文本擷詞工具 2020 版

操作手冊

撰稿人：黃韋菱、洪一梅

撰稿時間：2020 年 11 月

目錄

壹、工具網址

https://docusky.org.tw/DocuSky/docuTools/TermClipper/TermClipper2020.html

工具頁面如圖 1 所示：

圖 1 文本擷詞首頁

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：

1. 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

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。

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。

圖 6 文本中包含此詞夾片段

圖 7 加入種子詞夾

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。

圖 11 系統支援之巨集(縮寫)列表

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。

圖 13 巨集展開前

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。

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。

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。

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。

圖 18 擷取詞彙的增修

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。

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。

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。

圖 21 查找詞彙功能

肆、功能設定

本工具提供相關功能的設定，如圖 22。

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

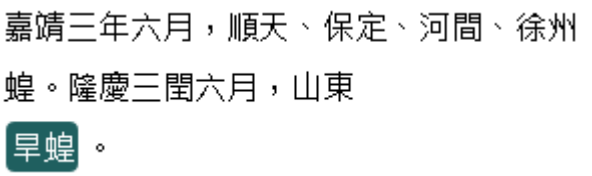
以下為本工具各種詞彙擷取方法的策略運用，使用者可依據文本性質及欲擷取詞彙的詞類及特性，參考並組合運用：

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。



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「⊥」

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

3. 篩選詞彙－將災異詞彙加入選用詞彙中。

(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

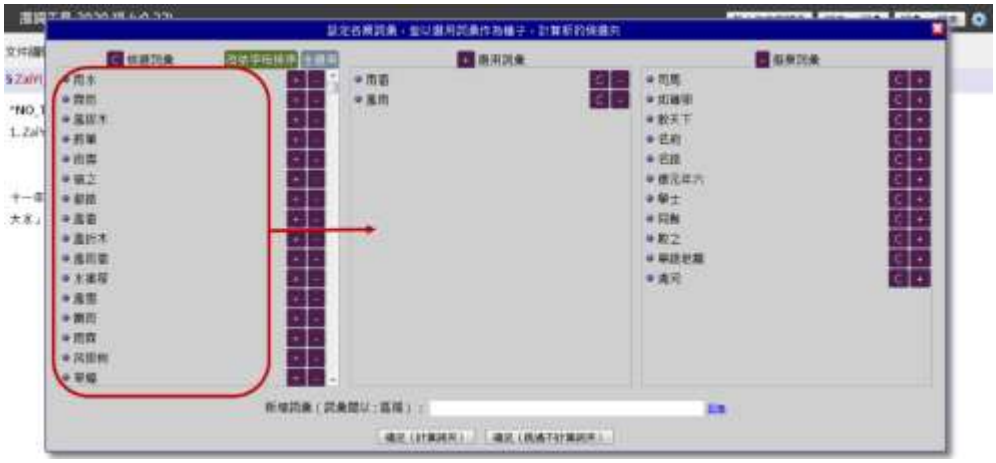
溢

固安。兩畿、浙江、河南俱連月大雨水。

操作策略：

1. 左夾輸入「大」；右夾空白；設定詞彙最長字數為「4」。

2. 篩選詞彙－將災異詞彙加入選用詞彙中。



三、巨集的策略

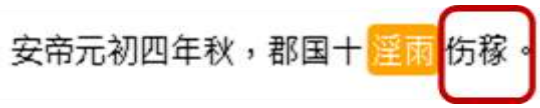
(一)情境：使用者有二份種子詞彙表，可能為自行整理，或由他方取得，欲利用來在同一份文本中擷取更多相關詞彙。

操作策略：使用自訂巨集#UDEF#兩次。

1. 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。

2. 在選用詞彙中會分別有使用二次自訂巨集的擷取結果。

(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_ㄅ#，替代文字有六個：X；Y；Z；ㄅ；ㄆ；ㄇ

1. 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_ㄅ#」，選用一份詞彙表，將文本中所有該詞彙表中的詞彙都以ㄅ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ㄅ替代。

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

3. 從候選詞彙中，將災損詞彙加入選用詞彙中。

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。

(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。

陸、練習用文本範例

下載網址：https://reurl.cc/WLdZK9

QRCode：

一、DocuXML

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