後,可於小視窗中,將該詞加入已擷取詞彙 中,並可註解該詞筆記。如圖 18。

MINIT # 2020 15 (v0.22			和人士主动称	1. HAL-HA HA-HAL O
eitige I. OR H. H	25 1 16 17 18 19 20	21 21 (40) \$P\$	第三部長:20-5(2 新建市:MAX:新建	· 伊斯
\$ Zarn p0017 \$ 118 1 M 18	16		#HILK	· 家務 [1919]
"NO_TITLE"				
11.2.2.2.1.1.2.2				
17.2aW_p0017				
成公五年「夏,漢11				
■ ・製業券回搬用三日平面	的时期建筑量"山田"?	● 常田・圓陽・田田・天昭常日・紫癜	自環、下進、百姓称失其所分	県、天然県高・野亡象也・第山省
着老・殺害部市及大干也・狭 二節	-			
・廣公以此・演绎之會・天下	11.47.1			
D*	加入結果者			
「「「「「「「」」」」「「「」」」」「「」」」「「」」」」「「」」」」」」」」	三支活動的、第一型集王室・蘭	中野語見れ、東京に為早山、豊富也三南(他相ど	(、古者三代命紀、蔡不趙富	(+南田親藤 (不離廣意)
NI CONTRACTOR				

如發現已擷取詞彙的錯誤或特例,也可選取該已擷取詞彙,在小視窗中將 該詞從已擷取詞彙中移除,或註解該詞筆記。如圖 19。

■頃工具 2020 EF (v0.22)	BA335	
文件编辑表: 1 - 20 14 - 24 - 26 - 36 - 19 - 20 - 21 - 22 - 430 4966 15.00	総合規則、第14日 修道県: MAX: 修道部	100-103 1
\$Zahi phot2 § 闭塞出限密数 10	吉托均衡	A.R. (2.4)
*NO_TITLE*		
17. ZeVi, p0017		
#公五年「夏·萍山	C	
2 。 劉梁瓚曰戰元三日不迭。官官前前對臣復贤之,八尚十劉阳以為山陽,氣也, 劉指及天下也,遂官募組	DEFENSION CH. T	「「「「「「「「」」」、「「」」、「「」」、「「」」、「」」、「」」、「」、「」
三時.	STER-END BY	
) 縄会以起。漁県之會、天下	20W:42A6mi 01	
R		
加勒國政,其後後,憲法連載;三席這會說,單一即是王室,筆件群說範則。與取以為梁山,要望出		144,不是通信。
RE.		
山川・四川〇山川・二之祭也・黄冠馬の道・登集成在第四・至十七年道王男田・黄書・中行道	和國公孫立師会。	

圖 19 擷取詞彙的移除與筆記

## 二、文件編號與查找詞彙

有些文本規模較大,在原 DocuXML 檔中已有分件,工具會依文件編號序區分頁 籤,使用者可直接點選頁籤查看指定的文件,如圖 20。

文本編詞 2020 版 (v0.22)	输入文本目錄子 現良 一 計算 一 計算 ( )			
文件编辑 1 -10 💶 2 3 4 8 8 7 8 9 10 +10 8776 732	RAMA 32-52 WER 254 WERE	MIN-DALL IN 10 - EMILIAN I D		
§ZaPrQuantitentutivees.txt_markum_p0001 § 已開放開業在本業型的出現次数:0	度找到景:	RN [8:8]		
*NO_TITLE* 1. ZaiYiGuariWorkJinow.td_markus_p0001 十一年「数・余 大力」 - 量評問以為時費、米比年為果正、第之間、百姓怒怒、魔氣盛、飯二爾慎水、劉内以為時来際公績優 墨。	· 建贝不应 · 新年編其臣主董博島	1、編人在指 - 种型尾翼 - 尾裂公之		

#### 圖 20 依文件編號區分頁籤

也可於「查找詞彙」處輸入欲查找之詞彙,查找結果除以[文件數:詞彙數]顯示於 後外,並會在文本中將該詞彙以顏色標示,且顯示該詞彙在文本中出現的次序編號。 持續點選**查詢**鍵,可依序追踪文本中的該詞彙;點選[文件數:詞彙數],將啟動小視窗, 依件羅列查找結果,直接點選指定件,可呈現該件文本及查找之詞彙,如圖 21。

· 建碱工具 2020 世 (v0.22)	输入支出的	17. 2	18-318 HA-318	
THANK I WE A JET A M A JE A A A A A A AN AN	WARE SO - MIRES	91.00	88-95 8	
\$7.8% p0005\$ \$\$\$#######2	唐氏消费: ×31		(RENDEN) (MER	
.w5'zurt.		247	Zoh pana	
5. 24% potes		245 251 251	Zahi pilati Zahi pilati Zahi pilati Zahi pilati	
文帝總三年秋 (111) 曹清不經三十五日。蜀田山 (111)、西九百时度。爰、唐茂家八千钟所。昭三百部人。先是,胡人即迎	平口菜粥用用。為上立	153.	2010 (413)	1
然上作,我科雅族,这族祖、祭稷。贾敏,死三族,是约,比两者		281	Zahtpattis	1
(名言)		258	Table office	14
-此果子。納達花裡,每双命間,但但此情,納斯不定補除人,進速於單位計定值。		254 255 264	Zahl pelvel Zahl pelvel Zahl pelvel Zahl pelvel Zahl pelvel Zahl som	
Design of the de				

圖 21 查找詞彙功能

本工具提供相關功能的設定,如圖 22。

<b>建築工具 20</b> 3	20 種 (v0.22)		林人北北的银石 资产一资产 的第一	
st Halling : 2	- 記念和設立 - 記念和設立 - ※ Donaday 最多期間文件編載(上版 200000): 2000	推开称		5
xmaE41	<ul> <li>(四川県県町市代設定:た所 2) (四県 2+ 5 石所 2)</li> <li>(四川県県町市代設定:第一四川県立小山県 1) (田村県町県</li> <li>(県知一級化:第三部中文研想之外の)(株別)(株) (都道)(広知時内)(株)</li> </ul>	(山): *並へ首 [編8]		四月一日
SE NMT-RH	<ul> <li>         ・</li></ul>	8287.00 (88.) (888.)		
##.\.318#	NAME - MEANE - NAMESIN	BE MANDA BRIAND REPORT	(8056)	

#### 圖 22 功能設定

#### 一、一般參數設定

使用者可自行決定欲載入的文本件數,上限為 200,000 件。

### 二、詞夾參數設定

詞夾模具長度,用來設定擷取詞彙之詞夾,可設定左夾、右夾及欲擷取詞彙的 字數,如為西文拼音文本,則為字元數。

詞夾模具的閥值,用來設定單一詞夾至少需出現多少個候選詞彙,才可成為候 選詞夾。

標點一致化,用來決定是否將文本中中文頓號之外的標點符號,都置換成特殊 符號(⊥),以利擷取詞彙時略過標點符號。

#### 三、進度存取

儲存,用來將使用者的操作進度儲存起來,但不包含文本內容。使用者可於

必要暫停作業時進行進度儲存。

載入,用來將之前的操作進度載入,使用者可利用此功能接續之前的作業進 度。

四、詞彙輸出

使用者可將工具操作結果,也就是所擷取到的詞彙匯出。。可單純匯出 UTF-8 純文字檔的詞彙表;也可選擇匯出含文件編號、詞彙、巨集及註解的 UTF-8 純文字檔。

#### 伍、進階運用

以下為本工具各種詞彙擷取方法的策略運用,使用者可依據文本性質及欲擷 取詞彙的詞類及特性,參考並組合運用:

#### 一、工作排程建議

本工具可探勘多種詞性與詞類,包括名詞、動詞、形容詞、量詞…等,但建 議每一次工作排程以一種詞性的詞類為主,例如:名詞中的人名。使用者有可 能在一次工作排程中注意到其它詞性的詞類或同詞性的不同詞類,此時可另行 註記,以備其它工作排程的規劃與運用。

#### 二、左右夾的策略

(一)情境:欲擷詞彙的左右皆為標點符號,如:在歷史自然災異的文本中, 災異詞彙的左右通常都為標點符號,可能是逗號或是頓號甚至為段落首 字。

> 嘉靖三年六月,順天、保定、河間、徐州 蝗。隆慶三閏六月,山東

#### 早蝗。

操作策略:

1. 詞夾參數設定 - 標點一致化,將標點符號皆置換為「⊥」

- 親友委教論室
● 諾克機與的長度最差: 拉克 2 試量 1- 4 石灰 2
• 諾貝權典的豐富設定: 第一該與至少出現 1 簡保遵訊彙
● 總點一致化:是否將中文編號之外的模點消號。朝實換以特殊符號(⊥): ●量 ○否

 左右詞夾輸入 - 左夾輸入「⊥」; 右夾輸入「⊥」; 設定詞彙最長字 數為「4」

-		27.E.M	NTER I STORAGE	2.000 p.#	
	18.9820共	教徒学母亲用	11人的相	1738A	古莱的普及湖中
C		and the second states	la selecter a second	1	
	期增訊與人可僅輸入在於成者	(病):在約: 詞列	<b>用用(表示因素)</b> 14	(1月) 新規20	<b>R</b> .2

3. 篩選詞彙 - 將災異詞彙加入選用詞彙中。

C HRITER	心然而成的冲 二回日	REIR STR		
<ul> <li>大衛用</li> <li>大衛用、米</li> <li>大衛用、米</li> <li>大衛用県</li> <li>大衛用県</li> <li>大衛</li> <li>大衛</li> <li>大衛</li> <li>大衛</li> <li>大衛</li> <li>大衛</li> <li>大衛</li> <li>大衛</li> <li>大都</li> <li>大都</li> <li>大都</li> <li>大衛</li> <li>大衛</li> <li>大衛</li> <li>大衛</li> <li>大衛</li> </ul>		*/= */= */= */= */=	<ul> <li>○</li> <li>○<th></th></li></ul>	
		WARE AND IN THE REAL PROPERTY OF	COLORES .	

(二)情境:欲擷詞彙的左或右銜接有特定詞彙,如:在歷史自然災異文本 中,災異詞彙前通常銜接有「大」,例:大「雨水」。

> 十一年六月,渾河 溢 固安。兩畿、浙江、河南俱連月天雨水。

操作策略:

1. 左夾輸入「大」; 右夾空白; 設定詞彙最長字數為「4」。

18.30	58. HE 6.45. T31.	117.6M2	1月1日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日日		ing in large
	a naza	()(1349)(L <del>.)</del> * 7_	100 A KANG 7 2000	<b>60</b>	20
(	新規時用(等価額入在用産8	(利) 主在和 大	(廣西四美):4 西北	1/ 1025( <b>R</b>	

2. 篩選詞彙 - 將災異詞彙加入選用詞彙中。

T.B. 3030 H 6.0.33	基定各质的	1、並以銀川約県作為柚子。21	算新的保護兵		
	247688 100	1 电相对来		■ 每果时来	-
□ 九			<ul> <li>1</li> <li>1</li></ul>		915 915 915 915 915 915 915 915 915 915
-	新雄語業(開業開以)調	M 3 2		-	
		ER COMMANY   Lang COMM	arman .		

## 三、巨集的策略

(一)情境:使用者有二份種子詞彙表,可能為自行整理,或由他方取得,欲利用來 在同一份文本中撷取更多相關詞彙。

操作策略:使用自訂巨集#UDEF#雨次。

 除在輸入種子詞彙處使用自訂巨集外,在工作過程,可在候選詞彙的新增 詞彙功能中,再使用一次自訂巨集。

文志開日	1 2020 Mi Ivo 221				CONTRACTOR OF CONTRACTOR	0 2010
288 \$280 4NO_T 1.220 +-8 28 28	<b>О</b> кана	83. (XII.7900) (2000)	Faller a disentituter(A-Alitzia) ■ statute + state: Jama IIII	C	E SKAS	300.0
2		-	教は、首用) 「 AUCHY JERLA			-

2. 在選用詞彙中會分別有使用二次自訂巨集的擷取結果。

24.00	2020 <b>W</b> (vn.22)	1				CONCERNMENT OF TAXABLE	0 1000
交片線 \$.2m *NO_1 1.2m 用級王 二年 - 2年 - 第 - 38	asua		* R.COT. U.R. * R.COT. U.S.		05	E CRIAR	intz
		NUMBER OF	and a mana	ei az iner	armaner:	**	

(二)情境:在文本敘述中,會有二種詞類相依出現的情況,例如:災異伴隨 著災損,「淫雨」「伤稼」。



操作策略:使用#REPLTERM\_勺#, 替代文字有六個:X;Y;Z;勺;攵; □

 當有二種詞類相依,使用者有其中一種詞類的詞彙表,可在種子詞彙 中輸入「#REPLTERM\_勹#」,選用一份詞彙表,將文本中所有該詞彙表 中的詞彙都以勹替代。例如:我有一份災異詞彙表,將文本中所有的 災異詞彙皆以勹替代。

27/16 27/16 21	構成素 - ●●●■本本素結構 Docuteri ■ ●●■本本結構 (17-8 文学 bol, stpar flienaes:2a(Yl_2000)*Sa(Y_2000)+dapase(Ne-bri> ==本本/教、完 (大会別、特徴現象理論、完美公定論。 == bol, stpar flienaes:2a(Yl_2000)*Sa(Yl_2000)+dapase(Ne-bri> em三年夏、温中、実際 次、本規模で抽象。 = p)、sapar flienaes:2a(Yl_2000)*Sa(YL_2000)+dapase(Ne-bri>		2、東中国共民主義指載。 (19) 100 (10) (10) (10) 100 (10) (10) (10) 100 (10) (10) (10) 100 (10) (10) (10) (10) 100 (10) (10) (10) (10) (10) (10) (
227/16 101	bit, sppn filename:2aiYi_0001's2aiYi_0001's2pprov(%-driver) 年 「秋、東 木、 (算件時に30時間、完計(年本県王、前之間、前性型道・開発層・台 人在前、外目間尾、高級公之選。 bit, sppn filename:2aiYi_0000's2aiYi_00002s4pprov(%-driver) s市、本出版四千餘家。 p3, sppn filename:2aiYi_0000's2aiYi_0001s4pprov(ho-driver)		2、 明日田王臣王王 州都 - 
24716 10	生「我、宝 (木」、董牛好に為4時勤、王比算為處正、罷之戰、罰性怒怒、關果語、並 (人在男、外后軍萬、萬時公之道。) 		21 - 17 - 17 - 17 - 17 - 17 - 17 - 17 -
24/16 101	1957. supus filesumes Zalfi_p0002*sZalfi_p0		0 4 3 4 10 4 3 4 10 5 4 5 10 5 5 10 5 10 5 10 5 10 5 10 5 10 5
22/16 (#2005/04/04/04/04/04/04/04/04/04/04/04/04/04/	e時三年夏・国中・美閣 1本・本出現四千種家。 	Contraction of the second seco	1000 00 00 00 00 00 00 00 00 00 00 00 00
1		a state	
7	時代:可説 (本・伊・福祉千六語論家・法共同八語論家・ 		
戦場工具 2020 垣 (い	0.32)	<b>NACES</b>	a constan constan
THE COLOR	51 12 11 12 15 16 17 11 11 -50 1738 111	Be48 -30-92	48-91. A
5.2mm p00343 200	Q 是E # 平田内11月21秋 #	DARR.	80 (FS)
-wo mie-			
34, ZaW (20034			
物得永元元年七月。都講 王 - 新 <mark>方 5</mark> 章杀人 - 已	1、 <mark>1</mark> 、小板、雪菜(用約)曰:「香港有利,收费起港,肥之 <mark>11</mark> , <mark>秋 13</mark> 日, 13别把土型,已除了餐;盐雾能将,接 <mark>15</mark> 条杀人,能将不餐,盐雾干燥、医15	·举人,而 <mark>为</mark> 。 <mark>为</mark> 《天言》说道不信,故谓秦 五山不说, <mark>为</mark> 除不能,故谓御宗,便 <mark>为</mark> 此入I	· 医 <mark>· 3 )5</mark> 米人,伊朗和维。1 6月,阳 <mark>5</mark> 永山,,世时,和

 再利用災異與災損的相依性,設定左詞夾為勺,請工具找出最長十個 字的災損詞彙。

TRAN	00.00.00	2268	IRA - 世纪原用的相子主要的现象。		
	E Harten		■ 和人名利弗子利利	Edination	10
(		3月):在称: y 第月載	教(条相記集)(54) 正司 (現年)(1986年)	RIETA :	-

3. 從候選詞彙中,將災損詞彙加入選用詞彙中。

TE MINISAN IN		「非常的意くまに満定的条件の場子」と	REDUCT			 r.
		○計算算法(当該)(注意)(注意)(注意)(注意)(注意)(注意)(注意)(注意)(注意)(注意		* * * * * * * * * * * * * * * * * * *	<b>.</b>	12 元
- 100 - 道王 - 光道平田人 - 光道照 - 光道照 - 光道照人				<ul> <li>月里</li> <li>子は</li> <li>前前</li> <li>二</li> &lt;</ul>		
	目1月2日第(7日第	(max:mm) (		2.8		
		HR (119209) - HR (1187	ATRACK -			

## 四、筆記的策略

(一)同詞異義的處理,不移除,以筆記方式註解,提醒標記時應進行消歧, 例如:崩。

離除工具 2020 線 (+0.22)		TRACE BUT
Consider ( 11) ( 14) ( 14 ) ( 14 ) ( 14 ) ( 14 ) ( 14 ) ( 14 ) ( 14 ) ( 14 ) ( 14 ) ( 14 ) ( 14 ) ( 14 ) ( 14 )	204 668 - 20 (2) 60 80 - 100 - 80	48-90 8
KZAN JODIZY DOMESTICAN IN	81/1#:	
and mure.		
17. Zalvi_Listi17		
▲公五年「夏・奈山 間」、市営権回避地三百不須・資産原料協売地之・75項・美市以及山際・泰世・ <mark>間</mark> 除・校性・大 開発五天下坊・奈美尊厳 三年	enii. naga · re · nunzuna	· 其然展览,整门事任,是以石榴的,即量
Reality and a constant of the second se	A	
10年間の、11月前、高生茶館、三市道書紙、単、中和王家、著の世記和同・家師に為井山、首切の	「「「「「「」」の「「」」の「「」」の「「」」の「「」」の「「」」の「」」の「	創設時編・不望清白・
RE	THAT IS . 1	
WALLED AND A PARTY AND A REAL PARTY OF A PAR		

(二)異詞同義的處理,以筆記的方式註解,提醒標記時應將其聚合,例如: 山崩與崩。

IEIRIT.# 202018 (v0.22)		64.5.8.0MP- 140	
RANNEL IL COLLER IN THE INCOLUNCE IN COLLERS	5x.88	21-51	49-90 A
SZAM BOUCES ADDRESS TO	<b>治</b> 1025年		
NO_IIIII.			
17.24H(_pmt7			
·美山五年 7夏(伊山			
🛄 。· 新田集日載可三日不由 - 專用的新探索供力 · JNA - 新向以加加用 · 表生 · 🚺保 · 代生 · F.*	(10日)日日 日日 日日 1日 1	NFE-另位直注-首:	このなく可以注意ないの意
加重起关节把,使做幕曲			
E			
- 廣田目標,進建2會,東下			
G.	A	Col Colorado	
·····································	· · · · · · · · · · · · · · · · · · ·	NE - NUME - 78	「清白」
	SITABULA 'B. F		
THI . TO BE . COME . BANKER . SANSTER . S. T.	Supplied and the second strength of the secon	100	
新闻·新闻·注意。7	1144		
Data and a second second			
	A CONTRACTOR OF A DESCRIPTION OF A DESCR		

## 陸、練習用文本範例

下載網址:<u>https://reurl.cc/WLdZK9</u>



文本:歷史自然災異記錄.xml

## 二、UTF-8 純文字詞彙表

種子詞彙表:災異詞彙表.txt