

標記與詞彙統計分析工具
(含 Palladio 視覺化工具)
操作手冊

目錄

目錄.....	1
一、 工具簡介、網址與範例檔下載.....	2
二、 首頁功能鍵介紹.....	2
三、 操作流程.....	3
1. START 開始.....	3
2. 載入文本或文獻集.....	3
(1) 方法一：上傳純文字.....	4
(2) 方法二：從 DOCUSKY 中上傳文獻集.....	4
3. TERMLIST 上傳詞彙表.....	5
4. ANALYSIS 開始分析.....	6
5. 統計結果與匯出.....	8
(1) BASIC TERM FREQUENCIES 詞彙詞頻統計.....	8
(2) CATEGORIZED FILE RESULT 依段落統計詞頻.....	9
(3) CATEGORIZED TERM RESULT 依詞彙類別統計詞頻.....	10
四、 匯出資料與視覺化呈現.....	11

一、工具簡介、網址與範例檔下載

標記與詞彙統計分析工具由謝博宇先生設計開發，是一款對文字資料進行詞彙統計的工具。當我們在進行文本研究的時候，常會好奇某些特定的字詞出現的頻率，藉由對這些語詞出現頻率的統計，來觀察作者對於某些概念的重視程度，或作者所要表達的意涵。

工具頁面如圖 1 所示：

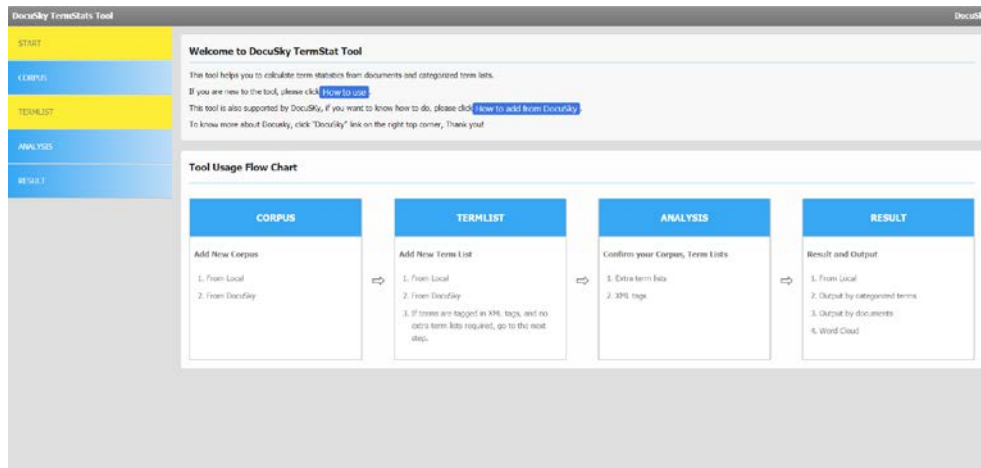


圖 1 工具首頁

本工具的網址為：

<https://docusky.org.tw/DocuSky/docuTools/TagStatsTool/index.html>

為了便於引導使用者使用本工具，使用者可以至

<https://tinyurl.com/y6nqbjyq>

下載本使用說明所使用的建庫檔「西遊記_DocuSky 範本.xml」，將此建庫檔在個人的 DocuSky 上建庫後，即可按本使用說明的操作方式依序操作各項功能。

二、 首頁功能鍵介紹

在工具頁面的左側是如何操作本工具的流程，分別為：「開始」(START)，「上傳文本或文獻集」(CORPUS)，「上傳詞彙表」(TERMLIST)，「開始分析」(ANALYSIS)，以及「結果呈現」(RESULT)。如果上傳的文本是經過「碼庫思(MARKUS):古籍半自動標記平台」或「批次標記工具」(ContentTagging Tool)標記的文本，則可省略「上傳詞彙表」的步驟，直接以標記過的詞彙進行分析。



圖 2 操作面板

三、 操作流程

1. START 開始

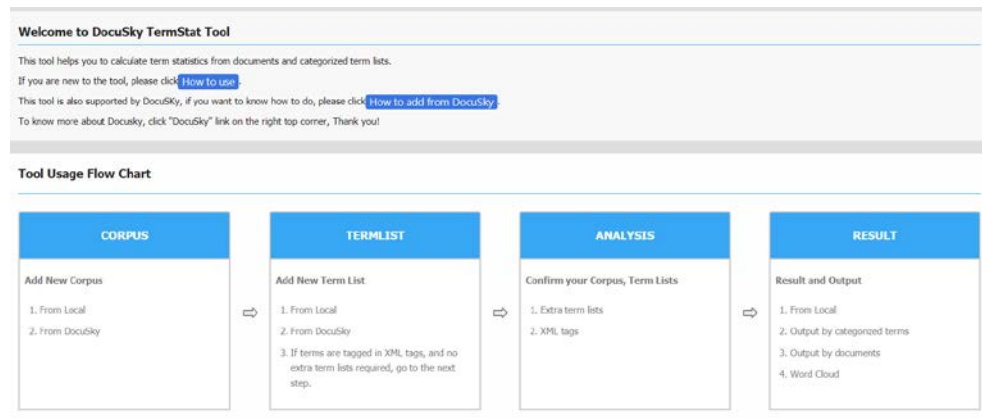


圖 3 進入工具後的頁面

在開始步驟中，有簡明的使用流程，使用者也可以直接點選 [How to use](#) 來連結本使用說明的 PDF 檔。關於如何由 DocuSky 載入文獻集，也會在本使用說明中說明。

2. 載入文本或文獻集

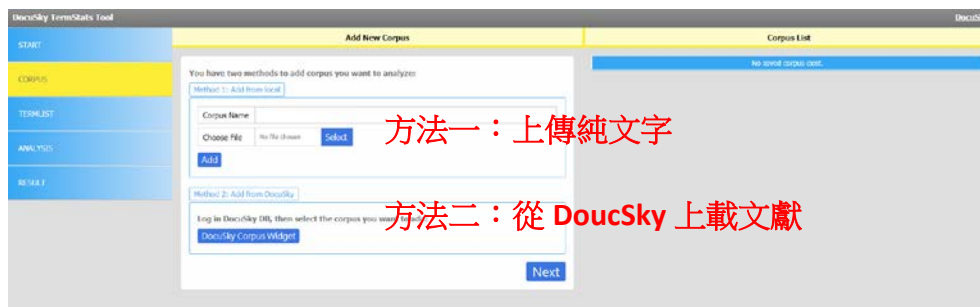


圖 4 載入文本或文獻集

點選左側的「CORPUS」按鈕，進入上傳文本或文獻集。本工具提供兩種文本上傳方式，一種是上傳純文字（txt 檔，UTF-8 格式）；另一種是從 DocuSky 中上傳文獻集。

(1) 方法一：上傳純文字


先在文獻集名稱（Corpus Name）中輸入自訂的文獻集名稱，然後點選 Select 按鈕開啟上傳檔案視窗。然後點選 Add 按鈕。

You have two methods to add corpus you want to analyze:

Method 1: Add from local

Corpus Name	西遊記	1. 輸入文獻集名稱
Choose File	Xiyuji.txt	2. 上傳純文字檔
	Select	3. 點選 Add 按鈕
Add		

圖 5 上傳純文字檔

此時，右方欄中將會出現文獻集的列表，若要刪除該份文獻集，則點選右側紅色的  圖示。使用者也可以一次上傳多份 txt 檔，並將多份檔案放在同一個文獻集當中，以利後續的詞頻統計分析。


Corpus List			
ID	Title	#Doc	Del
1	西遊記	100	

圖 6 完成文本載入

(2) 方法二：從 DocuSky 中上傳文獻集

點選 DocuSky Corpus Widget 按鈕連結 DocuSky 資料庫（需登入），然後點選載入即可將文本帶入工具中。

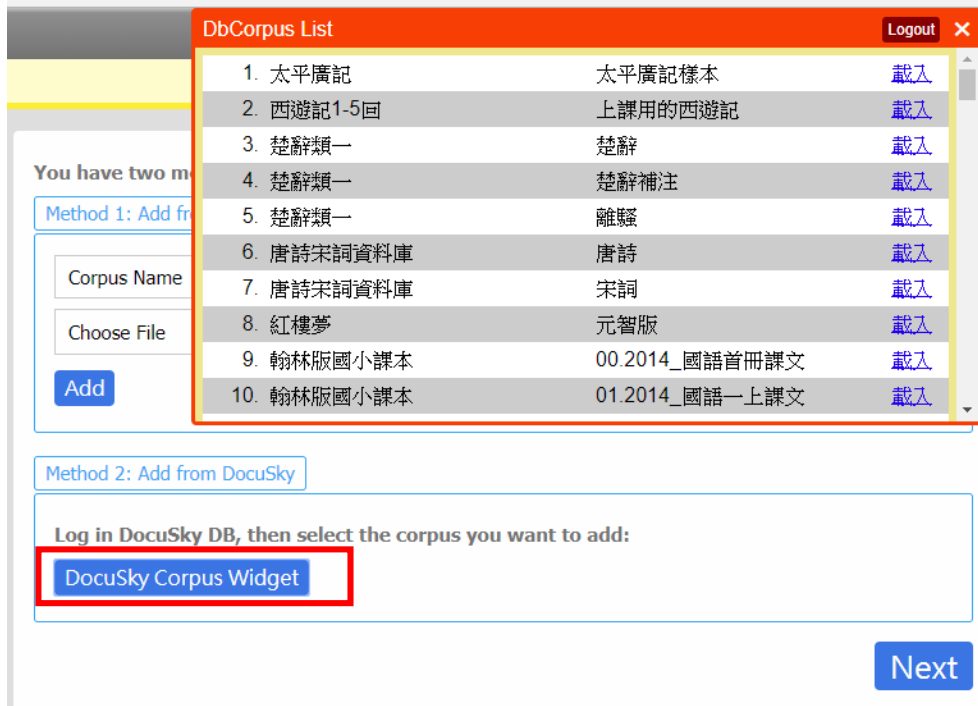


圖 7 從 DocuSky 載入文本

上述兩種方法中擇一完成文獻集上載後，按 Next 按鈕進入下一步。

3. TERMLIST 上傳詞彙表

上傳詞彙表的方式在系統的規劃中也有兩種，分別為上傳詞彙表（txt 檔，UTF-8 格式），以及從 DocuSky 載入詞彙表，惟目前後者的功能因故暫停使用。但若使用者從 DocuSky 中載入的文本已經經過標記，則可以標記過的詞彙直接進行分析。故此時便不用上傳任何詞彙表，直接按 Next 按鈕進入下一步。上傳的詞彙表一樣會顯示在右側的欄位中。



圖 8 上傳詞彙列表

由使用者直接上傳詞彙表 txt 檔的方式同前，這裡不再贅述。詞彙表 txt 檔的格式，必須每個詞彙以換行分開。

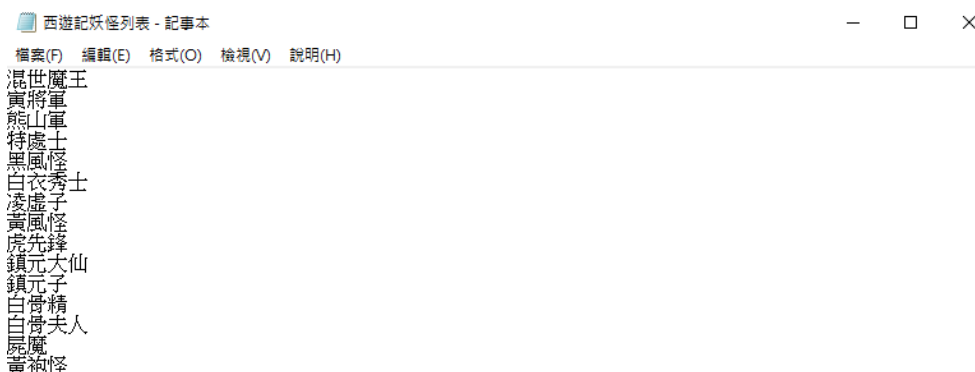


圖 9 編寫詞彙列表

上傳詞彙表到 DocuSky 存檔功能目前暫停使用，故不在此處贅述。

4. ANALYSIS 開始分析

開始分析的頁面分為兩種分析形式，Corpus and term lists 分頁表示，使用者可以利用剛剛上傳的詞彙表進行分析；Corpus with XML tags 則表示利用標記過

的詞彙進行分析。若是使用前者，確認上傳檔案無誤後，即可按 Run Analysis 進行分析。

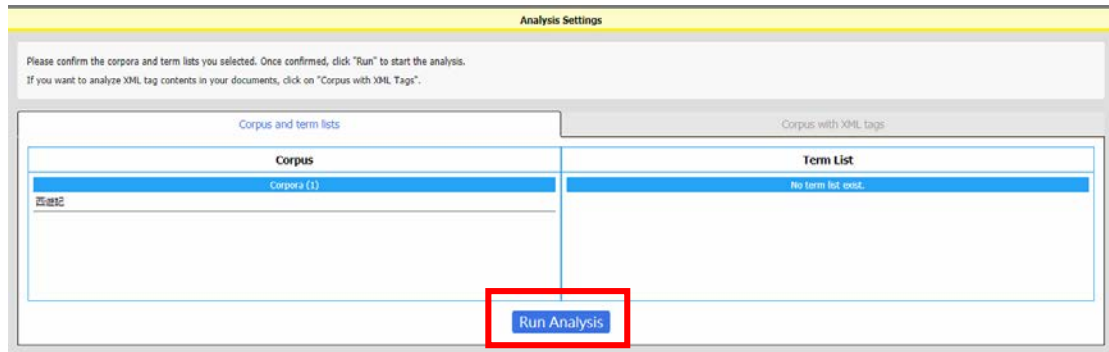


圖 10 以詞彙列表進行分析 (Corpus and term lists)

使用已經標記過的文本進行統計 (Corpus with XML tags) 時，下方會出現不同的標記類別供使用者選擇，依使用者的需求進行勾選即可。

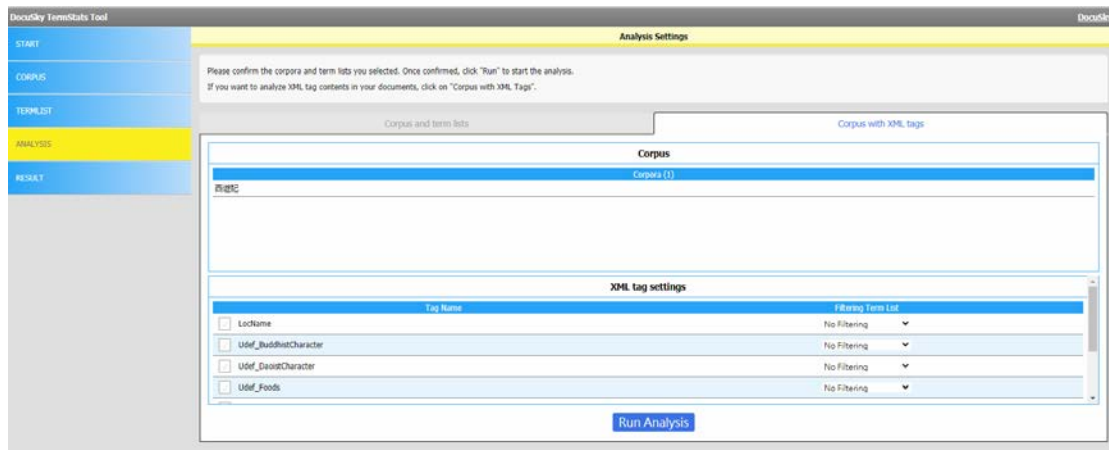


圖 11 以標記文本進行分析 (Corpus with XML tags)

以上圖為例，在載入的「西遊記_DocuSky 範本」這個文件集中，已有 LocName (地名)、Udef_BuddhistCharacter(佛教人物)、Udef_DaoistCharacter(道教人物)、Udef_MainCharacter (主角)等許多的標記。使用者可以按照目前需要進行統計的詞彙類別勾選一到多個類別進行統計。現以《西遊記》當中的佛、道教人物、妖怪 (Udef_Monster) 的詞頻進行分析。便可得到如下圖 11 的畫面：

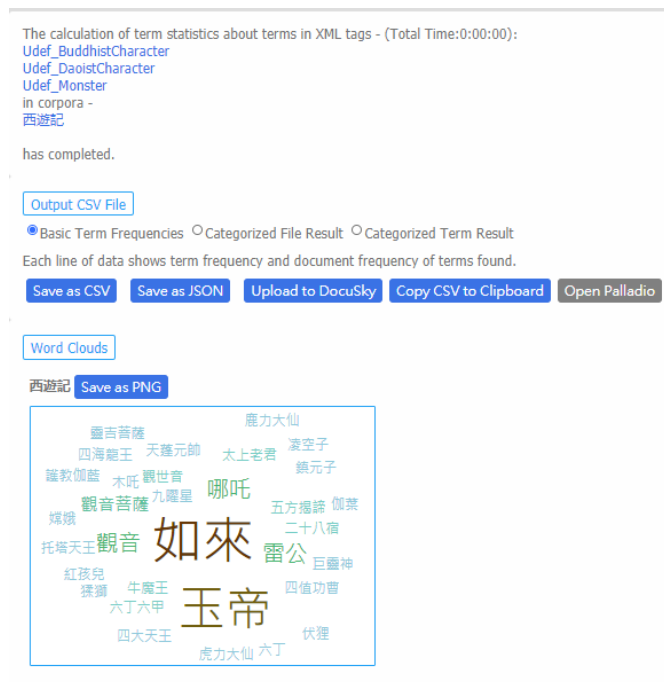


圖 12 統計結果

5. 統計結果與匯出

統計的結果，除可以文字雲（Word Clouds）呈現並點選「Save as PNG」將此圖檔下載。使用者可按三種不同方式進行詞彙統計，並將統計數據數據以 CSV（Save as CSV）、JSON（Save as JSON）檔案格式下載或上載到 DocuSky 上（Upload to DocuSky），或是直接將數據複製後匯出至史丹佛大學（Stanford University）所開發的視覺化平台 Palladio 進行後續的處理與呈現。

(1) Basic Term Frequencies 詞彙詞頻統計

最基本的統計數據是「Basic Term Frequencies」，輸出的資料將包括標記詞彙類別（Category）、詞彙（Term）、詞彙出現次數（TF）以及該詞彙出現在多少個段落（DF）。輸出的 CSV 檔案可以 Microsoft Excel 開啟，如下圖所示：

Category	Term	TF	DF
Udef_BuddhistCharacter	如來	207	38
Udef_BuddhistCharacter	觀音	65	25
Udef_BuddhistCharacter	觀音菩薩	50	31
Udef_BuddhistCharacter	觀世音	28	23
Udef_BuddhistCharacter	伽葉	10	2
Udef_BuddhistCharacter	靈吉菩薩	9	2
Udef_BuddhistCharacter	烏巢禪師	6	4
Udef_BuddhistCharacter	如來佛祖	3	3
Udef_BuddhistCharacter	如來佛	3	2
Udef_BuddhistCharacter	文殊菩薩	3	3
Udef_BuddhistCharacter	菩提祖師	2	1
Udef_BuddhistCharacter	釋迦牟尼	2	2
Udef_BuddhistCharacter	毗藍婆菩薩	1	1
Udef_BuddhistCharacter	大勢至菩薩	1	1
Udef_BuddhistCharacter	普賢菩薩	1	1
Udef_DaoistCharacter	玉帝	189	33
Udef_DaoistCharacter	哪吒	73	15

圖 13 Basic Term Frequencies 統計結果

使用者可就此表進行詞彙的統計與後續的利用。

(2) Categorized File Result 依段落統計詞頻

若選擇了「Categorized File Result」，則是以每個段落（filename）為基礎，統計各類詞彙在該段落中出現的次數。使用者也可以在匯出前選定要一併匯出的 metadata 欄位，以作為匯出之後資料的參照或後續利用。

如以載入的「西遊記_DocuSky 範本全標記」為例，在匯出時我們先選擇「Title」欄位一併匯出。

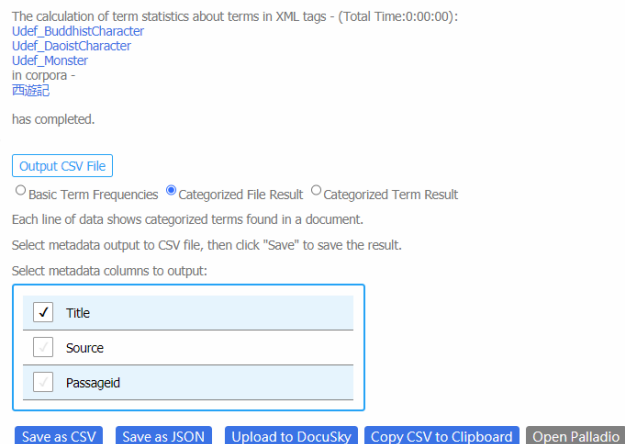


圖 14 以「Categorized File Result」方式進行匯出

匯出的 CSV 如下圖 15：

Category	Corpus	Filename	Title	TermsCount	TotalOccur	TermList	Detail
Udef_BuddhistCharacter	西遊記	西遊記p0001	靈根育孕源流出	1	2	菩提祖師	菩提祖師(2)
Udef_DaoistCharacter	西遊記	西遊記p0001	靈根育孕源流出	4	4	千里眼;順風耳;玉帝;搖光	千里眼(1);順風耳(1);玉帝(1);搖光(1)
Udef_Monster	西遊記	西遊記p0002	借徵菩提真妙理	1	1	混世魔王	混世魔王(1)
Udef_DaoistCharacter	西遊記	西遊記p0003	四海千山皆拱伏	18	27	四海龍王敖閻;秦廣王;玉帝	四海龍王敖閻(3);秦廣王(3);玉帝(3);東海龍王
Udef_Monster	西遊記	西遊記p0003	四海千山皆拱伏	3	3	混世魔王;猿猴;牛魔王	混世魔王(1);猿猴(1);牛魔王(1)
Udef_DaoistCharacter	西遊記	西遊記p0004	官封弼馬心何足	4	59	玉帝;哪吒;巨靈神;托塔天王	玉帝(28);哪吒(16);巨靈神(12);托塔天王(3)
Udef_Monster	西遊記	西遊記p0004	官封弼馬心何足	1	1	牛魔王	牛魔王(1)
Udef_BuddhistCharacter	西遊記	西遊記p0005	亂蟠桃大聖偷丹	1	1	觀音	觀音(1)

圖 15 Categorized File Result 統計結果

在上圖 15 中，「Category」欄為詞彙類別，「Corpus」欄為文獻集名稱，「Filename」則為段落檔名。「TermCount」是本段落（Filename）中，本類詞彙出現幾個？「TotalOccur」則是本類詞彙總共出現幾次？如在 Udef_BuddhistCharacter 這類中，出現了一個角色，即菩提祖師（TermsCount=1），而菩提祖師在第一章（西遊記 p0001）中被提到兩次（TotalOccur=2）。而出現的詞彙（TermList）為「菩提祖師」，在「Detail」欄中則可看出菩提祖師出現兩次（菩提祖師(2)）。

又如，道教人物（Udef_DaoistCharacter）在第一章出現了千里眼、順風耳、玉帝與搖光等四位道教神祇，且各被提到一次，因此 TermsCount 與 TotalOccur 分別都為 4。

(3) Categorized Term Result 依詞彙類別統計詞頻

「Categorized Term Result」是依照詞彙的類別進行詞頻的統計。使用者也可以在匯出前選定要一併匯出的 metadata 欄位，以作為匯出之後資料的參照或後續利用。

如以載入的「西遊記_DocuSky 範本全標記」為例，在匯出時我們先選擇「Title」欄位一併匯出。

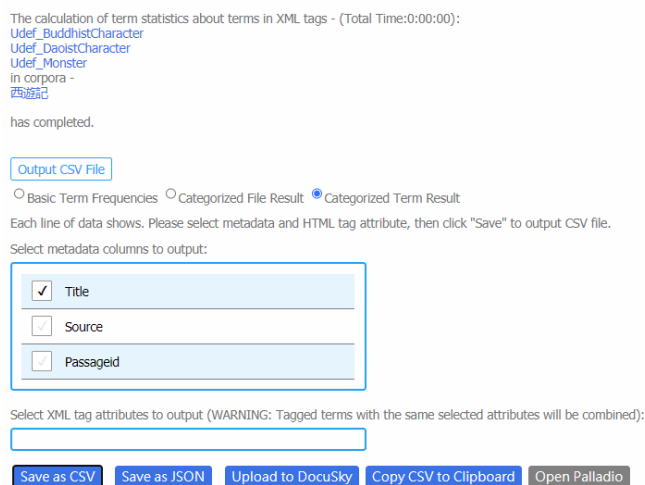


圖 16 以「Categorized Term Result」方式進行匯出

匯出的 CSV 如下圖：

Category	Corpus	Filename	Title	TagVal	Frequency
Udef_BuddhistCharacter	西遊記	西遊記p0001	靈根育孕源流出	菩提祖師	2
Udef_DaoistCharacter	西遊記	西遊記p0001	靈根育孕源流出	千里眼	1
Udef_DaoistCharacter	西遊記	西遊記p0001	靈根育孕源流出	順風耳	1
Udef_DaoistCharacter	西遊記	西遊記p0001	靈根育孕源流出	玉帝	1
Udef_DaoistCharacter	西遊記	西遊記p0001	靈根育孕源流出	搖光	1
Udef_Monster	西遊記	西遊記p0002	悟徹菩提真妙理	混世魔王	1
Udef_DaoistCharacter	西遊記	西遊記p0003	四海千山皆拱伏	東海龍王敖廣	2
Udef_DaoistCharacter	西遊記	西遊記p0003	四海千山皆拱伏	南海龍王敖欽	2
Udef_DaoistCharacter	西遊記	西遊記p0003	四海千山皆拱伏	北海龍王敖順	2
Udef_DaoistCharacter	西遊記	西遊記p0003	四海千山皆拱伏	西海龍王敖閻	3

圖 17 Categorized Term Result 統計結果

在上圖 17 中，「Category」欄為詞彙類別，「Corpus」欄為文獻集名稱，「Filename」則為段落檔名。「Title」是隨同匯出的 metadata 欄位。TagVal 為匯出的詞彙，「Frequency」則是該詞彙在本段落中出現的頻率。如在第三章中（西遊記 p0003）屬於道教人物的東海龍王、南海龍王、北海龍王各出現了兩次，而西海龍王敖閻則出現了三次。

四、 匯出資料與視覺化呈現

本工具所搭配的視覺化呈現工具為史丹佛大學所建置的 Palladio 視覺化平台，該平台可提供多樣化的視覺化呈現方式。因本工具與該平台的連結為 CSV 統計資料的匯出與關連分析圖的呈現，故僅就此部分進行說明，關於 Palladio 的其他功能請見：<http://hdlab.stanford.edu/palladio/help/>。

我們選用 Categorized Term Result 的統計結果後，直接選取「Copy CSV to Clipboard」即可將本統計結果以 CSV 格式暫存在剪貼板（Clipboard）中，此時再點選「Open Palladio」即可以另開分頁的方式在瀏覽器中開啟 Palladio。

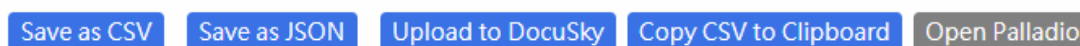


圖 18 複製並開啟 Palladio

進入 Palladio 之後，即可將已經暫存於剪貼板中的 CSV 資訊直接在 Palladio 的 Load .csv or spreadsheet 中，如下圖 19 所示：

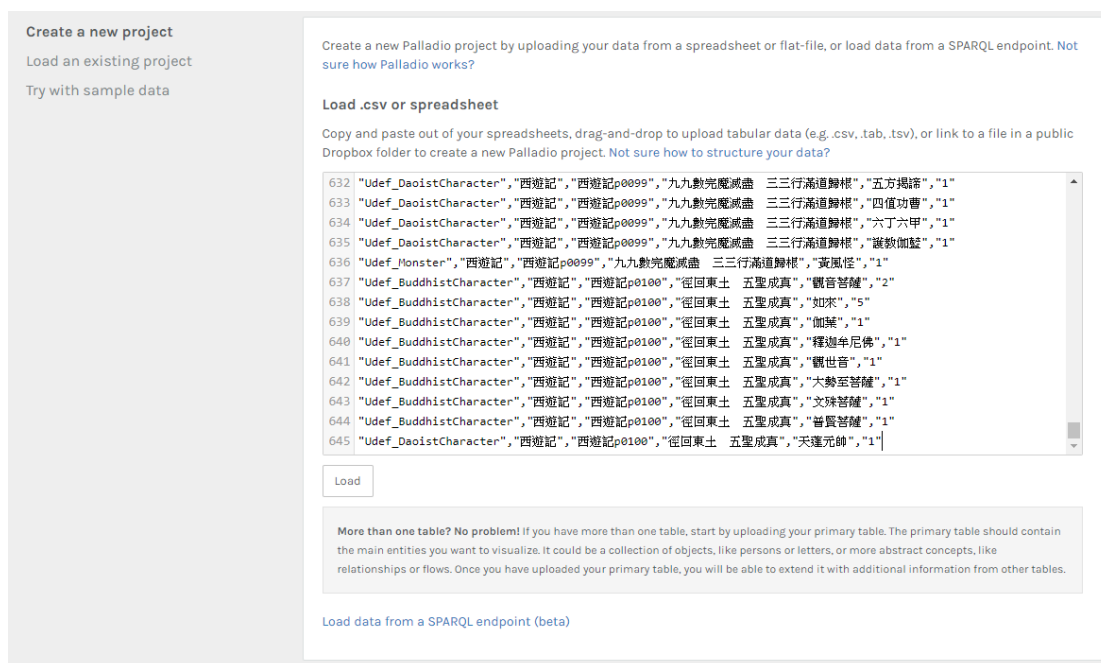


圖 19 在 Palladio 貼入複製的 CSV 資訊

點選「Load」按鈕即可將貼上的 CSV 資訊載入到 Palladio 當中。
進入到圖 20 的畫面後，使用者可點選「Graph」進入繪圖功能。

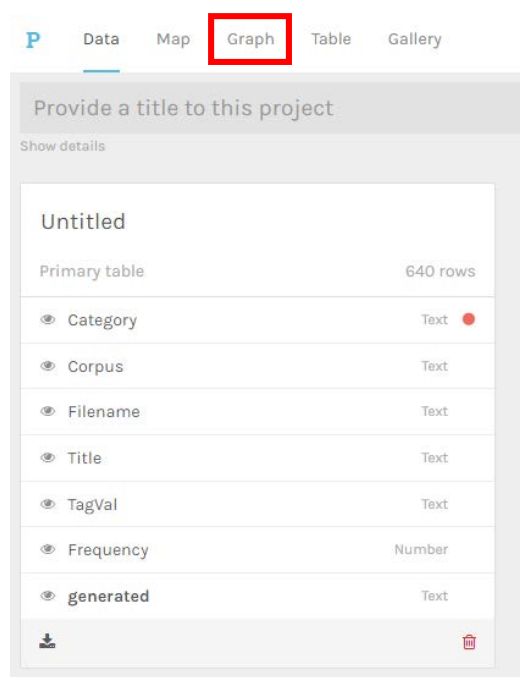


圖 20 點選 Graph 進入繪圖功能區

接下來使用者要決定哪兩項資料要當作是繪製關連圖的 Source（來源）與 Target（目標）。如在本案例中，Source 可以是各段落的 Title，也可以是各段落

的檔名 (Filename)，而 Target 則為被標記的詞彙 (TagVal)。當 Source 與 Target 選定之後，就會在繪圖區域生成詞彙關連圖。

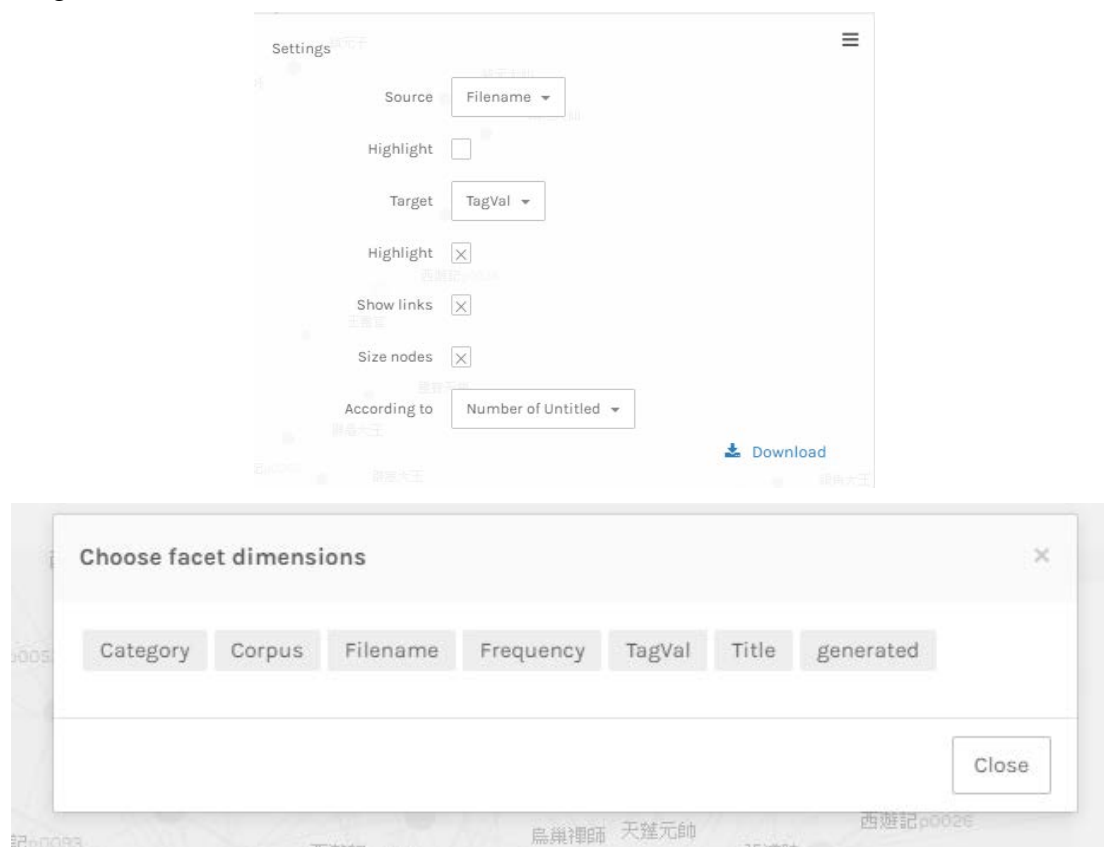


圖 21 選擇 Source 與 Target

使用者可按照個人的研究需求，在 Target 或 Source 中選擇一個點選 Highlight 使之醒目提示，也可以點選 Size nodes 使點的大小依照其頻率而有所不同。

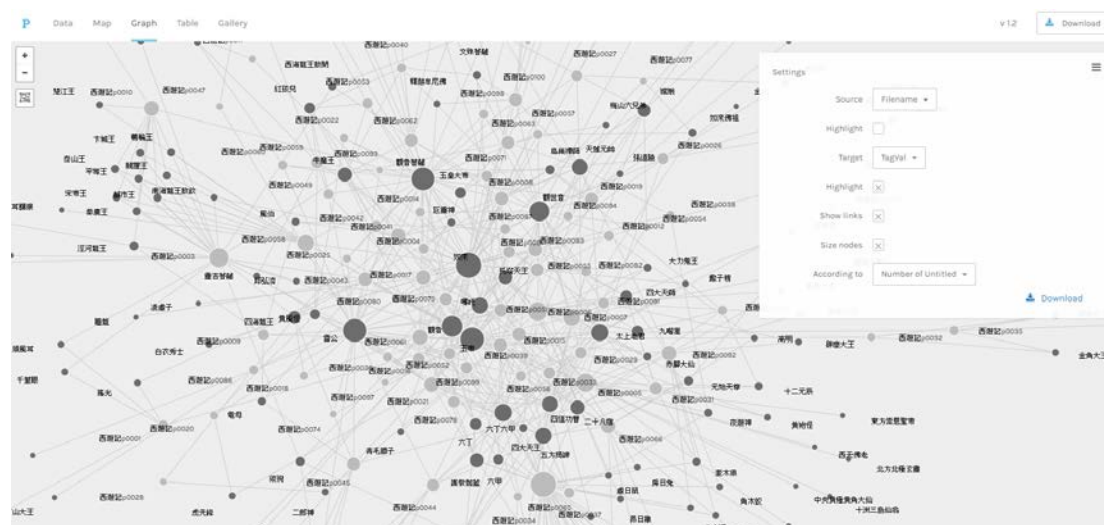


圖 22 點選參數後生成關連圖

不過，這樣的關連圖過於繁雜，我們可以利用篩選的機制把暫不需要的點隱藏起來。我們可以點選左下角的「Facet」開啟篩選機制。

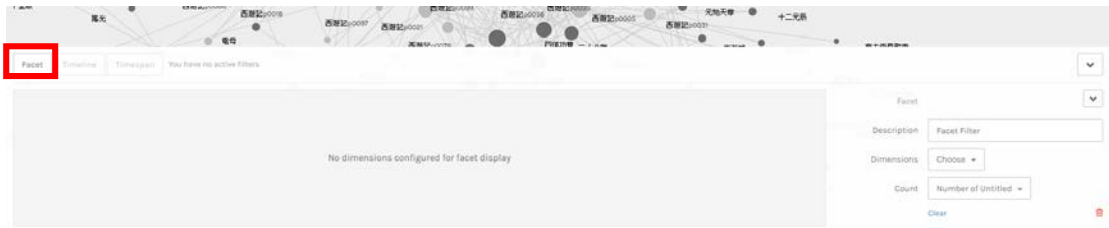


圖 23 點選 Facet 進行資料篩選

在「Dimensions」中我們可以選擇要進行篩選的 CSV 欄位，



圖 24 選擇要進行篩選的欄位類別

例如，我們只需要觀察《西遊記》中妖怪在各章節中出現的關連分析，我們就在這裡選擇「Category」(詞彙類別)，然後在篩選的選單中，點選 Udef_Monster，如此就可以先把道教人物與佛教人物先行隱藏，畫面也會變得比較清爽。

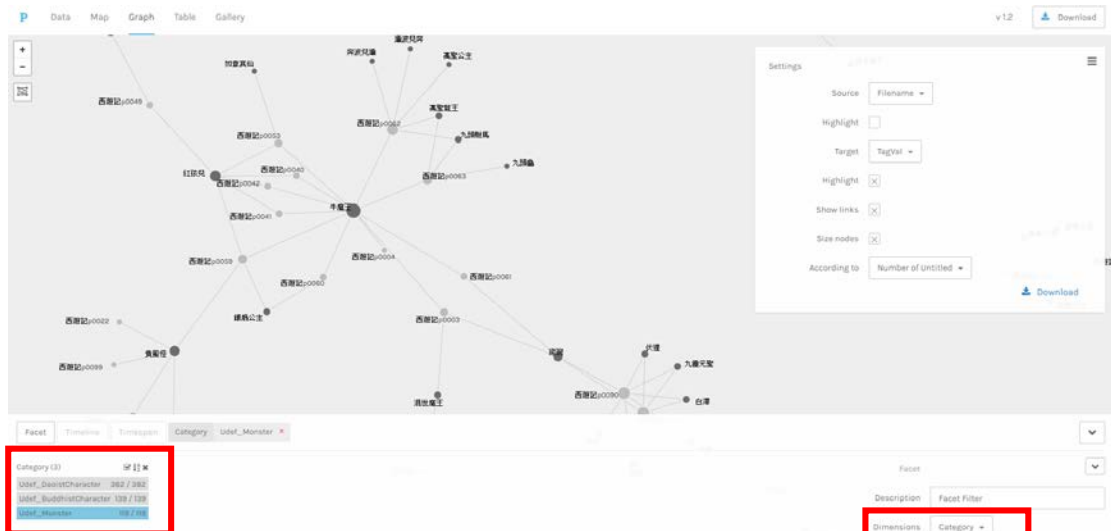


圖 25 篩選後的關連圖

之後，我們再靠著拖曳圖中的點，就可以呈現比較清楚的妖怪詞彙關連分析圖了。

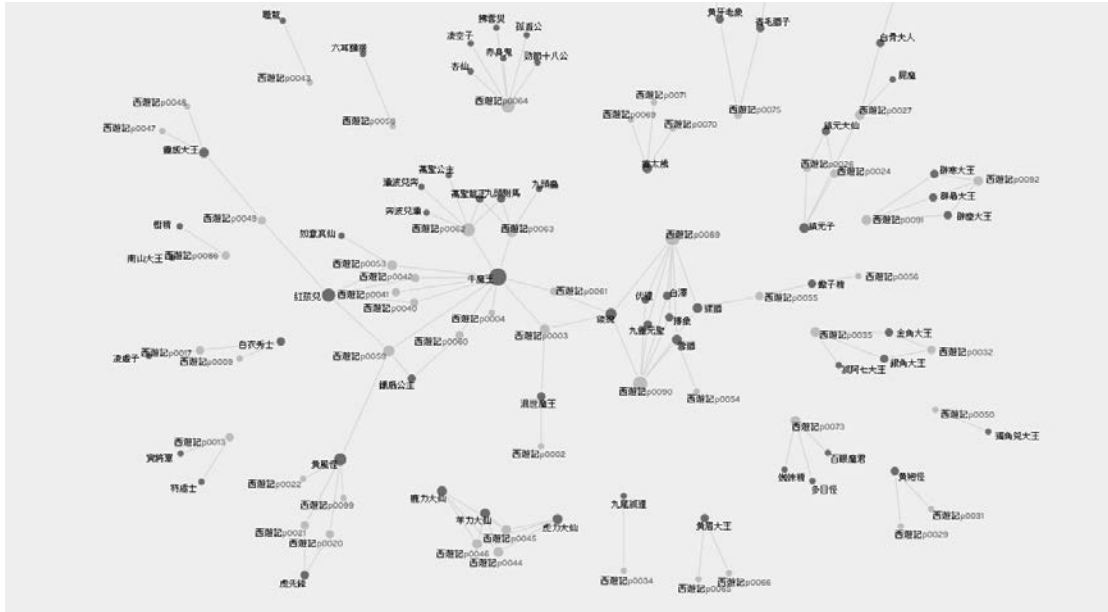


圖 26 《西遊記》妖怪詞彙關連分析圖

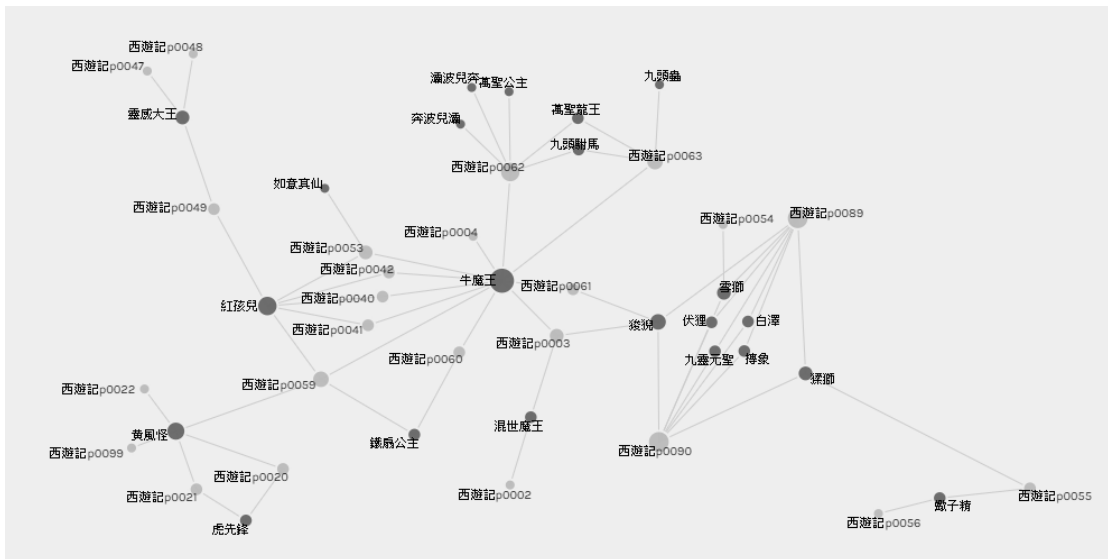


圖 27 以牛魔王為核心的妖怪詞彙關連分析圖

那麼，要如何看待圖 27 這張詞彙關連分析圖呢？我們發現，這是一張以牛魔王為核心的關連圖，由章節的節點所串起來的關係顯示，牛魔王與紅孩兒在西遊記 p0053、西遊記 p0042、西遊記 p0040、西遊記 p0041、西遊記 p0059 等章節中共同出現過，而又如鐵扇公主在西遊記 p0059 這一章中與紅孩兒、牛魔王共

同出現過。這種「共現關係」即能透過這樣的關連分析圖迅速呈現。並且可以利用更多不同類別的詞彙、CSV 欄位的關連性，提供使用者做更多的視覺化觀察。