國立政治大學數位內容碩士學位學程

Master's Program in Digital Content & Technologies

National Chengchi University

碩士論文

Master's Thesis

基於標記式主題模型之資料視覺化研究與實現

A Study of Data Visualization Based on Labeled Topic Model and Its Implementation

研 究 生：曾子芸

指導教授：陶亞倫、蔡銘峰

中華民國 一百零六 年 三 月

March 2017

106
碩士論文

基於標記式主題模型之資料視覺化研究與實現

國立政治大學
National Chengchi University

政治大學數位內容碩士學位學程

曾子芸

基於標記式主題模型之資料視覺化研究與實現

A Study of Data Visualization Based on Labeled Topic Model and Its Implementation

研 究 生：曾子芸　　Student：Tzu-Yun Tseng

指導教授：陶亞倫、蔡銘峰　　Advisors：Ya-lun Tao、Ming-Feng Tsai

國立政治大學

數位內容碩士學位學程

碩士論文

A Thesis

submitted to College of Communication

National Chengchi University

in partial fulfillment of the Requirements

for the degree of

Master

in

Digital Content & Technologies

中華民國 一百零六 年 三 月

March 2017

致謝

回首過去，總結起來研究所就是一段不停自我懷疑及自我審視的一段過程。回想起剛開始什麼都不會的我，經歷過許多熬夜寫程式到天亮的日子，每天最常見的風景就是書桌前面貼滿便利貼的白牆壁，終於，也慢慢地走到即將畢業的這一刻了。在此，想利用謝詞這小小的空間，來感謝許多人這一路上對我的提攜與幫忙。

首先要感謝的就是讓我進實驗室的蔡銘峰老師，以及另一位主指導教授陶亞倫老師。謝謝蔡老師當初知道我只會基本的 html、css 還願意讓我先去實驗室旁聽，如今回想起來真的是覺得老師人真的很好。也謝謝陶老師每次見面討論時，給我的許多寶貴建議，每次和陶老師討論完畢時，都覺得又有好多想法與收穫，而陶老師安定的話語，也讓總是常對自己作品懷疑的我能更有信心地繼續前進。

再來想要感謝在這一路上陪伴我的朋友們，寫論文的過程中總是孤獨的，但有了你們便讓我覺得不寂寞，謝謝 DCT 6 的全體同學，每一位都讓我如此的喜歡～再來是想謝謝和我一起參與數位人文計畫的夥伴們，這兩年來真的辛苦了，也想謝謝實驗室的其他同學，對我大大小小的幫助～很開心可以遇到一群如此善良又厲害的同學們，謝謝你們！

最後要感謝的就是我的家人們了！謝謝你們讓我可以無後顧之憂地專心當一個研究生，這段期間比較少回家真的覺得很不好意思，也謝謝你們從來不會給我過多的壓力催促我，讓我能有動力繼續前進！

其實想感謝的人還有好多好多，回想起來這一路上真的受到了太多人的幫忙，謝謝這些曾經幫助過我的人們，希望以後的我也能努力培養自身能力來幫助別人，以回饋這些我所收到的，難以回報的恩情。

曾子芸

國立政治大學數位內容碩士學位學程

March 2017

中文摘要

隨著文字資訊的爆炸式增長，越來越多的訊息開始以電子文本的形式儲存及傳遞。但隨著文本內容資訊量不斷地增加，使用者也越來越難以快速地掌握文本全貌。因此本研究試圖透過主題模型（Topic Models）、標記式主題模型（Labeled Topic Models）演算法－在自然語言處理領域裡文本探勘的方法，識別出大規模文本中潛藏的主題訊息之後，再利用圖像視覺化在資訊表達上的優勢和效率，透過各種視覺化圖案的呈現從不同的角度來探索文本，形成一種嶄新的大規模文本閱讀與分析方式。

本研究設計了兩階段實驗：第一階段任務導向性實驗、第二階段指定任務實驗，以及評估問卷來驗證本介面的易用性（Ease-of-use）和有用性（Usefulness）。並透過實驗問卷的分數結果驗證了，本研究所設計之介面在實務上的確能輔助專家學者進行文本相關研究，也能讓對文本熟悉程度不一的使用者在利用此介面探索文本的過程中，更快速地掌握大規模文本的事件全貌。

國立政治大學 National Chengchi University

Abstract

With the explosion of text information, there are more and more data being recorded and transmitted in the form of texts. However, as the amount of textual information becomes larger, how to effectively and efficiently realize the information also becomes more difficult. This study attempts to use the Topics Models, text-mining techniques to identify the important topics in the large textual information. In addition, this study also aims to use the techniques of data visualization to present the most informative and valuable details within the large texts.

There are two parts in this work: the first part is the introduction of text-mining algorithms and the second part is the design of the data visualization. Moreover, in the experiments, we also conduct several surveys to verify the proficiency and usefulness and the visualization design. The results of the experiments and surveys, supports that our design provides an effective and efficient interface for users to understand a large set of texts, even for the experts familiar with the corpus.

國立政治大學
National Chengchi University

目錄

致謝 **1**

中文摘要 **2**

Abstract **3**

第一章　緒論 1
1.1　前言 1
1.2　研究動機 1
1.3　研究目的 2
1.4　論文架構 3
第二章　相關文獻探討 4
2.1　主題模型 4
2.1.1　隱含狄利克雷分布 LDA 5
2.1.2　標記式隱含狄利克雷分布 Labeled LDA 5
2.2　主題模型視覺化 6
2.2.1　靜態主題模型視覺化展示 7
2.2.2　靜態主題模型視覺化總結 9
第三章　介面概念與設計 10
3.1　系統概念 10
3.1.1　文本主題與關鍵字 12
3.2　介面設計流程 14
3.3　系統介面 15
3.3.1　首頁（Index） 16
3.3.2　主題／主題代表字總覽（Grid） 18
3.3.3　主題遠近分佈（Scaled） 18
3.3.4　主題列表（List） 19
3.3.5　主題年代分布（Year） 20
3.3.6　文字索引（Index） 20
3.3.7　主題專題頁（Topic） 21
3.3.8　詞彙搜尋頁面（Search） 21
3.4　介面設計總結 22
第四章　介面實作技術 23
4.1　整體介面架構 23
4.1.1　整體顏色設計 23
4.2　視覺化圖表技術說明 24
4.2.1　各式主題模型視覺化技術呈現 25
4.2.2　其餘靜態圖表技術 29
4.3　文本主題遠近分析技術 30

國立政治大學 National Chengchi University

第五章　實驗設計與結果評估 33
5.1　實驗目的 33
5.2　實驗對象 33
5.3　實驗流程 34
5.3.1　實驗前測問卷 35
5.3.2　前置介面引導任務 36
5.3.3　第一階段易用性調查問卷 37
5.3.4　第二階段指定任務 38
5.3.5　第二階段有用性調查問卷 39
5.4　實驗結果分析與討論 41
5.4.1　系統可用性尺度量表（SUS）比較結果分析 41
5.4.2　第二階段有用性問卷內容結果分析 44
5.4.3　實驗前測問卷前後結果分析 47
5.4.4　開放式問題訪談結果分析 52
第六章　結論與未來展望 54
6.1　研究結論 54
6.2　未來發展與改進 54
參考文獻 55
附錄 58
附錄一 第一階段介面引導任務問題列表 58
附錄二 受測者開放性問題回饋 60

國立政治大學
National Chengchi University

圖目錄

圖 2.1 LDA 演算法的概率圖模型 5
圖 2.2 Labeled LDA 演算法的概率圖模型 6
圖 2.3 The Termite System 7
圖 2.4 Navigating Wikipedia with Topic Model 8
圖 2.5 LDAvis 電影分類建模圖 9

圖 3.1 本研究系統架構圖 11
圖 3.2 本研究介面設計流程圖 14
圖 3.3 本研究主要介面 16
圖 3.4 首頁說明 17
圖 3.5 首頁視覺化部分 17
圖 3.6 主題／主題代表字總覽 18
圖 3.7 主題遠近分佈視覺化 19
圖 3.8 主題列表頁 19
圖 3.9 主題年代分布視覺化 20
圖 3.10 文字索引頁 20
圖 3.11 主題專題頁 21
圖 3.12 數位工具平台搜尋頁面 22

圖 4.1 Material Design 線上工具 24
圖 4.2 以開放式數位平台頁面爲例實現 Material Design 之頁面呈現 . . . 24
圖 4.3 圓餅圖原始數據 25
圖 4.4 經 D3.js 轉換過後得到的圓餅圖新數據 26
圖 4.5 條列式呈現主題和主題底下關鍵字 26
圖 4.6 泡泡圖呈現主題和主題底下關鍵字 27
圖 4.7 主題占文本比例橫柱圖 27
圖 4.8 主題底下關鍵字比例橫柱圖 27
圖 4.9 面積圖原始資料格式 28
圖 4.10 經 D3.js 轉換過後得到的面積圖新數據 29
圖 4.11 面積圖文本主題視覺化 29
圖 4.12 Piety.js 小型迷你圖表 30
圖 4.13 Jensen–Shannon Result 31
圖 4.14 Principal Components Analysis Result 32
圖 4.15 文本主題距離遠近泡泡圖呈現 32

圖 5.1 實驗流程圖 34
圖 5.2 Technology Acceptance Model 科技接受模型 35
圖 5.3 互動式前置介面引導任務網頁 37
圖 5.4 System Usability Scale (SUS) 系統可用性尺度量表分數評估表 . . . 42

圖 5.5　受測者前測問卷前後結果分析－問題一 47
圖 5.6　受測者前測問卷前後結果分析－問題二 48
圖 5.7　受測者前測問卷前後結果分析－問題三 49
圖 5.8　受測者前測問卷前後結果分析－問題四 50
圖 5.9　受測者前測問卷前後結果分析－問題五 51

國立政治大學
National Chengchi University

表目錄

表 2.1　圖 Figure 2.1、Figure 2.2 所列符號一覽表 6

表 3.1　文本資料集資訊 13

表 5.1　實驗前測問卷問題列表 36
表 5.2　系統可用性量表（SUS）問題列表 38
表 5.3　指定任務問題列表 39
表 5.4　第二階段有用性調查問卷問題列表 40
表 5.5　系統可用性尺度量表（SUS）受測者分數列表 41
表 5.6　系統可用性尺度量表（SUS）問題分數列表 43
表 5.7　第二階段有用性問卷各項分數列表 45
表 5.8　實驗前測問卷結果分析－問題一 47
表 5.9　實驗前測問卷結果分析－問題二 48
表 5.10　實驗前測問卷結果分析－問題三 49
表 5.11　實驗前測問卷結果分析－問題四 50
表 5.12　實驗前測問卷結果分析－問題五 51
表 5.13　開放式問題列表 52

國立政治大學 National Chengchi University

第一章

緒論

1.1 前言

隨著文字資訊的爆炸式增長，日益加快的工作節奏，越來越多的訊息開始以電子文本的形式加以儲存及利用。各式各樣的文字資訊諸如郵件、新聞，甚至網站、論文期刊、電子書平台等等，其實都是以不同格式的電子文本來儲存主要訊息。日新月異的科技帶來了儲存訊息、獲取訊息的便利性，但持續不斷的累積也同時導致文本的數據量日漸龐大，遠遠超出人們的閱讀負荷，深陷在「訊息超載」的泥淖。面對這種情況，人們開始需要更高效的文字閱讀及分析方法讓閱聽者從文本集中挖掘所需要的部分，於是從文本集中抽取文本主題以概括文本的內容，也就是所謂「主題模型」的文本探勘方法也隨之產生。

然而在主題模型的基礎上，如何適當地呈現主題及主題內隱含之關係也是一個重要的問題。傳統的文本主題模型雖然在一定程度上已經實現了從文本中挖掘出主要的主題訊息，但這些主題訊息的結果是以主題向量及詞向量的形式存在，無法直接協助閱聽者對於文本事件合理分析、理解的應用需求，於是文本視覺化技術基於上述需求開始應運而生。它將文本中複雜的或者難以通過文字表達的內容和規律以視覺符號的形式表達出來，同時向人們提供與視覺資訊進行快速交互的功能，使人們能夠利用視覺感知處理能力快速獲取資料中所蘊含的關鍵訊息。因此，如何善加利用文本視覺化的特性與技術，精準且富含美感的詮釋文本主題模型，讓使用者更快速的掌握整體文本也將是未來的趨勢之一。

1.2 研究動機

經過和史學團隊一起研究的過程後發現，有別於以往史學家利用自身領域知識（Domain Knowledge）去研究、分析文本脈絡的研究方法，利用資訊科學領域中的資訊檢索（Information Retrieval）技術與自然語言處理（Natural Language Processing）的技術，以機器學習以及統計式學習的方法來觀測 文本中的脈絡，確實能有效且快速地挖掘出文本中機器認爲最重要的詞彙或是主題概念。可以說，研究團隊本身的專業認知，再加上機器學習結果的數據佐證，兩者之間有著相輔

相成的結果。

然而傳統主題模型所呈現出來的結果，往往以文字條列式爲主，而標記式主題模型等機器學習演算後的結果，常以機率分佈或是向量的形式，也就是一長串的數字符號來呈現。無論是以文字條列式、抑或是以演算法的結果直接呈現，其實都不容易直接觀看到結果背後所呈現的意義，時間一久也很容易忘記文字條列或是演算法的內容。更有甚者，若非本身爲專業領域人員，對於一般使用者而言，想要直接透過演算法的結果來來理解文本，更是難度相當高的一件事情。基於以上種種原因，本研究試圖建立一個以標記式主題模型演算法結果所得出的文本主題視覺化介面。希望透過此視覺化介面的操作之後，可以將演算法的結果以視覺化的圖形方式呈現，並且提供能與使用者互動的介面，讓使用者自行挖掘及探索文本內容，同時在互動過程中做立即性的呈現。

除此之外，在與史學團隊的合作過程也發現，當研究者發現了感興趣的詞彙，或是在某個主題底下出現他們預料之外的詞彙的時候，研究者經常會反覆進行文本搜尋的動作，藉由搜尋此詞彙在文本的位置，來還原當時歷史情境，及藉由詞彙上下文來尋找其脈絡，因此，文本搜尋系統與視覺化介面的結合可以說對研究者而言，是相當快速又方便的一種新研究模式，所以本研究也希望能藉由此視覺化平台，結合文本搜尋系統，成爲一個輔助專家研究學者的工具，開啓研究團隊另一種研究的新渠道。

1.3 研究目的

根據先前學者的研究 [16]，透過主題模型的資料視覺化，希望能夠達到本研究以下目的：

1. 各式專門領域文本的內容結構。
2. 該領域文本中的核心主題與詞彙。
3. 以傳統學者的研究方法尚未找出的相關研究議題。
4. 對於具有時間序列性文本的演進與發展。

基於以上的目標，本文試圖建立一個以標記式主題模型結果所得出的資料視覺化介面。希望透過此視覺化介面的操作之後，可以將文本的標記式主題模型以視覺化的圖形方式呈現，並且提供能與使用者互動的介面，讓使用者自行挖掘及探索文本內容，成爲一個輔助專家研究學者的工具，並能在畫面上做立即性的呈現。而主要具體之視覺化目標如以下所述：

1) 瞭解文本裡各主題分佈情形：

以標記式主題模型跑完大規模文本之後，我們可以得知每個主題在文本裡出現的頻率，本視覺化介面將這個數字以直方圖和河流圖來表現，其中直方圖所代表的是每單一個主題在各年代所含的比重多寡，河流圖則代表所有主題在各年代所分布呈現的趨勢比重。

2) 了解每個主題在文本裡所代表的意義

同樣以標記式主題模型跑完大規模文本之後，我們可以得到每個主題底下的一組詞向量，代表每個主題是由哪些詞彙所組成的，本視覺化介面則將這組數字以長條圖的長度及右側數字來表示，使用者查看長條圖寬度大小即可迅速了解每個主題在文本裡所代表的意義及代表性詞彙。

3) 瞭解詞彙在特定文本裡的獨特代表性

同樣的詞彙，在日常生活中以及在特定文本中所代表的或許是不一樣的意思，若使用者非專業該領域學者，可能無法理解詞彙所代表的意思，因此本視覺化介面希望能將視覺化內容和文本更加緊密結合在一起，透過文本再搜尋的動作直接引導詞彙至原文本所在內容，幫助使用者更快速理解相關文本片段與內容。

1.4 論文架構

本論文的章節內容說明如下：

第一章爲說明與緒論，闡述本研究之目的、動機與試圖解決的問題。第二章將說明 LDA 、 Labeled-LDA 各主題模型之簡介、文本視覺化簡介、相關範例文獻等等，與本研究相關的文獻與範例說明。第三章將闡述本視覺化介面的介面設計與概念。第四章將闡述本視覺化介面的詳細實作內容。第五章將實驗此視覺化介面對於專家研究學者的實際幫助，透過專門設計的量表與問卷，與歷史學者一起設計具有公評力之題目測驗，收集使用者之回答數據與使用心得。第六章，終章，爲本研究之結論與展望，將探討本研究對傳統主題模型的貢獻，在視覺化方面的心得展望，與對未來研究的考慮與建議。

國立政治大學 National Chengchi University

第二章

相關文獻探討

隨著巨量資料興起，我們試圖通過圖形化的數據和交互式設計來展示文本裡隱含的訊息，這是文本視覺化的發展趨勢之一。而針對如何將語料庫文本進行視覺化處理，本質上其實也包含了文本探勘中的主題模型問題。爲了完成文本資料視覺化的任務，首先我們需要對文字資料處理、主題模型、文本視覺化和相關技術有基本的了解，所以在本章節的前半部，我們將會從主題模型開始介紹，並說明在本研究中所使用的主題模型和標記式主題模型演算法，並於後半部介紹文本視覺化的相關文獻及使用工具。而對於最終的文本靜態視覺化分析結果，會在之後章節有更詳盡的介紹。

2.1 主題模型

無結構的文字資料無法直接用於資料視覺化，因此，採用適合的文字度量方法從文字中分析結構化的資訊非常重要。主題模型（Topic Model）[1]就是用來在一系列文檔中發現抽象主題的一種統計模型。直觀來講，如果一篇文章有一個中心思想，那麼一些特定詞語會更頻繁的出現。比方說，如果現在有篇闡述「狗」的文章，那「狗」和「骨頭」等詞可能的出現頻率會比較高。主題模型正是試圖用數學來體現文檔的這種特點，統計文檔內的詞語，根據統計的信息來斷定當前文檔含有哪些主題，以及每個主題所占的比例各爲多少 [17] 。

主題模型之概念興起於二十世紀之 90 年代，其代表模型包括 Deerwester 等 [8]於 1990 年提出的隱含語義分析（Latent Semantic Analysis, LSA）, Hofmann 等 [11]提出的概率潛在語義索引（Probabilistic Latent Semantic Indexing, PLSI）, Blei 等 [3]提出的隱含狄利克雷分布（Latent Dirichlet Allocation, LDA）以及 Teh 等 [22]提出的層次化狄利克雷過程（Hierarchical Dirichlet Process, HDP）等 [21], 其中又以 LDA 模型的影響最爲廣泛。在本研究中，主要是以採用 LDA 主題模型進行研究，以下將針對 LDA 演算法做更進一步的介紹。

[1]Topic models are a suite of algorithms that uncover the hidden thematic structure in document collections. These algorithms help us develop new ways to search, browse and summarize large archives of texts.–引用自David M. Blei，網址爲 https://www.cs.princeton.edu/ blei/topicmodeling.html

2.1.1 隱含狄利克雷分布 LDA

LDA [3] 是一種非監督機器學習技術，可以用來挖掘大規模文檔集或語料庫中潛藏的主題資訊。爲了減少需要學習的參數數量，LDA 假設每一篇文檔所蘊含主題是隨機變數，滿足一定的概率分佈。Figure. 2.1所示爲 LDA 的概率圖模型，其中，陰影圓圈表示已觀察到的隨機變數，非陰影圓圈表示隱變數，箭頭方向表示兩個隨機變數之間的條件概率依賴關係。

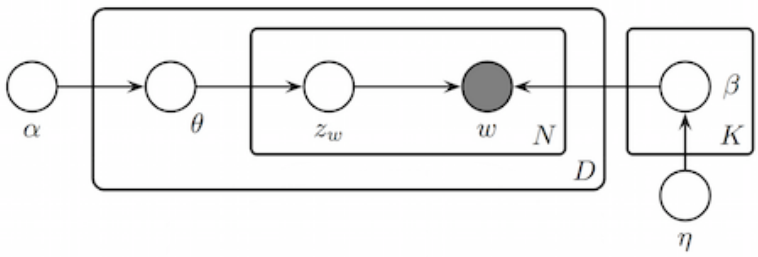


圖 2.1: LDA 演算法的概率圖模型

對於語料庫中的每篇文檔之主題模型，LDA 定義了如下生成過程:

1. 對於每篇文檔 d，從狄利克雷分佈 $Dir(\alpha)$中採樣得到該文檔的主題比例 θ_d
2. 從狄利克雷分佈 $Dir(\beta)$中產生所有 K 個主題 φ_k
3. 對於文檔 d 中每個單詞 w，從主題比例 θ_d 中採樣一個主題分配 z
4. 從主題 Φ_z 中生成對應的單詞 w
5. 重複上述過程，直至生成文檔中的每一個單詞

2.1.2 標記式隱含狄利克雷分布 Labeled LDA

一般而言，在文本裡面往往會有一些伴隨資訊，如一篇新聞文本可能會被標注上藝術、物理和臺北等帶有意義的不同標籤，它們可爲其主題建模提供十分重要的資訊，以明模型獲得更具解釋性的主題。爲此，一些可以利用這些標籤資訊的監督式主題模型方法被提出，主要是希望能儘量保證文檔生成主題與標籤資訊存在一定的匹配性。

基於上述原因，Ramage 等[16] 提出的標記式 LDA（Labeled LDA）在傳統 LDA 建模中加入了一個簡單約束，即所生成的主題只能來自標籤集合。依照 Ramage 等人所提出的標記式隱含迪利克雷分佈（ Labeled LDA ），我們將可以簡單地用預先觀察到的文章主題（即人工標記的主題分類）去限制主題模型。其模型圖如 Figure. 2.2 所示，其中所使用的符號與前述傳統 LDA 相同，而另外增加的符號是 Λ，Λ_d 表示該篇文章所標記的主題種類，爲一長度爲文本數，內容僅

包含〔0,1〕的整數陣列，加入此限制後，文章與主題之間的分布 θ 不再只與 α 有關而是同時受限於 Λ 值。如此一來，即可確保所有文章字詞所分配到的主題 $z_{d,i}$ 皆被該文章原有之主題標記所限制。

由於此種避免了將詞彙隨機分配到文本內所有主題的算法，將字詞的主題範圍強迫限定在所屬文檔標記的主題之內，因此良好的利用了原文本被人工標注的各種意義標籤，其所分類的效果和傳統 LDA 模型比起來也有顯著的良好效果，因此本研究選擇利用 Labeled LDA 爲此次文本視覺化的文本處理主要算法，並利用 Labeled LDA 所算出來的文本字詞向量結果進行後續的各式交互式視覺化分析。

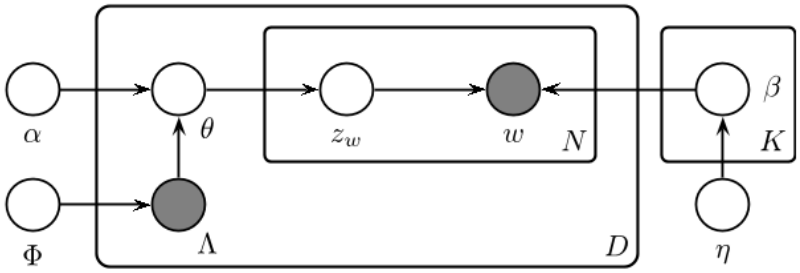


圖 2.2: Labeled LDA 演算法的概率圖模型

表 2.1: 圖 Figure 2.1、Figure 2.2 所列符號一覽表

$Dir(\alpha)$	狄利克雷分佈 (for document topic distributions)
$Dir(\beta)$	狄利克雷分佈 (for topic word distributions)
D	語料庫
N	文檔中的詞語數
K	主題數
θ_d	文檔 d 的主題分佈
$z_{d,i}$	文檔 d 中第 i 個詞的主題
$\phi_{zd,i}$	從取樣中主題所對應的詞語分佈
$w_{d,i}$	最終生成詞語
Λ	所標記的主題種類
Λ_d	文章 d 所標記的主題種類

2.2 主題模型視覺化

基於文檔內容的主題模型視覺化是研究語料庫中每個文檔的主題分佈、每個主題的詞項分佈、文檔間的關係、主題間的關係，並通過視覺化方法將這些資訊

展現出來，讓從事相關研究工作者能從宏觀上快速瞭解和掌握大規模文檔集的主題和主要思想。而針對文本視覺化方法呈現方式的不同，可分爲動態視覺化以及靜態視覺化，動態視覺化又可再細分爲時間序列的主題展示和動畫主題展示，靜態視覺化則著重在研究語料庫中的主題關係 [14] 。本文以下將主要針對文本的靜態視覺化展開研究。

2.2.1 靜態主題模型視覺化展示

史丹佛大學視覺化團隊的 Chuang J、Manning C D、Heer J 在 2012 年提出了 Termite[6] ，用圖形的方式來展示對大規模文檔集主題建模的結果。如 Figure. 2.3 所示， Termite 展示了利用 LDA 建模產生的主題內容，其中縱軸表示文檔集中詞彙表裡面的所有詞項，橫軸是經過主題模型建模後產生的主題編號，若主題包含相對應的關鍵詞，則在交會位置上畫一個圈圈，圈圈的大小表示關鍵詞在相應主題中的權重。除此之外，Termite 還引入了自己設計的鍵能算法（The Bond Energy Algorithm），把經常一起出現的詞彙排在一起，例如 social 和 networks ，增強了語義便於使用者理解。較爲可惜的是 Termite 缺乏了一種對 LDA 模型的全域視角，只顯示了算法裡權重最高的前三十五個字，無法看見文檔裡其他的詞彙，也沒有展示主題、文檔間的關係。

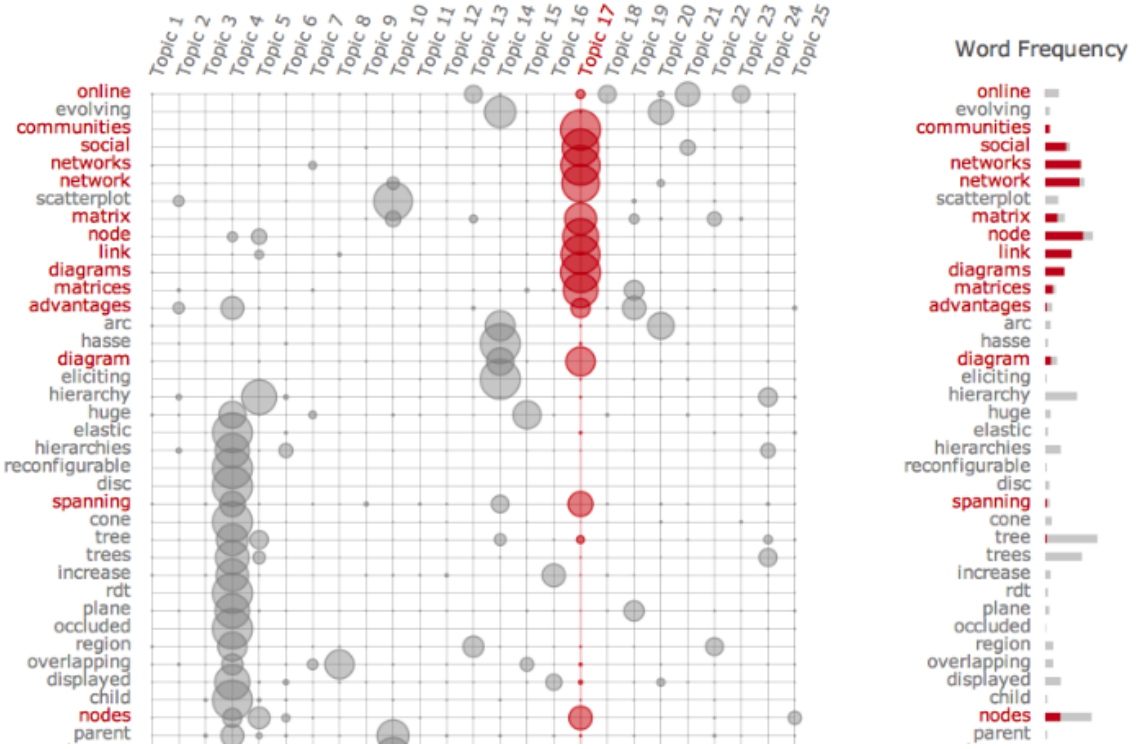


圖 2.3: The Termite System

僅展示主題的部分關鍵詞並不能完全展現出主題內容，所以，如 Figure. 2.4 所示，2012 年，Chaney A J B, Blei D M 提出了 VTM（Visualizing topic models）[5]。VTM 不僅展示了主題－主題間的關係，主題－文檔的關係，還展示了主題－詞項間、詞項－詞項間的關係，更詳細展示了文本進行 LDA 建模後的輸出成果。

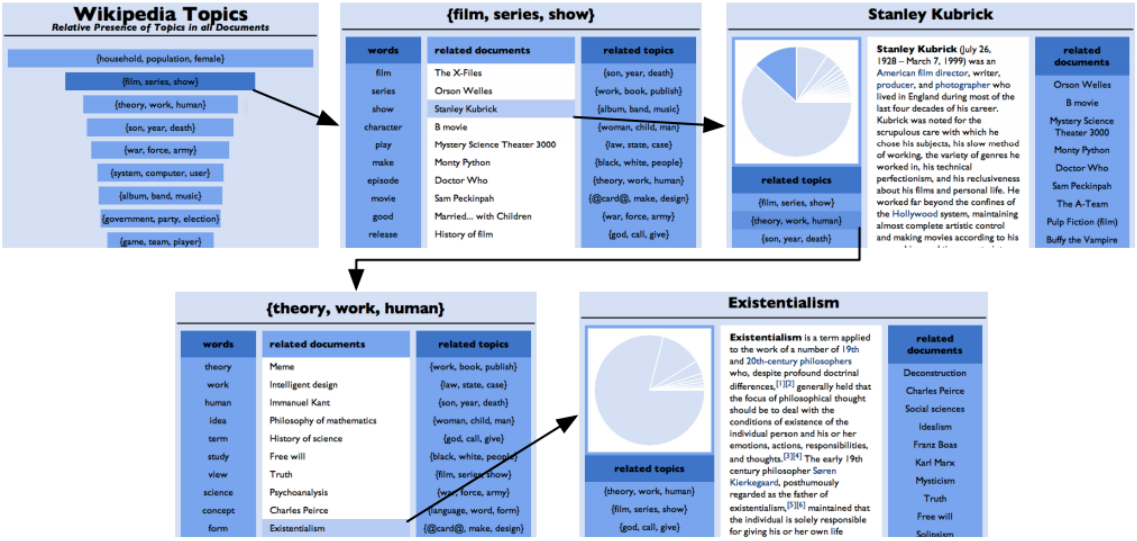


圖 2.4: Navigating Wikipedia with Topic Model

由圖中可見，最左邊第一部份的圖展示了 VTM 基於 LDA 建模後對大規模語料庫產生的主題，每個長條柱都代表了一個主題，由熱門程度依序往下排序。使用者點擊某一主題後就進入該主題展示頁面，該頁面包含了主題關鍵詞、主題相關文本，以及與該主題相關的主題。使用者可以點擊主題相關文本進入文本介紹頁面。該頁面則包含（1）文本的主題分佈比例，以圓餅圖顯示，點擊圓餅圖上的主題，則進入相對應的主題介紹頁。（2）文本的詳細內容。（3）該文本的相似文本，以長條圖代表任一文本。VTM 的主題相似度是基於主題詞項分佈的相似度計算得出，而文本間的相似度則由文本主題比例之間的相似度計算得出，整體介面的具體操作過程[2]就如同 Figure. 2.4 所述。

到了 2014 年，Carson Sievert 和 Kenneth E. Shirley 提出了 LDAvis [20]：一個互動式的主題模型視覺化方法。LDAvis 結合了 R 語言和 D3.js，以 LDA 主題模型對語料庫進行主題建模後，產生視覺化的結果分析工具。LDAvis 的作者認爲以詞彙屬於主題的機率大小來依序選取表示主題的詞項，並不是最適合的結果，於是在這裡，LDAvis 的作者提出了一個算法：

$$\text{relevance}(\text{term } w \mid \text{topic } t) = \lambda \times p(w \mid t) + (1 - \lambda) \times p(w \mid t) / p(w)$$

這算法代表的是某個詞語主題的相關性，由 λ 參數來調節。如果 λ 接近 1，

[2]This navigator can be found at http://bit.ly/wiki100–引用自http://ajbc.io/projects/papers/ChaneyBlei2012.pdf

那麼在該主題下更頻繁出現的詞，跟主題更相關；如果 λ 越接近 0，那麼該主題下更特殊、更獨有（Exclusive）的詞，跟主題更相關。所以使用者可以通過調節 λ 的大小來改變詞語跟主題的相關性。 下圖 Figure. 2.5 表示基於 LDA 模型的電影語料庫（IMDB）主題建模結果，每一個圓圈表示一個主題， LDAvis 的作者將每個主題出現的頻率，用圓圈的大小來表示，而主題之間的關連性，作者用主成分分析 (Principal Component Analysis) [23] ，將主題分佈投影到第一大主維度和第二大主維度這兩個方向上，所以每個主題（圓圈）相互之間的位置遠近，就表達了主題之間的關連性。

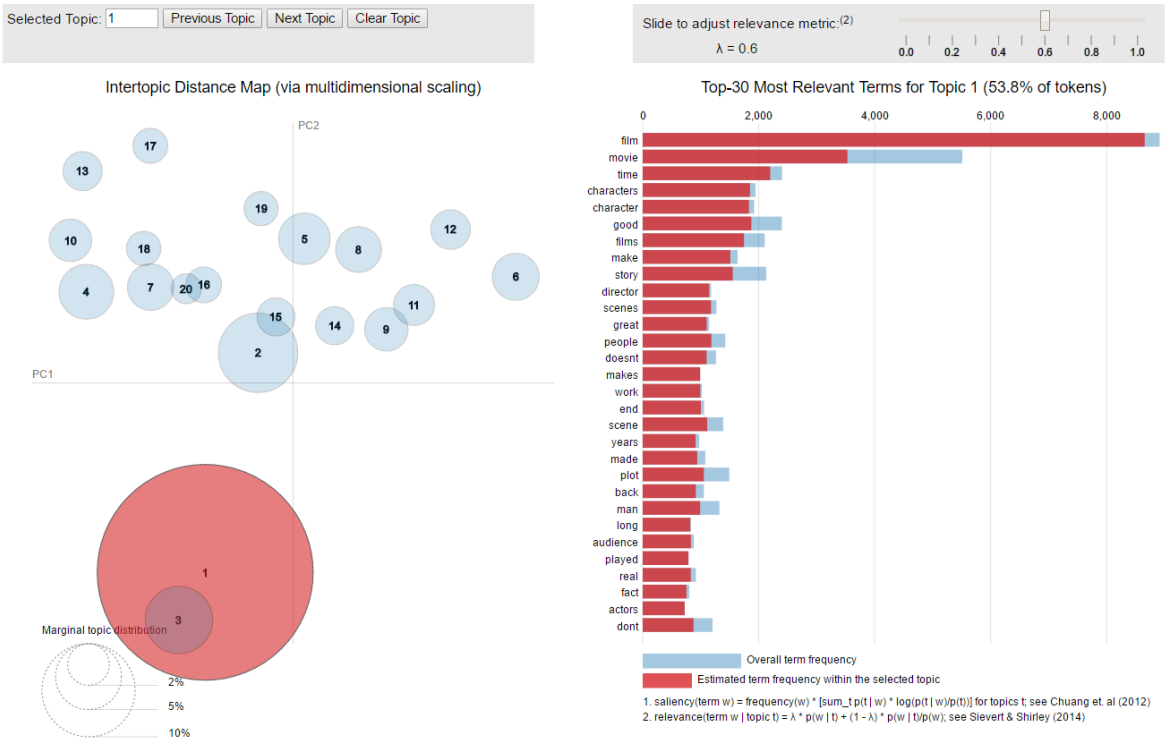


圖 2.5: LDAvis 電影分類建模圖

2.2.2 靜態主題模型視覺化總結

主題模型作爲一種能夠從語義上對大規模文檔集抽取主題的模型，提供了一種讓研究者較以往更有效率研究文檔集主題的方法 [25] 。優秀的視覺化工具可以很優秀的展示文檔、主題、詞項三者間的關係，也能爲後續的視覺化研究提供良好依據。通過本章的綜述研究可以發現，基於 LDA 主題模型的視覺化已經非常成熟，但是基於 Labeled LDA 標記式主題模型的視覺化在目前國內外的研究中可以說還是相當的少見。如何將和 LDA 主題模型相比之下多出的主題標籤資訊轉換成良好的資料視覺化應用將會是 Labeled LDA 標記式主題模型的視覺化所面臨的核心挑戰，而這也是在本研究中欲探討的主要問題之一。

第三章

介面概念與設計

隨著數位內容崛起，面對各式各類的學術資料庫與大量數位文獻，我們發現有越來越多的文史學者選擇和資訊科技人員一起合作，以數位文本作爲研究的文獻類型。身爲站在提供資訊工具幫忙文史學者的角度，我們能提供的常是一些人力難以達及的、所謂量化方面的協助，例如以主題分類演算法、分群演算法等等，這也是跨領域的最大優點之一。

但隨之而來的一些缺點，就是演算法結果有時太過晦澀，例如 LDA 演算法的大量字詞向量結果，無法直接支撐閱聽者對於文本事件合理分析、理解的應用需求,，於是本研究介面希望可以透過視覺化的方式，讓文史學者一目了然演算法跑完文本之後的結果，進而更方便讓使用者挖掘出文本不同的面向。這是本介面最主要的第一部分。

又從觀察使用者的需求發現，研究人員時常會對文本內部執行”再搜尋”的動作，回頭查找此詞彙同時出現在文本哪一部分，或是查找此作者同時寫了哪幾篇文獻等等，因此本介面的第二部分，即是和文本搜尋頁面做一結合，讓使用者查看完視覺化的結果之後，能針對主題模型內的詞立即返回文本內部，做再搜尋之動態有機連結，將文本視覺化和文本本身更緊密的連結在一起。

3.1 系統概念

本研究主要爲設計一個結合開放式數位文本平台的視覺化介面。下圖（Figure. 3.2）爲本研究的系統概念圖，主要概念爲三大部分，分別是（１）開放式數位平台、（２）文本視覺化介面、（３）後端文本處理。

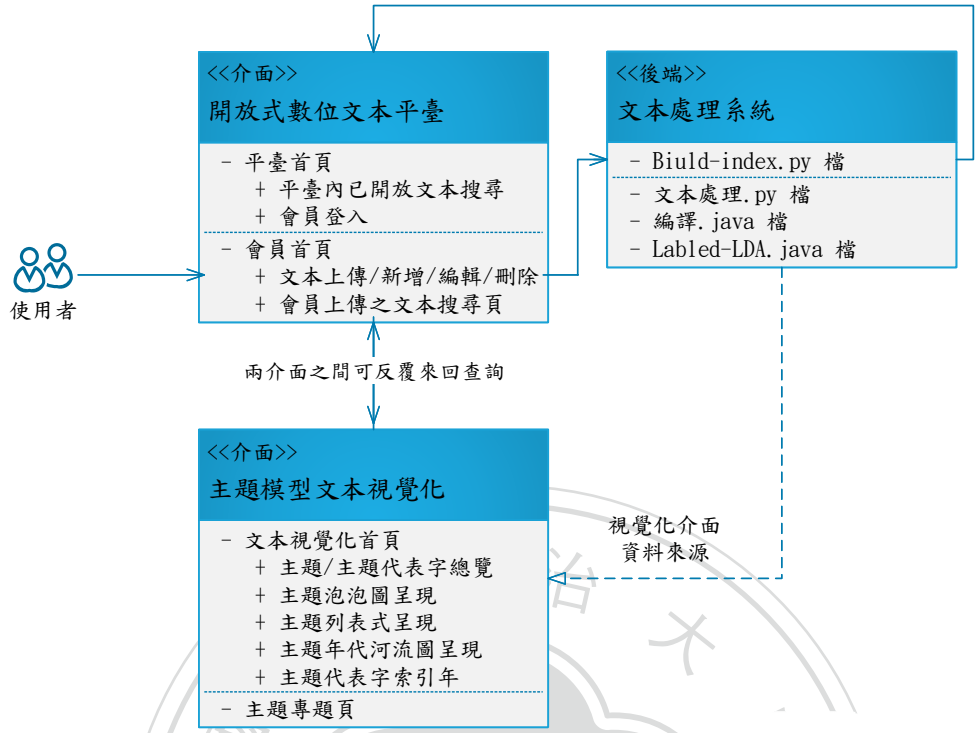


圖 3.1: 本研究系統架構圖

在開放式數位平台裡面，使用者可以上傳自己目前手邊擁有的數位文本，上傳到平台之後，除了幫助使用者儲存文本之外，還可以幫助研究學者建立起專門爲文本製作的搜尋引擎，方便使用者重複進行文本再搜尋的動作，以本研究爲例，研究團隊上傳了數位化《自由中國》文本至本平台後，可以自由搜尋包括在自由中國裡面所出現的任意關鍵字、年代、作者、專欄等等，也可在平台中同時加上上述所有搜尋條件對文本進行搜尋。再來除了搜尋之外，也可以連到自由中國的視覺化介面，去探索利用視覺化方式來闡釋《自由中國》的另一種新的文本解讀方法。

在文本視覺化介面裡面，主要利用各種不同形式的視覺化圖型，來解釋各個不同的主題概念，例如以圓餅圖來呈現整份《自由中國》裡面的主題分布；以河流圖來呈現各個主題在《自由中國》裡面各年度的趨勢走向分布；以長條圖來呈現主題底下的文字分佈比例等等。當使用者在視覺化介面看到比較不懂的字詞時，本研究設計讓使用者可以在視覺化介面與數位化平台文本搜尋引擎之間來回移動，讓使用者可以立即了解此詞彙是出現在文本的哪一部分，進行文本再搜尋的動作，搜尋過後也可以在回來視覺化介面繼續探索下一個使用者有興趣的部分。

此處後端文本處理系統引用自前人研究（陳奕安. (2016). 適用於中文史料文本之標記式主題模型分析方法研究.）。後端文本處理主要發生在使用者上傳數位化文本之後，除了幫文本進行斷詞、建立索引、建立人工關鍵字報表等等製作

搜尋引擎所需要的前置動作以外，另一方面也利用以 Python 程式語言撰寫的腳本將文本處理成 Labeled LDA 演算法所需要的格式，再利用以 Java 程式語言撰寫的 Labeled LDA 演算法來計算處理過後的《自由中國》文本。計算完畢之後，本研究的視覺化介面再利用計算之後的主題向量、詞向量結果，進行最終的視覺化呈現。

3.1.1 文本主題與關鍵字

首先，以上述章節提到之 Jieba 中文斷詞系統讓電腦自動斷完整份《自由中國》文本後，爲了因應 Labeled LDA 演算法需有人工標註之主題標籤以訓練主題模型之條件，此次一起合作的史學研究團隊特別從兩本著作：

1. 《自由中國》全二十三卷總目錄暨索引 / 薛化元主編 [1]
2. 《自由中國》與民主憲政 : 1950年代臺灣思想史的一個考察 / 薛化元著 [2]

裏頭之內容，以”民主憲政”爲概念中心挑選出適當主題，再輔佐以”經濟”、”文藝”等主題類別加以完善整體概念。其次，爲了提高訓練模型的準確度，在每個主題底下，研究團隊又從以 Jieba 斷詞工具斷出來的大量詞彙中，人工篩選出每個主題中具代表性或權重應予以提高的人工關鍵字，篩選後再餵回原有的 Labeled LDA 演算法，進行多次迭代運算後方以得到最佳的模型成果。其詳細的資料集內容則如表 3.1所示：

國立政治大學 National Chengchi University

[1] 薛化元. 《自由中國》全二十三卷總目錄暨索引. 台北：遠流出版社, 2000.
[2] 薛化元. 《自由中國》與民主憲政 : 1950年代臺灣思想史的一個考察. 稻鄉, 1996.

表 3.1: 文本資料集資訊

資料集名稱	自由中國
文本主題數	24 (包含主題類：Common) *未被規範到特定主題之文本歸類處
主題分類	Topic 0 Common Topic 1 刊物的立場與反省 Topic 2 自由民主的基本概念 Topic 3 法治 Topic 4 表現自由／出版法問題 Topic 5 其他基本權問題 Topic 6 責任閣制／責任政治 Topic 7 行政中立——國民黨退出軍警特 Topic 8 司法 Topic 9 立法院問題 Topic 10 監察 Topic 11 考試 Topic 12 在野黨 Topic 13 國民黨體質改造／國民黨問題 Topic 14 地方自治 Topic 15 地方選舉問題 Topic 16 軍隊 Topic 17 教育／救國團 Topic 18 外交／聯合國問題 Topic 19 總統三連任問題／修憲／國大 Topic 20 反共救國會議 Topic 21 反共 Topic 22 經濟／財政 Topic 23 文藝
總關鍵詞數	11,021
總文章篇數	3,651
已標記文章篇數	1,973
文本總字數	16,171,034
已標記文字總字數	7,569,758
文本資料區間	1949年11月20日－1960年9月4日

國立政治大學 National Chengchi University

3.2 介面設計流程

文字的視覺化是一種呈現文本結果的方法，其主要工作流程有三個部分（陳爲, 沈則潛, and 陶煜波. (2014) Data visualization : the most effective communication with data and design.）：（１）文字資訊採擷（２）視圖繪製（３）文字資料視覺化。具體流程則如下圖（Figure. 3.2）所示：

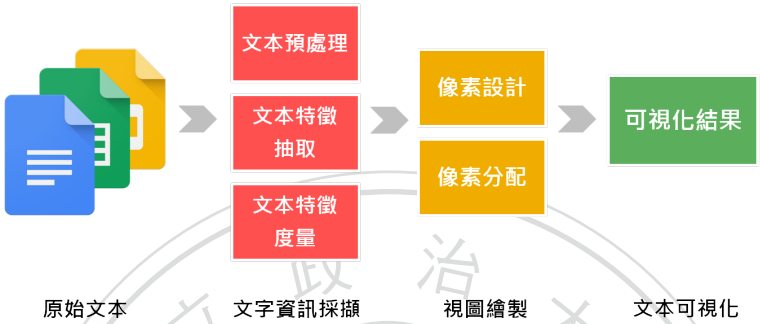


圖 3.2: 本研究介面設計流程圖

1) 原始文本：

此次視覺化的文本來源爲《自由中國》雜誌，由從中國大陸來台灣的自由主義者以及政治人物所出版。《自由中國》雜誌最初的構想是在中華民國政府撤退來台之前發生的。當時有一部分國民黨黨員和自由主義知識分子，認爲要堅定反共合法性，就必須要有一宣揚自由民主的言論機關，因此胡適、雷震、杭立武、張佛泉等人，研議創辦《自由中國》雜誌。但不久後中華民國政府暫退台灣，《自由中國》雜誌無法在大陸地區發行，遂於 1949 年 11 月在臺北創辦，由胡適擔任發行人，主要的編輯是雷震和殷海光。

2-1) 文字資料採擷－文本預處理

文字資訊的分析通常基於文字內容進行，然而，原始文字通常存在著無用且甚至干擾的資訊，以中文文本爲例，中文漢字是以字爲基本單位，因此我們以 Jieba 斷詞工具斷完整份文本之後，有些高頻的一字詞會被斷出來，例如：我、和、的等等較高頻率的一字詞。這些所謂的停用字[3]（Stop Words）不但不能代表文本意義，且會稀釋掉關鍵詞的密度，影響之後進行 Labeled LDA 演算法的演算結果。

2-2) 文字資料採擷－文字特徵的抽取

文字分析工作有時需要人工的制定的附帶標籤來輔助度量，這些特徵資訊以此次的《自由中國》文本來說，我們制定了語意級的主題標籤，希望能將文

[3] Zou, Feng, et al. ”Automatic identification of Chinese stop words.” Research on Computing Science 18 (2006): 151-162.

本透過這些主題模式加以詮釋，成爲另一種解讀文本的方法。另外也制定了人工關鍵字，透過人爲調高這些關鍵字的權重，去調整之後的演算法模型。讓整個主題底下的涵蓋主軸更爲明顯，主要意涵能更爲鮮明的跳出來。

2-3) 文字資料採擷－文本特徵度量

在有些應用環境中，使用者可能會對在多種環境下或從多個資料來源所取出的文字特徵的深層分析有興趣，以本研究來說，我們利用 Labeled LDA 演算法將抽取後的文字特徵做進一步的特徵度量，其他常見的文本特徵分析還有向量空間模型（Vector Space Models, VSM）、TF-IDF（term frequency–inverse document frequency）、詞袋模型（Bag-of-words, BoW）等等。

3) 視圖繪製－像素設計／像素分配

此階段將文字採擷所提煉出來的資訊轉爲直觀的視覺化圖型。在視覺圖型的輔助之下，使用者可以快速的取得資訊。視圖繪製常常包含兩個部份；像素設計以及像素分佈方法。優秀的像素設計需要準確無誤的承載文字的資訊特徵，如雷達圖（Radar Chart）、本次介面所使用的圓餅圖（Pie Chart）等等。像素分佈則要求有效又不失美感的分布像素在畫布上，使得視覺化表達能符合人類的感知。舉例來說，本次介面中所使用的「Pack Layout」即是其中之一。Pack Layout 利用階層性的結構把資料分類，並且在排版上使用資料數據中的「大小」來決定代表各個元素的圓圈大小，算出各主題圓圈的大小與座標之後，幫助我們整齊的排列在規範好的畫布空間裡面。

4) 文本資料視覺化

經由上述過程後得到最終的文本可視化結果，詳細的介面成果會在之後的篇幅做更詳盡的解說。

National Chengchi University

3.3 系統介面

本系統主要介面如圖（Figure. 3.3），爲了方便指認，此處以數字標註記號，從編號 1 到編號 6，共分成六大區塊。爲了讓使用者更明確的知道現在瀏覽的頁面爲哪個部份，我們以顏色做爲區分，讓各個區塊用途更明顯的分割出來，另方面也能暗示使用者目前的所在頁面其對應功能爲何。

首先是編號 1，當使用者一進來的第一個頁面－首頁 (Index)，再來是編號 2 綠色區塊的主題總覽 (Topic Overviw)、編號 3 黃色區塊的主題分布之泡泡圖 (Topic Bubble Chart)，編號 4 藍色區塊的主題列表顯示 (Topic List)、編號 5 紅色區塊的主題的年代趨勢河流圖 (Stream Graph of Topics over Year)，以及最後編號 6 橘色區塊的文字索引 (Term Index)部分。底下，將會針對各個區塊說明更詳盡的設計概念。

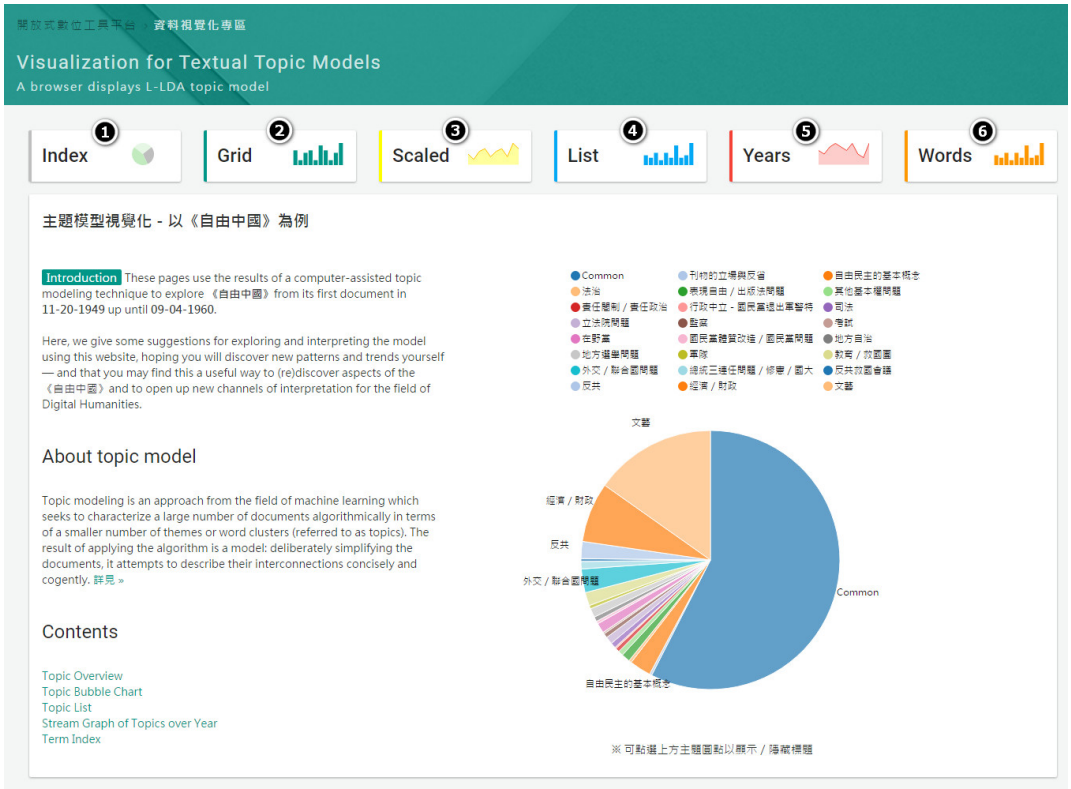


圖 3.3: 本研究主要介面

3.3.1 首頁（Index）

首先，因首頁相較其他頁面呈現了比較多的內容，考慮到使用者視覺動線的設計，因此將版面分爲左半邊與右半邊，以符合大多數人的閱讀習慣－先閱讀完左手邊的說明（Figure. 3.4）之後，再來體驗右邊的視覺化圖表（Figure. 3.5）。

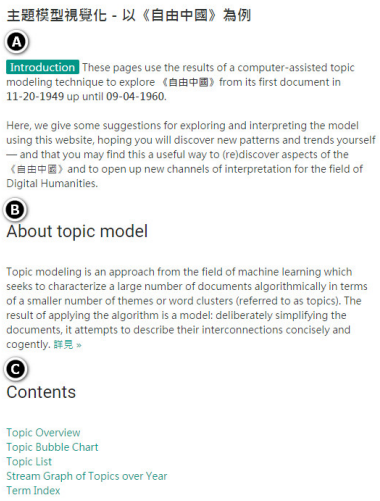


圖 3.4: 首頁說明

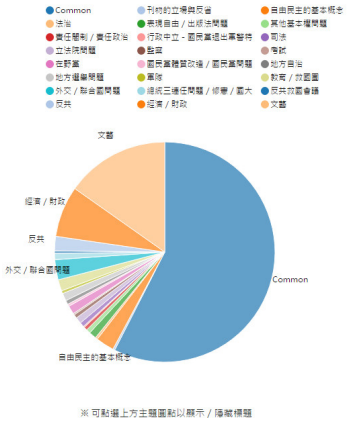


圖 3.5: 首頁視覺化部分

在左手邊的說明部分，由上往下又區分爲爲三個版面（如 Figure. 3.4 所標示的 A 、 B 、 C 三部分）。

A 部分試圖對初來的使用者說明此次視覺化文本《自由中國》的背景介紹，並說明這是一種以資訊科學的方法來輔佐人文史學家分析、閱讀文本的另一種嶄新觀點與方法。 B 部分則是爲第一次來訪的使用者，以及對本研究所使用的 Labeled LDA 演算法不熟悉的使用者提供的 Labeled LDA 演算法概略說明，因篇幅有限，故在簡短說明之後額外提供更詳盡的說明連結，待使用者有興趣或有空時仔細閱讀查看。 C 部分則是本視覺化介面的內容大綱，將滑鼠移上各個內容標題時，會有小小的提示框簡介，讓使用者對接下來即將瀏覽使用的視覺化介面有一些初步的認知。

而右手邊的視覺化部分，則大概介紹了在此次研究的《自由中國》文本涵蓋了哪一些主題，每個主題占總文本中的比率大概如何。由於大部分的文檔還是被歸類在主題 0 （Common 類），故若使用者想瞭解除了主題 0 之外的主題分布，則可點選上方主題標題，將主題 0 隱藏起來，下方圖表則會立即更新爲沒有主題 0 的圓餅圖表以方便使用者查看。

國立政治大學 National Chengchi University

3.3.2　主題／主題代表字總覽（Grid）

此頁面（Figure. 3.6）主要列舉了主題與主題代表字之間的相互關係，首先，在跑完主題建模之後，我們可以知道每個主題出現的頻率，因此我們以圓圈大小來表示主題占整份文本中的頻率。將滑鼠移到各個主題圓圈之後，中間顯示的文字清單則代表在此主題底下的關鍵詞彙，依照計算出來的主題－詞彙向量分佈，從向量最高到最低由上往下排列詞彙。另外，使用者若想知道此詞彙是否也同時出現在其他主題，也可將滑鼠移至字彙上面，即可顯示同樣包含此字彙的主題圓圈。

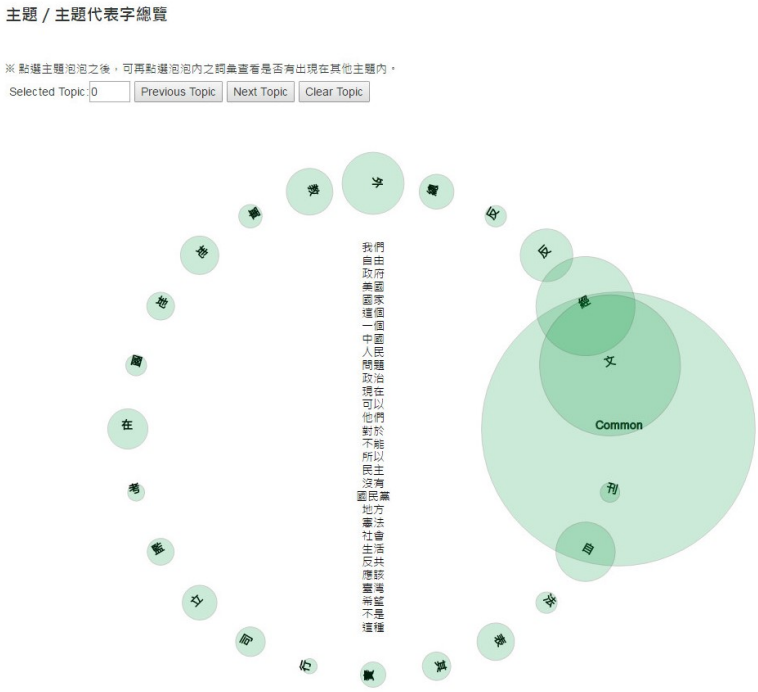


圖 3.6: 主題／主題代表字總覽

3.3.3　主題遠近分佈（Scaled）

此頁面（Figure. 3.7）主要呈現以 Labeled LDA 演算法計算出各主題分布之後，主題之間的遠近關係。主題泡泡之間的位置遠近，表達了主題之間的接近性，主題遠近的算法主要是採用多維尺度分析（Principal components analysis，PCA），提取出主成分做維度，將主題分佈到這兩個維度上，詳細的計算過程會在下一章節做更詳盡的解說。若使用者點擊任一個主題泡泡都可進入更詳盡的主題專題頁。

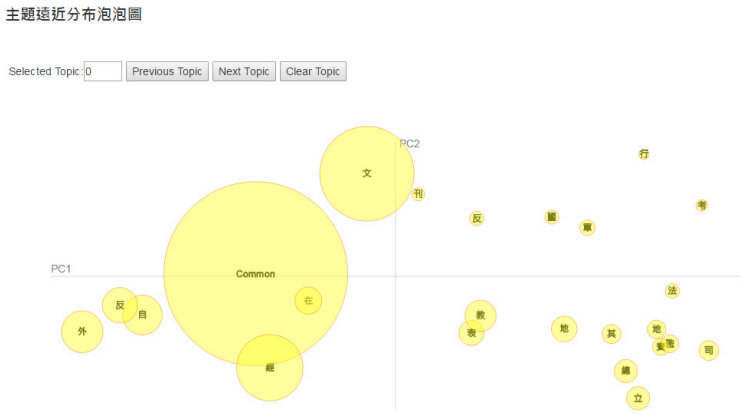


圖 3.7: 主題遠近分佈視覺化

3.3.4 主題列表（List）

在主題列表頁採用了表格方式來介紹各個主題及主題底下出現最頻繁的前十個詞彙。爲了幫助比較，每一行表格都可透過點擊行標題以進行排序（如 Figure. 3.8 所示）。

圖 3.8: 主題列表頁

最左邊是每個主題的名稱，共有 24 個主題，可按最上方的行標題切換主題排列順序。再來是主題年代分佈的微型直方圖，分別表現每個主題從 1949 年到 1960 年的比例分佈（因版面篇幅大小限制，此處每個微型直方圖的 y 軸並不一在同一個比例上。）同樣也可經由點擊最上方行標題切換圖型排列順序。

第三行呈現的則是在每個主題底下的詞彙向量組合裡最高的前十個詞彙，例如在主題 3（自由民主的基本概念）底下，詞彙向量的詞由高至低的詞分別爲：自由、民主、國家、政治、人民、思想、社會、政府、個人、反共。

最後一行呈現的則是每個主題占《自由中國》文本的比例多寡，以藍色長條圖的視覺化及右側數字來表示，可經由點擊最上方行標題來切換占最多比例的主題或者占最小比例的主題欄位。若使用者點擊任一個主題列都可進入更詳盡的主題專題頁。

3.3.5 主題年代分布（Year）

此頁面（Figure. 3.9）主要呈現以 Labeled LDA 演算法計算出文本各年代的主題機率分布，使用適合表現年代分佈趨勢的河流圖（Streamgraph）來呈現。每條河流的高低都代表了該主題在某一年的比例，並使用紅色系的漸層來表現不同的主題。上方設計了一按鈕來切換顯示有主題 0（Common 類）和沒有主題 0 的差別，預設顯示爲無包含主題 0。點擊任一河流都可讓使用者進入更詳盡的主題專題頁。

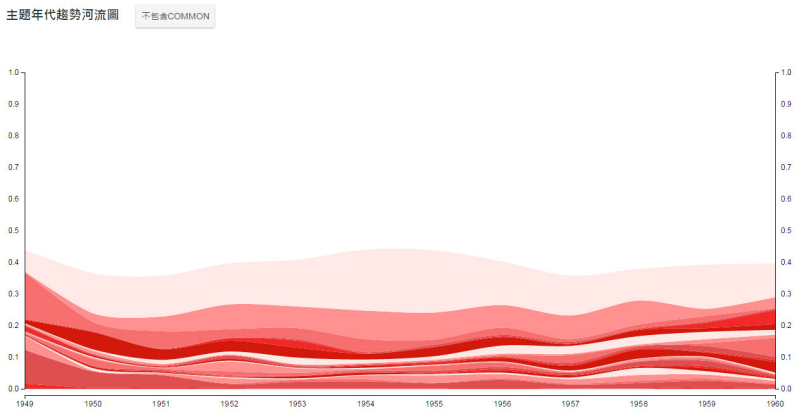


圖 3.9: 主題年代分布視覺化

3.3.6 文字索引（Index）

此頁面（Figure. 3.10）列舉了所有在主題底下被模型訓練出來的詞彙，其他同樣在文本裡被斷出來的詞彙但沒有包含在任一主題底下的，就沒有顯示在此頁面。將滑鼠移到詞彙上方時，可顯示此詞彙是被包含在哪幾個主題底下，點擊詞彙則可連結到數位文本平台搜尋頁面，進入搜尋該詞彙後的查詢結果呈現。

圖 3.10: 文字索引頁

3.3.7　主題專題頁（Topic）

在《自由中國》中，由專家所制訂的 24 個主題都有屬於自己的主題專題頁（如 Figure. 3.11 所示），目的是對於各主題做更加全面的詮釋。左側是在該主題底下，詞彙向量組合最高的前十個詞彙，也就是最能代表此主題意涵的主要詞彙，在這裡使用長條圖來表示主題中每個詞彙的頻率。右上方是各個年代的該主題年度比例，使用直方圖來表示，點選圖表上任一年份的直方圖，即可在下方顯示該年度底下的主題關鍵字，可以由此觀察出同個主題在不同年份底下，其關鍵字的變遷情形。

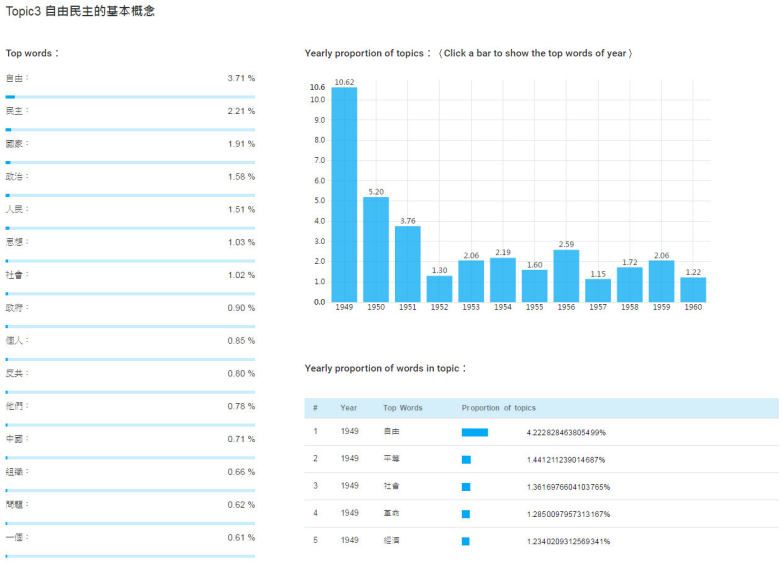


圖 3.11: 主題專題頁

ational Chengchi Universit

3.3.8　詞彙搜尋頁面（Search）

使用者可從主題專題頁的左側主題詞彙，或是文字索引頁，選擇有興趣或是想進一步了解的詞彙，點擊後進入詞彙搜尋頁面（如 Figure. 3.12 所示）。此頁面會列出所有包含此詞彙的《自由中國》文本段落，使用者可從文本脈絡中推測出此詞彙大概代表甚麼涵義，以及此詞彙當下的時空背景等等。 另外，若使用者還想再做更進一部的查詢，也可以選擇左邊的條件選單，增加額外的搜尋條件，例如查詢特定作者、專欄、或是年代、日期等等額外搜尋條件。

圖 3.12: 數位工具平台搜尋頁面

3.4 介面設計總結

文字資料的視覺化不同於其他具有空間屬性（像是地理資訊視覺化）的資料特性，缺少了空間位置等結構化的資料。因此，如何將沒有空間結構屬性的文字資料轉爲讓使用者有興趣去探索的二維視覺化結果是此次介面設計中主要的核心問題。另外，本介面的核心重點除了提供一個能與使用者友善互動的視覺化介面以外，如何加入進一步的分析查詢功能，將文本視覺化與文本本身更緊密的連合在一起，也是本視覺化介面在設計上主要的考量問題之一。

National Chengchi University

第四章

介面實作技術

本章主要介紹本研究視覺化介面的實作技術探討，共分爲三個部份。第一部分爲整體介面架構；第二部分爲本介面所有視覺化圖表之技術說明；第三部分爲主題遠近分析的計算說明。

4.1 整體介面架構

本視覺化介面整體架構是採用前端 CSS 框架「BootStrap」。BootStrap 是一個基於 HTML、CSS 和 JavaScript 的框架，由 Twitter 的 Mark Otto 和 Jacob Thornton 所開發。由於它全開源且支持全響應式網頁的特性，所以很適合用來開發符合各種不同大小設備的頁面，避免因爲兼容性而導致的大量重複勞動，也是目前最受歡迎的前端 CSS 框架之一 [7]。本視覺化介面所採用的 BootStrap 版本爲 v3.3.6，目前最新的版本爲 v4.0，但仍在 Alpha 測試當中。

4.1.1 整體顏色設計

本視覺化介面整體顏色設計是採用 Google 在 2014 年所推出的 Material Design [1]（質感設計，又譯爲材質設計、材料設計），是由 Google 所開發的設計語言。根據其設計的 Guideline [2] 內容，Material Design 的色彩靈感是來自於設置在寧靜場合之中強調色的標示，參考當代建築、道路標誌、地面標線以及體育場地給的啓發。強調明顯的陰影及聚光，使用意想不到的鮮豔色彩 [9]。

Material Design 提供了 500 種基本顏色作爲基本色票以提供 Android 系統、網頁及 iOS 系統使用。設計者可選擇三種以內的顏色當作基本色，並額外選擇一種顏色作爲強調色。其線上工具可以非常快速的讓設計者任意選擇兩種基本色和強調色，並自動搭配、顯示在頁面右上方的預覽圖裡，如圖（Figure. 4.1）。另外，除了顏色大方向的呈現以外，也會提供同色系但更深、更淺的顏色組合，以及可用於文字、圖示、分隔線、主要文字、次要文字等等的配色建議。

[1] https://material.io/guidelines/material-design/introduction.html
[2] https://material.io/guidelines/

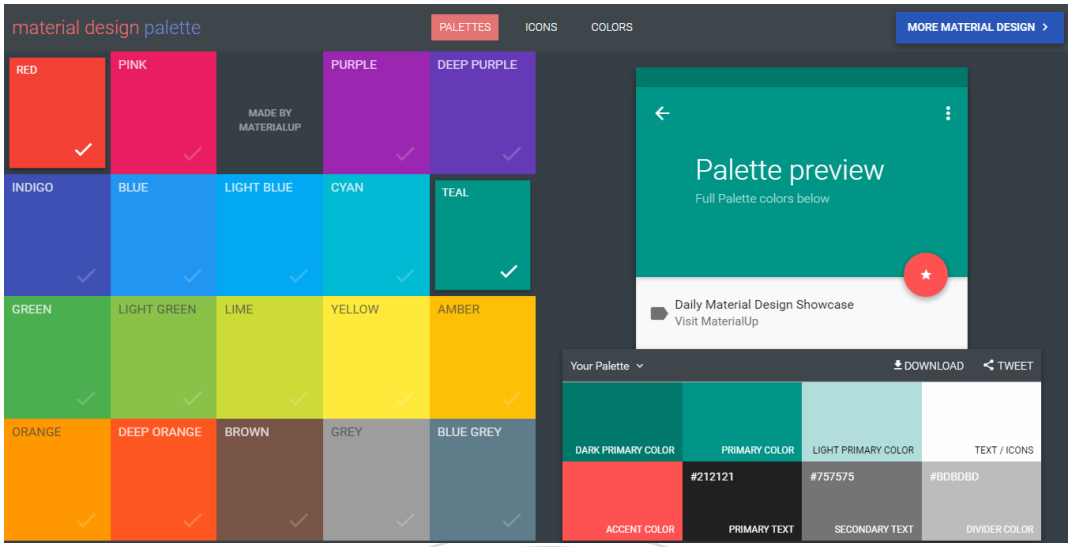


圖 4.1: Material Design 線上工具

以本研究介面之開放式數位平台（Figure. 4.2）爲例：其介面之基本色（圖片標記 A ）爲 Teal 色盤底下的 Primary Color（色碼：#009688），而介面之強調色（圖片標記 B ）則爲 Red 色盤底下與之對應的 Primary Color（色碼：#F44336）。標題文字和次要文字（圖片標記 C、D）則分別使用 Material Design 所規範的文字顏色：（色碼：#212121、#757575）。而放在基本色上面的文字顏色（圖片標記 E ）則採用 Guideline 所規定的純白色文字（色碼：#FFFFFF）。

圖 4.2: 以開放式數位平台頁面爲例實現 Material Design 之頁面呈現

4.2 視覺化圖表技術說明

本次視覺化介面裡所有的互動圖表皆採用「D3.js」來繪製。 D3.js（or just D3 for Data-Driven Documents）是一個用於可操縱數據資料與圖形顯示的 JavaScript 函式庫，它允許綁定任意數據到文本對象模型（Document Object Model， DOM），即 DOM 物件中。然後將數據進行運算轉換後，應用至文檔（Document） 中，運用 CSS3、HTML 以及 SVG 來實現視覺化展示。適用於 D3.js 的數據可以是各種

格式的，最常見的有 JSON 格式、逗號分隔值（CSV）或是 GeoJSON 格式等等。D3.js 始於 2011 年，而開發者則來自於 2009 年史丹福大學的視覺化計劃團隊成員－ Jeffrey Heer 教授和其實驗室底下的學生所共同開發 [4]。本視覺化介面所採用的 D3.js 版本爲 v3.1.5，目前最新的版本爲 v4.4.2，由於兩者寫法不同，故開發者通常在兩者之間選擇其中一種來實現整體的視覺化介面。

4.2.1 各式主題模型視覺化技術呈現

本視覺化介面之主題比例分布採用圓餅圖（或稱餅狀圖）的視覺化方式呈現。因爲圓餅圖常用於描述量、頻率或百分比之間的相對關係，又是最容易讓使用者直接理解的圖表之一，故採用圓餅圖的方式來呈現在「自由中國」文本中各個主題的比例分布情形。其圓餅圖的原始資料格式如下圖 Figure. 4.3 所示（此處以文藝類數據爲例展示 D3.js 在圖表繪製中所扮演的功能）：

圖 4.3: 圓餅圖原始數據

經過 D3.js 的圓餅圖布局函式轉換之後：

```
var pie = d3.layout.pie()
    .value(function(d) { return d.value; });

""""
in d3.js, If your function requests both d and i,
then d will refer to the data,
and i will correspond to the index.
""""
```

Coding 4.1：D3.js Pie Layout 3.x API

就會得到轉換原本文藝類 value 後的新數據，可拿來將其進行繪製成圓餅圖（如 Figure. 4.4 所示）。

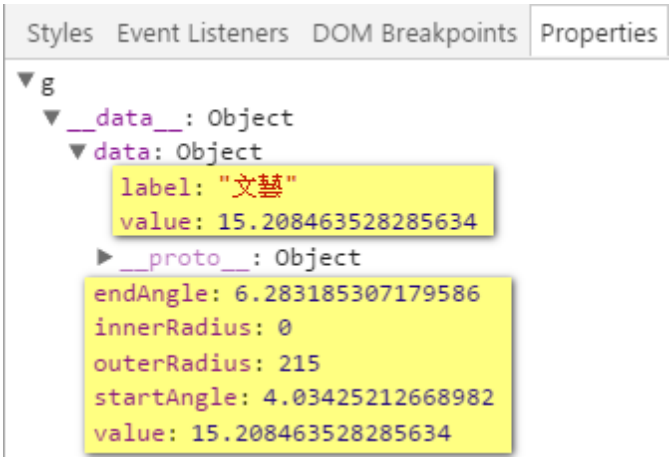


圖 4.4: 經 D3.js 轉換過後得到的圓餅圖新數據

如上圖所示，5 個整數被轉換成了 5 個物件（Object），每個物件都有變數起始角度（startAngle）和終止角度（endAngle），還有原資料（屬性名稱爲 data）。而這些都是繪製圓餅圖所需要的資料。

另外在本視覺化介面中，主題和主題底下關鍵字的呈現有兩種方式，分別是將主題以泡泡圖的方式呈現，並將主題關鍵字依機率大小由上而下排列顯示在中間，當滑鼠指向不同的主題泡泡時，中間的主題關鍵字就會隨之改變。另外當使用者按下任一主題關鍵字時，也支持使用者外連到開放式數位平台做立即性的查詢。另外一種呈現方式則是以條列式的方式呈現主題和主題底下關鍵字（如 Figure. 4.5 、 Figure. 4.6 所示）。

圖 4.5: 條列式呈現主題和主題底下關鍵字

上述介紹的兩種視覺化呈現方式其原始資料來源 data 皆是由 Labeled-LDA 主題模型跑完整份自由中國文檔後再進行後續整理所得到的資料，也就是說兩者所對應的資料來源是同一份以 JSON 格式所儲存的檔案。該檔案紀錄了包括當滑鼠移到中間的主題關鍵字時，隨之對應的主題圓圈顯示與否的視覺化呈現、主題圓圈的半徑大小、各關鍵字所對應的主題等等。

在條列式呈現中的主題占文本比例橫柱圖（如 Figure. 4.7），以及主題底下關鍵字比例橫柱圖（如 Figure. 4.8 ），是以 SVG（可縮放向量圖形, Scalable Vector Graphics）來繪製。SVG 是用於描述二維向量圖形的 XML 標記語言。基本上，SVG 用於圖形，而 XHTML 用於文字。此外 SVG 也有 DOM、事件，並且可由 JavaScript 存取，故可讓開發者建立豐富的動畫和互動式圖形。

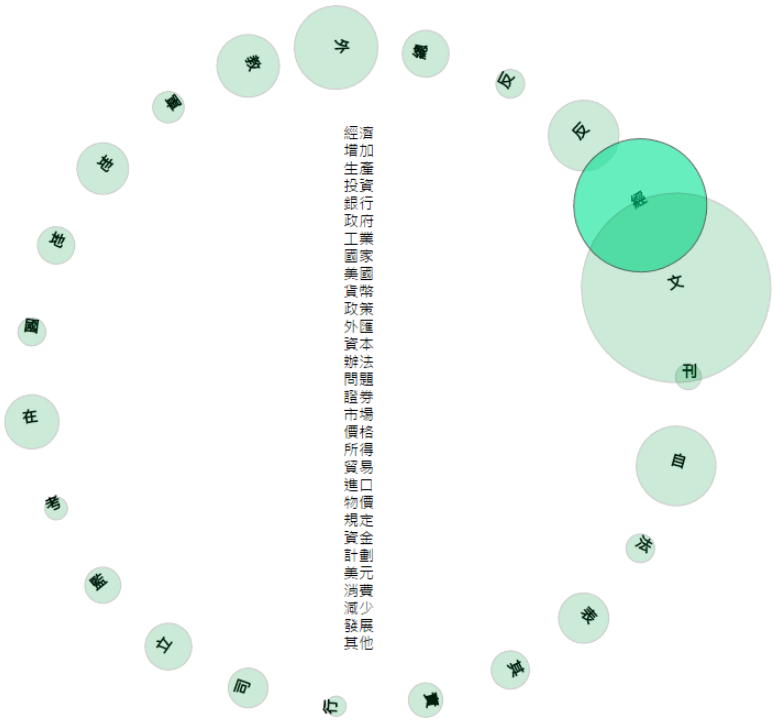


圖 4.6: 泡泡圖呈現主題和主題底下關鍵字

National Chengchi University

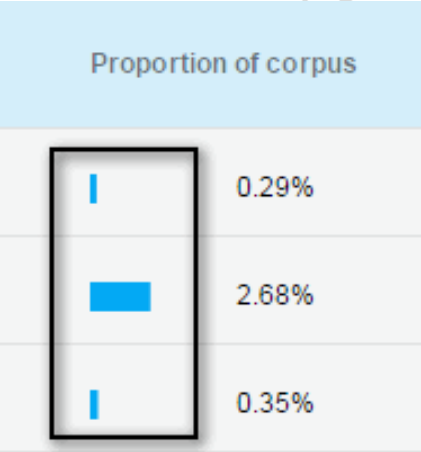


圖 4.7: 主題占文本比例橫柱圖

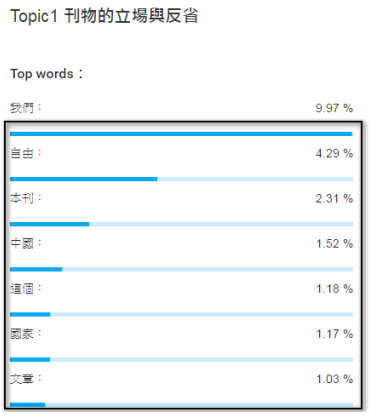


圖 4.8: 主題底下關鍵字比例橫柱圖

而關於不同主題隨著年代的變化，在本研究裡則以面積圖（Area Chart）的方式來呈現（如 Figure. 4.11 ）。面積圖常用於表示隨時間使用數字或百分比（在此情況下又被稱爲堆疊面積圖（Stacked Area Chart））的累計總計，並使用圖型面積顯示相關屬性中的時間趨勢 [24] 。它能顯示一個圖型的量化數據，在軸和線之間通常會用顏色將其填滿。同時它也是一個可交互式的視覺化圖表，揭示了在 1949 年到 1960 年自由中國文本的主題趨勢和所占的比例。點擊特定的主題河流將進一步鏈接到開放式數位平台頁面。其面積圖的原始資料格式如下（如 Figure. 4.9 ）：

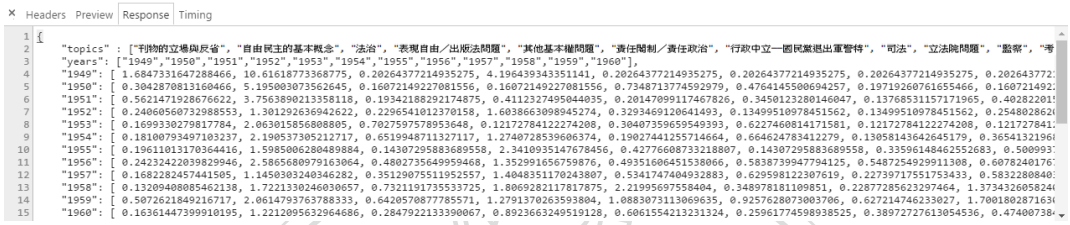


圖 4.9: 面積圖原始資料格式

再利用 D3.js 的路徑數據生成函式幫我們轉換原始資料使之成爲可以畫成面積圖的新數據。

```
var area = d3.svg.area()
    .interpolate("cardinal")
    .x(function(d) {
      return x(d.year);
    })
    .y0(function(d) {
      return y(d.y0);
    })
    .y1(function(d) {
      return y(d.y0 + d.y);
    });

"""
in d3.js, If your function requests both d and i,
then d will refer to the data,
and i will correspond to the index.
"""
```

Coding 4.2：D3.js Area Layout 3.x API

上述 D3.js 程式碼中的 .interpolate（"cardinal"）可以幫助我們把數據轉換成「樣條函式」（Cardinal Spline）。樣條是一種特殊的函數，由多項式分段定義。在計算機輔助設計和計算機圖形學中，樣條通常是指分段定義的多項式參數曲線。由於樣條構造簡單，使用方便，擬合準確，並能近似曲線擬合和交互式曲線設計中複雜的形狀，使得樣條在上述領域中是曲線的常用表示方法 [18] 。另外在繪圖軟體上經常使用的「貝茲曲線」即爲典型的樣條函式之一。

經 D3.js 轉換過後的數據可得到一個新的 y 座標（y0），若以圖（ Figure. 4.10 ）爲例，意即在 1949 年主題爲「經濟／財政」的面積圖其 y 座標要由原始 y 座

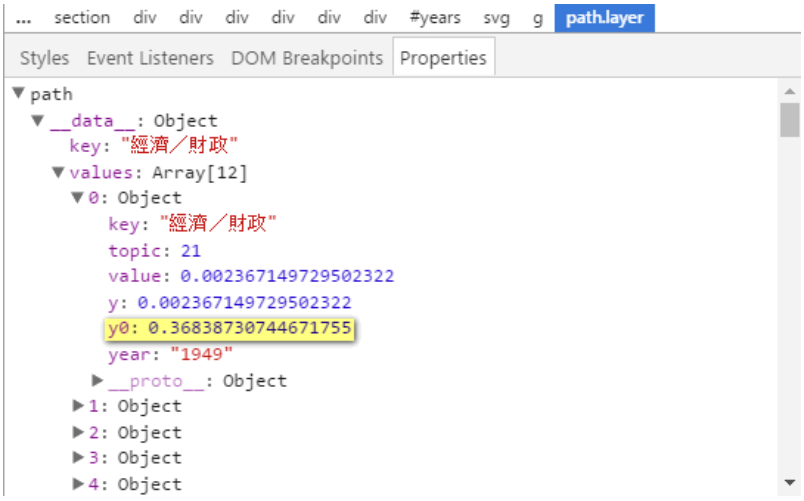


圖 4.10: 經 D3.js 轉換過後得到的面積圖新數據

標往上繪製到新的 y0 座標，若將程式碼中的 y0 改成 y1，將面積圖將會從往上繪製，變成往下繪製。

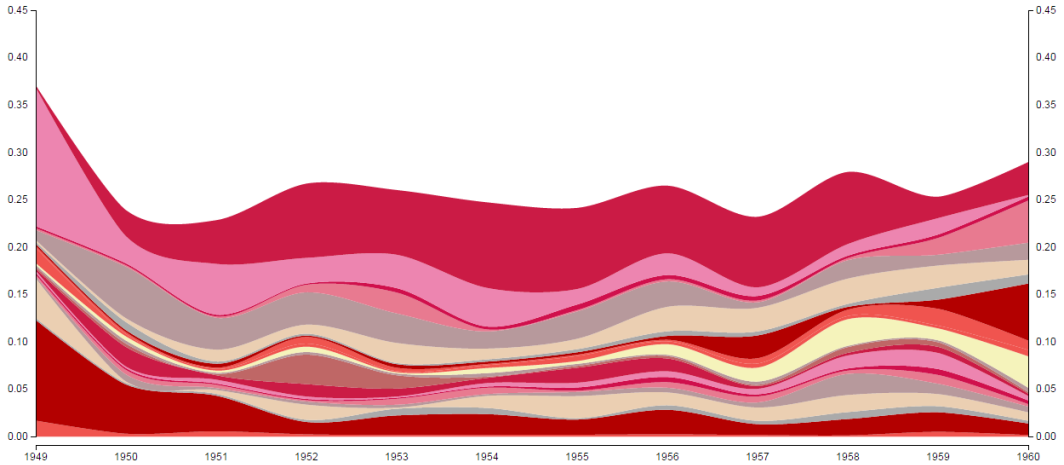


圖 4.11: 面積圖文本主題視覺化

4.2.2 其餘靜態圖表技術

在本視覺化介面裡面，除了可互動的交互式圖表，也有其他靜態的小型迷你圖表（如 Figure. 4.12 ），此類圖表可以在空間不多時傳達良好的視覺意涵給使用者，以節省若用文字表達所需的大量空間。在本視覺化研究裡的靜態小型迷你圖表都是使用「Piety.js」(註)所製作。Piety.js 是一個迷你圖表製作工具，只要給定資料就可以在瀏覽器上呈現所需要的圖案。而資料的綁定來源除了可在 HTML 裡直接給定數據以外，也可以透過 JavaScript 來設置資料。 Piety.js 目前支援使用者繪製長條圖、折線圖、以及圓餅圖等多種圖表，也支援使用者利用事件回傳更新圖表、自定意圖表色彩等等進階功能。

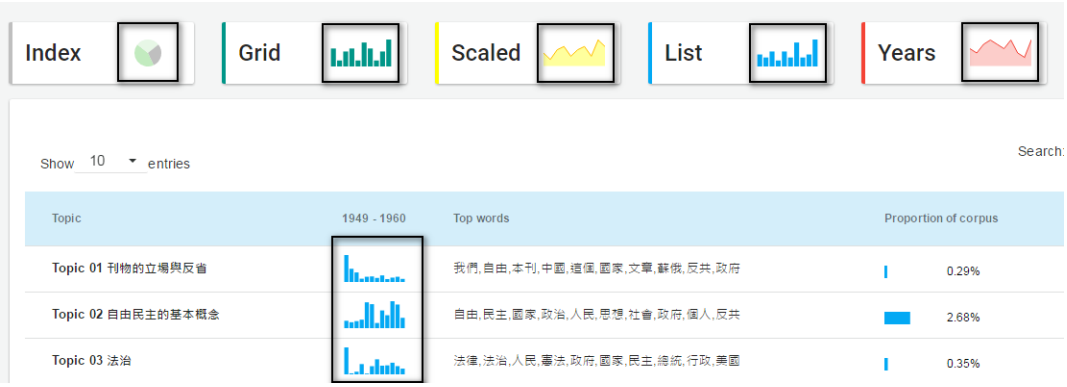


圖 4.12: Piety.js 小型迷你圖表

4.3 文本主題遠近分析技術

此視覺化介面的做法是參考之前在第 2.2 節提到的 LDAvis [20]：一個互動式的主題模型視覺化方法。基於 LDA 主題模型對語料庫進行主題建模後，將每個主題出現的頻率，用圓圈的面積大小來表示，所以圓圈的大小表示主題出現的頻率。而主題之間的關連性，作者先是利用了 Jensen–Shannon 散度（Jensen-Shannon Divergence）[3]去測量兩個機率分佈之間的相似性（在本研究中，指的是依序計算倆倆之間在自由中國文本裡 24 個主題分佈的相似性）。再利用主成分分析法（Principal Components Analysis，PCA）[4]，一種分析、簡化數據集的技術，經常用於減少數據集的維數，同時保持數據集中的對變異數貢獻最大的特徵。將經過計算後的 24*24 維的主題矩陣降維至二維空間，並分佈投影到第一大主維度和第二大主維度這兩個方向上以求得每個主題的（x , y）座標。所以每個主題圓圈在畫面上所呈現出來的位置遠近，就表達了主題之間的關連性。

[3]Fuglede, Bent, and Flemming Topsoe. ”Jensen-Shannon divergence and Hilbert space embedding.” Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on. IEEE, 2004.

[4]Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine. 1901, 2 (6): 559–572.

```python
from math import log
from numpy import array

class JSD:
    def KLD(self,p,q):
        if 0 in q :
            raise ValueError
        return sum(
            p * log(p / q) for (p,q) in zip(p,q) if p != 0
        )

    def JSD_core(self,p,q):
        M = [0.5 * (p + q) for p, q in zip(p,q)]
        return 0.5 * self.KLD(p,M) + 0.5 * self.KLD(q,M)
```

Coding 4.3：JSD Python 算法實現

此處的 p、q input 分別代入以 Labeled-LDA 主題模型跑完自由中國文本之後所得到的 Topic-Term Distribution（也就是跑完 Labeled-LDA 主題模型之後所產生的 .phi 檔），去計算倆倆主題分佈之間的距離。下圖爲跑完 Jensen–Shannon散度（Jensen-Shannon Divergence）後所得到的結果。

```
dist_matrix:
24x24 distance matrix
IDs:
'0', '1', '2', '3', '4', '5', '6', '7', '8', '9', '10', '11', '12', '13', '14', ...
Data:
[[ 0.          0.43871391  0.4565464   0.48522451  0.50727014  0.47562078
   0.46652081  0.44713493  0.46349229  0.47213877  0.46978571  0.45482529
   0.45264649  0.44965028  0.45339272  0.44821366  0.4336386   0.43615317
   0.43247391  0.431192    0.43027164  0.42623307  0.42601558  0.42465581]
 [ 0.43871391  0.          0.47869468  0.49029352  0.50711994  0.50663934
   0.50100192  0.48926669  0.49461057  0.48520803  0.49199971  0.48596537
   0.47184224  0.47343963  0.47035716  0.46799889  0.46463925  0.45812732
   0.45511454  0.45452433  0.45329845  0.45080229  0.45054959  0.44968183]
 [ 0.4565464   0.47869468  0.          0.25912155  0.46167846  0.43838821
   0.42003042  0.4083397   0.40652124  0.39710022  0.38801532  0.39510474
   0.37522687  0.37476816  0.36578494  0.36342058  0.35620465  0.35131162
   0.34568975  0.34280682  0.34307628  0.33992347  0.33936922  0.33855445]
 [ 0.48522451  0.49029352  0.25912155  0.          0.41169279  0.38969995
   0.36009153  0.36012042  0.35864527  0.33531744  0.3270127   0.33272547
   0.30721489  0.30637221  0.28878925  0.29081406  0.28526045  0.27515132
   0.26823485  0.26670528  0.26473047  0.26043805  0.26007348  0.25879467]
 [ 0.50727014  0.50711994  0.46167846  0.41169279  0.          0.38820959
   0.36938056  0.37020783  0.36415919  0.33396226  0.34250136  0.33993862
   0.31593976  0.30571907  0.3081347   0.30216467  0.29658201  0.27890919
   0.27971122  0.27930224  0.27731125  0.27304227  0.27262105  0.27157272]
 [ 0.47562078  0.50663934  0.43838821  0.38969995  0.38820959  0.
   0.34814608  0.32765856  0.31702369  0.31280672  0.26715906  0.29380469
   0.24144796  0.2705709   0.27936733  0.26797013  0.26418042  0.25316385
   0.24499029  0.24606094  0.24353198  0.23964902  0.23860501  0.23812921]
 [ 0.46652081  0.50100192  0.42003042  0.36009153  0.36938056  0.34814608
   0.          0.32425115  0.30934646  0.27165124  0.27007249  0.27443403
   0.24438952  0.2415233   0.2371063   0.2248805   0.21892686  0.20531975
   0.19540946  0.19459659  0.19231867  0.18734993  0.18722484  0.18564625]
```

圖 4.13: Jensen–Shannon Result

得到矩陣結果之後，爲了降低維度，並投影繪製到二維空間，故再進一步的利用主成分分析法（Principal Components Analysis，PCA）減少數據集的維數，同時保持數據集中的對變異數貢獻最大的特徵。

```python
    """
    Principal Components Parameters
    ---------------------------------
    distributions : array-like, shape ('n_dists', 'k')
    Matrix of distributions probabilities.

    Returns  pcoa : array, shape ('n_dists', 2)
    ---------------------------------
    """
    dist_matrix = DistanceMatrix(dist.squareform
        (dist.pdist(distributions.values, _jensen_shannon))
    )
    if skbio_old:
        data = PCoA(dist_matrix).scores()
        mds_res = data.site[:,0:2]
    else:
        mds_res = pcoa(dist_matrix).samples.values[:,0:2]
```

Coding 4.4：PCA Python 算法實現

下圖爲跑完主成分分析法（Principal Components Analysis，PCA）後所得到的結果。


```
{'mds_res': [[-0.17531122436420396, 0.037013980297954496], [-0.05215008949322095, 0.19157847060791633], [-0.2611786407175894, -
0.043121652659304725], [0.14054698198479368, 0.003557749579065722 3], [-0.011725991976306147, -0.0780033152623528], [0.094526533
10575462, -0.0803414803135211], [0.13196418277703448, -0.10475292228259726], [0.11910537194295058, 0.26909329491788286], [0.168
25255299530434, -0.11239907486967562], [0.11465033731352677, -0.20487457471313622], [0.13872960634319212, -0.0988744885631982
5], [0.16295824049370763, 0.1699697692993765], [-0.13537510198447247, -0.01503843856588306 6], [0.04916887421103366 4, 0.14723711
221463023], [0.12864823975714332, -0.07116235893484174], [0.05859069306335107 4, -0.07011747420043851], [0.07620418990231961, 0.
126472243419366 66], [-0.004819752026565037, -0.04477370844489898], [-0.3068237799322424, -0.07576835405512435], [0.105391254739
00336, -0.15212556469370628], [-0.007644330863362812, 0.1444849304467531], [-0.2782068781904051, -0.024146862967565037], [-0.16
456717155214629, -0.14564598575112928], [-0.09093409752860107, 0.2317387054944276]]}
```


圖 4.14: Principal Components Analysis Result

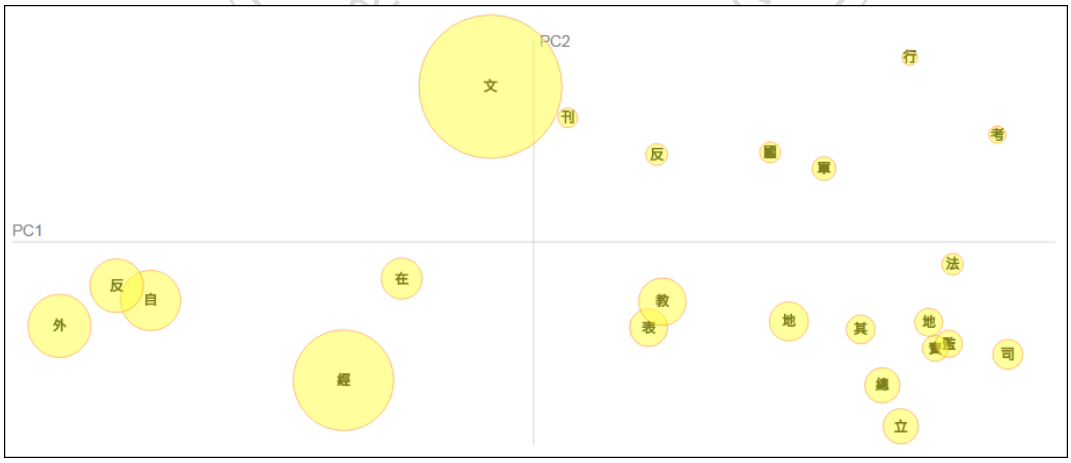


圖 4.15: 文本主題距離遠近泡泡圖呈現

第五章

實驗設計與結果評估

根據本研究在第三章和第四章的視覺化介面設計與實現，爲了探究此介面的設計對於實務上輔助文本研究的可行性，以及探究對於文本熟悉程度不一的使用者快速掌握大規模文本的幫助與否，本研究設計了一套實驗流程，透過實驗評估結果來瞭解本介面對使用者的實際價值，以及其背後的使用行爲、使用需求與動機。 在本實驗中，使用者先透過數個熟悉介面的前置引導任務之後，再進行讓使用者自行探索的指定任務。而上述任務的使用情形以及結果評估會在本章節進行後續說明。

5.1 實驗目的

本研究的實驗目的如下：

1. 對於已經熟悉「自由中國」文本的使用者，本視覺化介面是否可以在實務上實際的輔助使用者進行分析、研究，並藉由本視覺化介面和開放式數位平台的結合，進行資料的探索與收集，進而挖掘出和原本研究脈絡不盡相同、並可供探討之新型態議題。

2. 對於原本非該領域專家學者，也不具有歷史背景的使用者們，使否可以藉由本視覺化介面，快速的掌握大規模文本的大致內容方向與核心意涵，並觀察文本資料視覺化對於提升文本內容瞭解程度的有用性，以及各式視覺化呈現對使用者的幫助與接受程度。

5.2 實驗對象

因應上述之實驗目的，本研究的實驗對象將分爲兩群，分別爲具有歷史背景以及對「自由中國」文本進行研究過的「組別一」，以及完全沒有讀過「自由中國」文本，本身也非相關領域的「組別二」。各群人數皆爲五人，總實驗人數爲十人。本實驗將會從這兩群在文本熟悉程度上有高低落差的的使用行爲差異來進行分析，觀察其使用需求、行爲及使用感想以評估本研究所設計的實驗目的。

其中受測者的數量是參考開創網站可用性（Website Usability ）的使用者研究專家－雅各・尼爾森（Jacob Nielsen）的研究 [15]，尼爾森分析自己做過的 83 次網站研究，並把 5 人、10 人、20 人的訪問結果畫成圖表之後發現，在訪問 5 個人之後，已經可以觀察到網站裡 85% 的問題，而與其花時間再找 20 人找出剩下的 15% 問題，不如先處理已經發現的問題，之後再做測試，會是比較值得且較有投資報酬率的做法。

5.3 實驗流程

實驗正式開始前，爲求受測者能夠先初步了解本實驗之內容與目的，會先講解一份有關實驗內容的前導簡報，該份簡報內容共十三頁，每次講解時間大約爲十分鐘。簡報的內容總共分爲兩部分，第一部分爲主題模型簡介，讓使用者對主題模型的概念不至於這麼陌生，也比較容易瞭解何謂「主題模型的文本資料視覺化」。 第二部分爲本視覺化介面的各個視覺化呈現基本介紹，讓使用者對本研究介面有個初步的瞭解。在講解結束後則開始正式實驗，實驗過程主要共分爲五個階段：（１）實驗前測問卷，該問卷包含受測者基本資料填寫與一份實驗前測問卷、（２）第一階段介面引導任務、（３）第一階段易用性調查問卷填寫、（４）第二階段指定任務、以及（５）第二階段問卷填寫，包含有用性調查問卷、實驗前測問卷的再次填寫、以及開放式問題等（如下圖 5.1 所示）。詳細的問卷內容會在稍後的章節裡做更詳盡的說明。

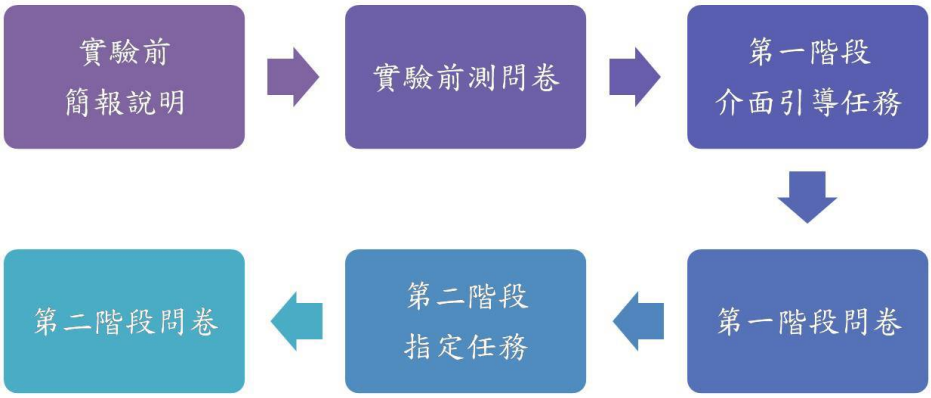


圖 5.1: 實驗流程圖

而爲了瞭解本視覺化介面是否能達成上述的實驗目標，問卷主要內容的設計參考了由美國學者戴維斯（Fred D. Davis, 1986）根據理性行爲理論（Theory of Reasoned Action，TRA）[19] 發展而來，用於解釋和預測人們對科技技術的接受程度的技術接受程度的模型－技術接受模型（Technology Acceptance

Model，TAM，又稱科技接受模型）[1]，專門探討使用者在接受一個新資訊系統時候所產生的行爲。

技術接受模型主張：人對資訊科技的使用受其行爲意圖的影響，也進而影響對資訊科技系統使用的情況（Davis, 1989）。該理論認爲，當用户面對一個新的技術時，感知有用性（Perceived usefulness，簡稱 PU ）和感知易用性（Perceived ease-of-use，簡稱 PEOU）是兩個主要的決定因素 [12] ，如下圖 Figure. 5.2 所示。

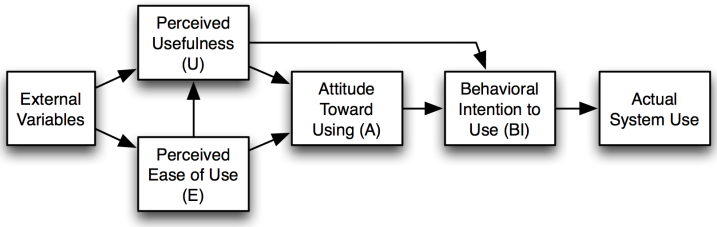


圖 5.2: Technology Acceptance Model 科技接受模型

故在第一階段，當使用者做完介面引導任務之後，本研究以評估系統易用性（ease-of-use）的目標來設計問卷，在第二階段當使用者完成指定任務後，本研究以評估系統有用性（usefulness）的目標來設計問卷。希望透過上述包含易用性與有用性的綜合評估，更深刻的瞭解使用者在本視覺化介面上面的用户體驗與使用感受。

5.3.1 實驗前測問卷

該份問卷包含了四題（問題 1、2、3、4）用來紀錄使用者基本資訊的問題（如表 5.1），以及兩題（問題 5、6）紀錄使用者在實驗之前，對利用文本資料視覺化輔助文本研究、以及對於用主題模型輔助文本研究的看法偏向同意與否。在實驗結束後會請使用者再填寫一次實驗前測問卷，測量實驗前和實驗後，當使用者對於文本資料視覺化、主題模型在文本上的應用的熟悉程度提升之後，對於文本資料視覺化、主題模型輔助文本研究的看法，以及對於「自由中國」文本、主題模型視覺化等的熟悉度是否會略有改變。

[1]Davis, Fred D., Bagozzi, Richard P. and Warshaw, Paul R. (1989) User acceptance of computer technology: a comparison of two theoretical models. Management Science. 35(8), p.985.

表 5.1: 實驗前測問卷問題列表

問題 1	對「自由中國文本」的熟悉與瞭解程度爲？
問題 2	對「主題模型」的熟悉與瞭解程度爲？
問題 3	對「文本的資料視覺化熟悉與瞭解程度爲？
問題 4	是否有使用過任何搜尋系統搜尋任何電子化文本？
問題 5	從「文本文字的量化特性」的角度去挖掘歷史問題是可行的。
問題 6	文本視覺化有助於發現一些在文本中被突顯出來的概念或問題點。

除了問題 4 請受測者填寫使用次數以外，其餘問題皆使用李克特五分量表（Likert Scale）[2] 來測量受測者對於該問題描述同意或不同意的程度。典型的李克特量表將受試者的判斷結果分爲五種：非常不同意、不同意、無意見（即：不同意也不反對）、同意、非常同意。根據陳述語句的傾向給予各等級不同分數。對正向陳述而言，答案越正向分數越高（如：「非常同意」爲 5 分，「同意」爲 4 分，以此類推）；對負向陳述而言，答案越負向分數越高（如：「非常不同意」爲 5 分，「不同意」爲 4 分，以此類推）[13]。

5.3.2 前置介面引導任務

爲了讓使用者能更加熟練的操作每個部份的介面功能，本研究設計了一個互動式的引導式問題網頁（如圖 5.3 所示），爲了方便作答，將問題設計在視覺化介面的上方，讓使用者在螢幕上即可完成操作任務，不用一直低頭看紙本或來回切換網頁，並能確保讓使用者能對每個視覺化介面都做到基本的操作訓練和說明。在訓練的過程中，施測者會在旁觀察並且紀錄受測者在操作時，所遭遇的問題。

該引導式問題網頁共包含了十九個問題，每部分的介面操作問題約在兩至三題之間，問題當中的第（4）、（13）、（14）、（19）題爲熟悉此視覺化介面與開放式數位平台之間的互動和使用操作。此實驗流程主要目的是希望能透過這十九個介面操作的小問題讓受測者在完成任務的過程當中更加的熟悉本視覺化介面。而詳細的問題內容爲避免篇幅過長，將收錄在附錄 A 的部分供作參考。

[2]Likert, R. (1932). A Technique for the Measurement of Attitudes. Archives of Psychology, 140, 1–55.

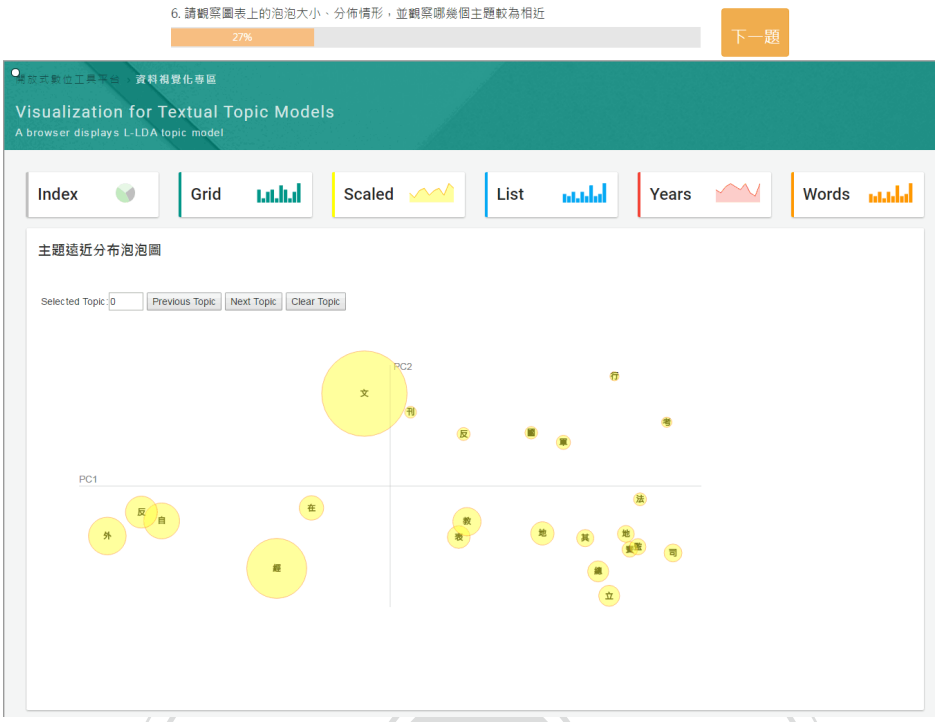


圖 5.3: 互動式前置介面引導任務網頁

5.3.3 第一階段易用性調查問卷

本研究第一階段問卷的目標旨在評估當受測者進行完前置介面引導任務之後，對於本視覺化介面的易用度是否符合使用者的需求。爲了評估此需求，本階段問卷採用系統可用性量表（System Usability Scale , SUS）[1]來量化本視覺化介面的易用程度。系統可用性量表（System Usability Scale , SUS）是由 John Brooke 在 1986 年創建至今，一直被廣泛應用在快速測試產品系統界面、桌面程式與網站界面上。此量表被認爲是可信賴且可有效評估系統可用性的工具 [10]，其主要用途爲檢測受測者在操作系統後的可用性主觀感受，也測量受測者在使用該系統上的效率和使用經驗。

如下表（5.2）所示，SUS 問卷包含了 10 個題目，作答方式同上份問卷一樣使用李克特五分量表（Likert Scale）。根據陳述語句的傾向給予各等級不同分數。在 SUS 量表中，個別題目的分數沒有意義，加總後的分數才是 SUS 分數。在問卷中，第 1、3、5、7、9 爲正向題，第 2、4、6、8、10 題負向題，所以在計算的時候，正向負向的計算方式會不一樣。

表 5.2: 系統可用性量表（SUS）問題列表

問題 1	我會願意經常使用這個系統。
問題 2	我認爲這個系統太過於複雜， 有多餘的設計。
問題 3	我認爲這個系統很容易使用。
問題 4	我想我需要有人幫助才能使用這個系統。
問題 5	我覺得這個系統的功能整合得很好。
問題 6	我覺得這個系統有太多不一致的地方。
問題 7	我可以想像大部分的人很快就可以學會使用這個系統。
問題 8	我覺得這個系統使用起來很麻煩。
問題 9	我很有自信能使用這個系統。
問題 10	我需要學會很多額外的資訊，才能使用這個系統。

正向題計算方式：

第 1、3、5、7、9 題，將各題的分數減 1，得到每題的分數，例如第一題原始分數是 3 分，減 1 後是 2 分。

負向題計算方式：

第 2、4、6、8、10 題，用 5 減去每一題的分數，例如第二題是 4 ，5 減 4 等於 1，這題分數就是 1 分。
最後將所有題目分數加總乘以 2.5 ，得到 SUS 總分（此分數非百分比）。

5.3.4 第二階段指定任務

經過上述的前置介面引導任務後，確認受測者對於本視覺化介面已經有一定程度的瞭解，就開始繼續下個階段的指定任務。此實驗階段的設計目的，是想探討當受測者在瞭解以及熟悉本視覺化介面各部份的功能及代表意義之後，是否能利用此平台挖掘出原本可能沒挖掘到的新研究議題，或是找到一些令人有興趣的「歷史殊異點」[3]，或甚至更進一步的找出此平台是否還有能拿來輔助其他領域的用途。

此階段的指定任務設計爲：在不限制操作時間的情況下，讓受測者自由探索本視覺化介面。當探索完畢之後，希望受測者能寫下他探索到且有興趣，或有疑問的內容，可能是某年份底下的關鍵詞；也可能是某年份的主題走向，或甚至根據平台所跑出來的文本量化內容，寫出一小段推測當時社會可能發生的事件，或

[3]金觀濤、邱偉雲、劉昭麟，〈「共現」 詞頻分析及其運用—以「 華人」 觀念起源爲例〉，《第三屆數 位典藏與數位人文研討會論文集》，2011 年12 月1-2 日，國立臺灣大學，頁214。

評論當時的社會風氣。

在此一部分的實驗中，並不設限受測者一定要寫出甚麼內容，因爲每位受測者的背景資訊都不同，也都有不同的觀點和各自使用此視覺化介面的方法，所以希望讓每位受測者自由發揮。但若是受測者眞的想不出來要寫甚麼，本研究也提供了一些指定任務（如下表 5.3 所示），當受測者沒有特別想法的時候，就可以根據指定任務來操作本介面，並寫下受測者在完成任務的過程中的任何想法。

表 5.3: 指定任務問題列表

問題 1	請受測者點擊表現自由/出版法問題之主題列表。
問題 2	由主題年代長條圖比例長條圖可否發現哪幾年特別突出？
問題 3	請使用者觀察 1958 年底下的主題年代關鍵字，觀察後寫下任何想法。
問題 4	請再點擊欲探討之關鍵字在 1958 年的相關文章，並寫下任何想法。

5.3.5 第二階段有用性調查問卷

當受測者做完上述指定任務後，就可開始填寫第二階段問卷。第二階段問卷的內容主要針對系統可用性（usefulness）來做設計，主要評估各個不同介面對受測者的有用程度爲何以及本視覺化介面整體綜合有用性的程度爲何，並包含兩題開放式問題以瞭解受測者對此視覺化介面的看法，例如任何的優缺點或特色、以及對系統的建議。此階段的問卷共有 16 題，第 1 題到第 8 題爲各細項評估，第 9 題到第 14 題爲整體介面的綜合評估，最後兩題則爲開放式問題。（如下表 5.4 所示），除了最後兩題的開放式問題，其他問題皆同樣由李克特五分量表（Likert Scale）來評測，爲了防止受測者便宜行事，全部填寫相同答案，問卷中的第（4）、（6）、（8）、（9）、（11）、（14）題爲負向題。實驗結果經過量表分數計算後會再加以計算平均值及標準差，希望能透過此階段的問卷答案得到更爲精準的受測者評估以及回饋。

表 5.4: 第二階段有用性調查問卷問題列表

問題 1	我認爲首頁（Index）的設計是對使用者有幫助，且實用的。
問題 2	我認爲主題總覽頁（Grid）的設計對使用者有幫助，且實用的。
問題 3	我認爲主題泡泡頁（Scale）的設計是對使用者有幫助，且實用的。
問題 4	我認爲主題列表頁（List）的設計是對使用者毫無幫助的。
問題 5	我認爲主題專題頁（Topic）的設計是對使用者有幫助，且實用的。
問題 6	我認爲文字索引頁（Word）的設計是沒有意義的。
問題 7	我認爲年代河流頁（Year）的設計是對使用者有幫助，且實用的。
問題 8	我認爲文本搜尋頁（開放式數位平台）是沒有意義的。
問題 9	我覺得將視覺化分析結合文本的探索是無趣的。
問題 10	我覺得此視覺化介面和傳統的文本閱讀方法相比， 更能讓使用者快速掌握文本。
問題 11	我認爲視覺化介面和開放式數位平台的結合是沒有意義的。
問題 12	我願意再回來使用此視覺化介面和開放式數位平台去查看其他文本。
問題 13	我認爲此視覺化介面裡面所運用到的各類視覺化圖表是過於複雜的。
問題 14	我認爲此視覺化介面的整體設計介面是一目了然的。
問題 15	此視覺化介面有什麼特別吸引您的功能或特色？
問題 16	對此視覺化介面的建議？請敘述說明。

5.4 實驗結果分析與討論

根據本研究 5-3 節提到的兩階段問卷，第一階段問卷旨在評估本視覺化介面的易用性，第二階段則旨在評估本視覺化介面的有用性，探討受測者在使用本視覺化介面時的行爲操作和使用心得是否有符合研究者所預期。本章節將依序先針對 SUS 系統可用性尺度量表作結果分析，再來是第二階段系統有用性的結果分析，最後是實驗前測問卷在實驗前和實驗後的結果分析比較。在結果分析的過程中，除了問卷量化後的數值結果，本章節也將使用受測者的回饋內容做出質化的描述分析。而詳細的開放式問題結果因篇幅問題，將會收錄在附錄 B 的部分。

5.4.1 系統可用性尺度量表（SUS）比較結果分析

本節將會對受測者在第一階段前置介面引導任務結束後，所填寫的系統可用性尺度量表（以下簡稱 SUS ）做出結果分析。SUS 主要用來把使用者感受的質化資料藉由此量表轉成可分析的量化資料，藉此方式來了解使用者對系統主觀的滿意度。此量表採用正反問題各半措辭的方式，目的就是要增加問卷的客觀性。表 5.5 爲第一階段受測者所填寫的 SUS 問卷結果。

表 5.5: 系統可用性尺度量表（SUS）受測者分數列表

受測者	SUS 分數
A	80.0
B	62.5
C	67.5
D	72.5
E	77.5
F	67.5
G	70.0
H	72.5
I	62.5
J	72.5
最大值	80.0
最小值	62.5
平均值	70.5
標準差	5.7494

國立政治大學 National Chengchi University

Bangor 等人 [2] 從接近 3,500 個調查中發現 SUS 被運用於網頁介面爲最多（41 %），其次爲行動電話介面（17 %）。而 SUS 於網頁介面平均得分爲 68.2 分。且爲了協助相關研究者使用與解釋系統可用性量表結果，Bangor 等人將分數分爲五個等級（A、B、C、D、F）（如圖 Figure. 5.4 ），90 分以上爲 A Score 、80 分以上爲 B Score ，70 分以上則通稱爲「可接受範圍」（C Score），而在等級 C 中又再細分爲 72 分以上爲好（Good）等等。

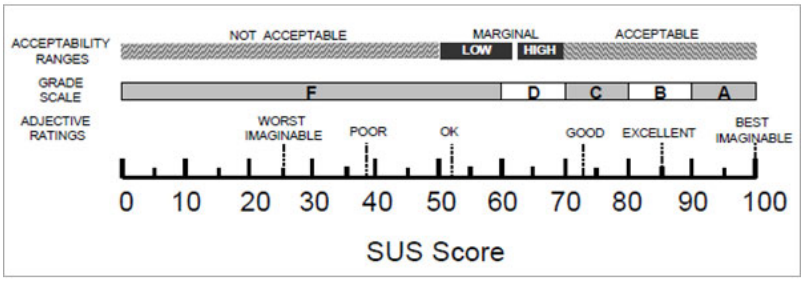


圖 5.4: System Usability Scale (SUS) 系統可用性尺度量表分數評估表

經過問卷分數整理再轉換成 SUS 量表分數之後，受測者所填寫的量表平均分數爲 70.5 分，剛好在可接受範圍（C Score）之內，但仍未達到等級 C 內之 72 分（Good）的水準，不過已超過了一般網頁介面的 SUS 平均得分 68.2 分，但仍應朝 72 分以上（Good）的水準去做努力跟改善。

再仔細查看每項題目的答題情形，會發現第一題（我會願意經常使用這個系統）是所有問題裡面平均值最低，而且標準差最高的一題，故未來若要提高本視覺化介面的 SUS 量表分數，第一題是最需要去改進的地方，因爲此題只有本身爲歷史背景的受測者給到了 4 分的高分，而其餘非歷史背景的使用者皆給出了 0 分至 1 分的評價，如何讓此介面能讓更多不同背景領域的使用者願意使用，將會是未來最大的改進方向之一。至於其他和系統操作有關的問題（例如問題 8、9）都得到了比較高分的平均分數，可以說在此視覺化介面中，受測者對於一些基本的操作行爲在使用上並沒有發現甚麼太大的問題。

	題目	樣本數	最大值	最小值	平均值	標準差
問題一	我會願意經常使用這個系統。	10	4	0	1.6	1.36
問題二	我認爲這個系統太過於複雜，有多餘的設計。	10	3	2	2.6	0.48
問題三	我認爲這個系統很容易使用。	10	3	2	2.8	0.40
問題四	我覺得我需要有人幫助才能使用這個系統。	10	4	2	3.2	0.74
問題五	我覺得這個系統的功能整合得很好。	10	3	2	2.8	0.40
問題六	我覺得這個系統有太多不一致的地方。	10	4	3	3.8	0.40
問題七	我可以想像大部分的人很快就可以學會使用這個系統。	10	3	1	2.4	0.80
問題八	我覺得這個系統使用起來很麻煩。	10	4	3	3.6	0.48
問題九	我覺得我很有自信能使用這個系統。	10	4	3	3.3	0.45
問題十	我覺得我需要學會很多額外的資訊，才能使用這個系統。	10	3	3	3	0

表 5.6: 系統可用性尺度量表（SUS）問題分數列表

5.4.2 第二階段有用性問卷內容結果分析

此階段的問卷共有 16 題，第 1 題到第 8 題爲各細項評估，第 9 題到第 14 題爲整體介面的綜合評估，最後兩題則爲開放式問題。這一階段主要評估本視覺化介面的有用性及各個功能是否能有效幫助到使用者輔助文本研究或更快速的瞭解大規模文本。在此階段同樣使用李克特五分量表（Likert Scale）來評測。實驗結果經過量表分數計算後會再加以計算平均值及標準差，希望能透過此階段的問卷答案得到更爲精準的受測者評估以及回饋。

下表包含了受測者進行指定任務時總花費的時間，以及各細項和綜合的視覺化介面有用性問卷分數。其中受測者（A）、（B）、（C）、（D）、（E）爲具有歷史背景之使用群族，而受測者（F）、（G）、（H）、（I）、（J）則爲非相關背景科系之群族。

國立政治大學
National Chengchi University

受測者編號	花費時間(分：秒)	首頁	主題總覽頁	主題泡泡頁	主題列表頁	主題河流頁	主題專題頁	文字索引頁	文本搜尋頁	綜合問題
A	80:49	3	3	5	3	3	5	5	5	4.40
B	28:55	2	2	3	3	1	4	2	5	3.30
C	36:42	3	2	4	3	3	5	3	5	4.20
D	45:30	2	2	3	3	3	4	2	5	3.10
E	58:21	2	2	3	3	3	5	3	5	3.60
F	39:10	3	2	4	3	2	5	2	4	4.10
G	26:57	3	2	3	4	2	5	2	3	3.90
H	23:48	3	3	4	3	3	5	3	3	3.80
I	47:12	2	2	3	3	3	4	2	5	3.60
J	68:23	3	2	3	4	3	4	4	4	4.00
平均值	45:34	2.60	2.20	3.50	3.20	2.60	4.60	2.80	4.40	3.80
標準差	–	0.49	0.40	0.67	0.40	0.66	0.49	0.98	0.80	0.38

表 5.7: 第二階段有用性問卷各項分數列表

從上表 5.7 可以看到，受測者對於覺得有用的視覺化介面整體來說還蠻有共識的，被認爲最有用的介面分別是平均 4.6 分的主題專題頁，和平均 4.4 分的文本搜尋頁。而若再依使用群族細分，會發現具有歷史背景之群族（受測者 A、B、C、D、E）在文本搜尋頁的有用性全數皆給了 5 分的高分。受測者 E 甚至在第一次看到文本搜尋頁時立刻表示:「要是我當時寫論文有這個就好了！這個網頁是甚麼時候有的？現在可以用了嗎？」其他人（受測者 A、C、D）則表示有時候就是突然想翻一下文本但手邊沒有文本，或是想搜尋某個關鍵字或某個概念。但想不起來在文本哪裡的時候，就會使用文本搜尋頁。

相較於上述熟悉文本的組別，非歷史背景之受測者（編號 F、G、H、I、J）對於文本搜尋頁的有用性則並沒有給到非常高的分數，除了受測者 I 表示，他本身有在寫小說，此頁面能讓他做校稿用，例如查「我」這個字，如果在同一頁出現太頻繁，就可以用別的字來代替，可以幫助我避免一直用一樣的字來代表主角。 其他受測者（編號 F、G、H、J）則對主題專題頁的有用性給了較高的評價。受測者 G 表示:「主題專題頁是我認爲最有用的地方，可以看到那年底下的關鍵字，覺得配合歷史文本還蠻有用的，文本搜尋頁相較之下則讓我覺得不太實用。」

從表中也可觀察到標準差最大的項目爲「文字索引頁」，僅有受測者 A 對於此頁面給了極高的滿分評價，受測者 A 表示: 我在文字索引頁（Word）的頁面看到曹丕，直覺認爲在影射蔣中正，點進去文本搜尋頁查看原文之後果然驗證了我的看法，我覺得這是在當時是一篇非常大膽的文章，不知道作者有沒有被抓去關。另外，我在（Word）的頁面裡看到「法寶」，點進去看發現有這兩個字的文章都跟國際政治有關，但這似乎牽扯到語義學的概念了，可能當時國際政治的評論很愛用法寶二字。整體來說我覺得這頁對我還蠻實用的。受測者 J 則表示在這一頁裡面有好多人名，不知可否統一出來成一個人名的區域，因爲我對這些沒聽過的人名比較有興趣，會想要去查查看這些人的資料。

從上面的受測者回饋，可以發現本視覺化介面對不同領域的使用者而言，有用的地方皆不盡相同，而對於非歷史背景的使用者而言，未來此視覺化介面或許也能轉成小說文本校稿之類的其他用途，這是一個很好的回饋方向。但對於平均分數較低的首頁及主題河流頁，在幫助使用者的有用性上還有再加強的空間，無論是設計介面，還是所表達的內容（例如受測者 B 表示:「河流圖的顏色難以閱讀，建議在河流上面增加關鍵字項目」），都還有再加強的空間，其改善方向應該要朝讓整體介面功能更加全面的符合各種不同背景的受測者，以提升受測者的使用愉悅度和有用性感受。

5.4.3 實驗前測問卷前後結果分析

在實驗開始前，本研究會先請受測者填寫一份實驗前測問卷。該份問卷包含了四題用來紀錄使用者基本資訊的問題，以及兩題紀錄使用者在實驗之前，對利用文本資料視覺化輔助文本研究、以及對於用主題模型輔助文本研究的看法偏向同意與否。

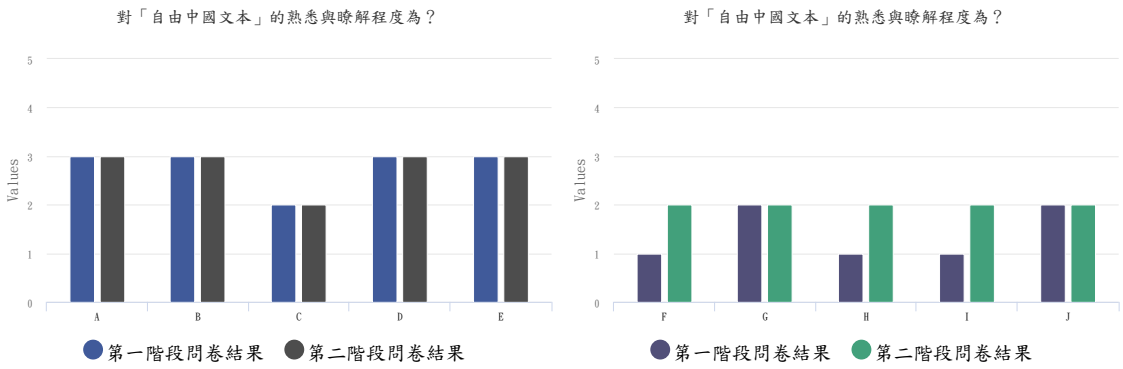
而當實驗結束後，會請受測者再度填寫一次此實驗前測問卷，目標是希望瞭解受測者在實驗前後，對於文本內容瞭解的提升程度爲何，以及對於文本資料視覺化和量化文本分析的態度，是否會有實驗前後上的轉變。

根據問卷的題目，以下將依序討論各個問題的問卷結果，並依照受測者背景的不同，分群（群組 1）、（群組 2）畫成圖表後進行互相比較與分析。

表 5.8: 實驗前測問卷結果分析－問題一

對「自由中國文本」的熟悉與瞭解程度爲？	群組 1 平均值	群組 1 標準差	群組 2 平均值	群組 2 標準差
第一階段結果	4.6	0.8	1.4	0.49
第二階段結果	4.6	0.8	2	0

※ 群組 1（受測者 A、B、C、D、E）爲具歷史背景領域之受測者，群組 2（受測者 F、G、H、I、J）爲非相關背景領域之受測者。



(a) 群組 1 前後受測結果　　(b) 群組 2 前後受測結果

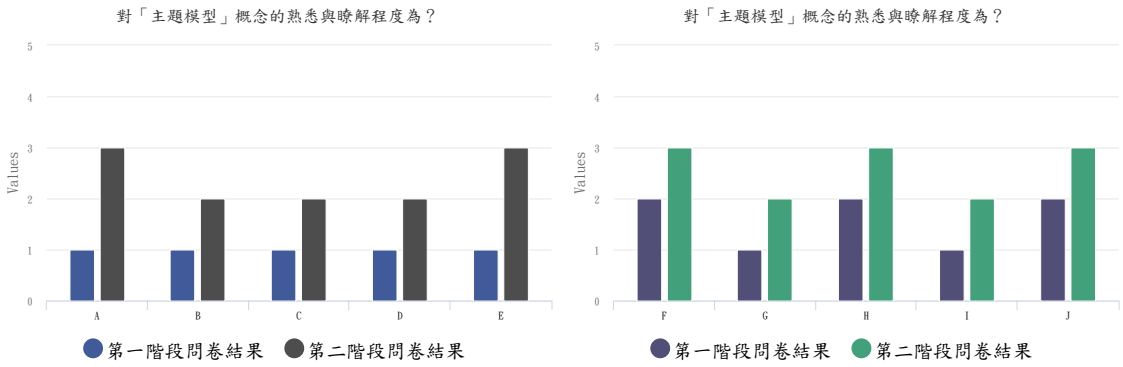
圖 5.5: 受測者前測問卷前後結果分析－問題一

從上圖可以看出，原本就熟悉「自由中國」文本的組別，在實驗結束之後，對自由中國的文本內容並沒有提升其瞭解程度，受測者 B 表示，因爲問卷的級距太小了，如果是十等份了話就是從瞭解程度 2 提升到瞭解程度 3 左右。而對於原本不熟悉自由中國文本，非歷史背景領域的受測組別而言，在做完實驗之後，有三位受測者都表示比較熟悉自由中國文本在說些甚麼了，原本可能只是聽過「自由中國」這本雜誌的名字而已，實驗結束之後大概瞭解裡面包含了哪些核心概念和歷史思想。

表 5.9: 實驗前測問卷結果分析－問題二

對「主題模型」的熟悉與瞭解程度爲？	群組 1 平均值	群組 1 標準差	群組 2 平均值	群組 2 標準差
第一階段結果	1	0	1.6	0.49
第二階段結果	2.4	0.49	2.6	0.49

※ 群組 1（受測者 A、B、C、D、E）爲具歷史背景領域之受測者，群組 2（受測者 F、G、H、I、J）爲非相關背景領域之受測者。



(a) 群組 1 前後受測結果　　(b) 群組 2 前後受測結果

圖 5.6: 受測者前測問卷前後結果分析－問題二

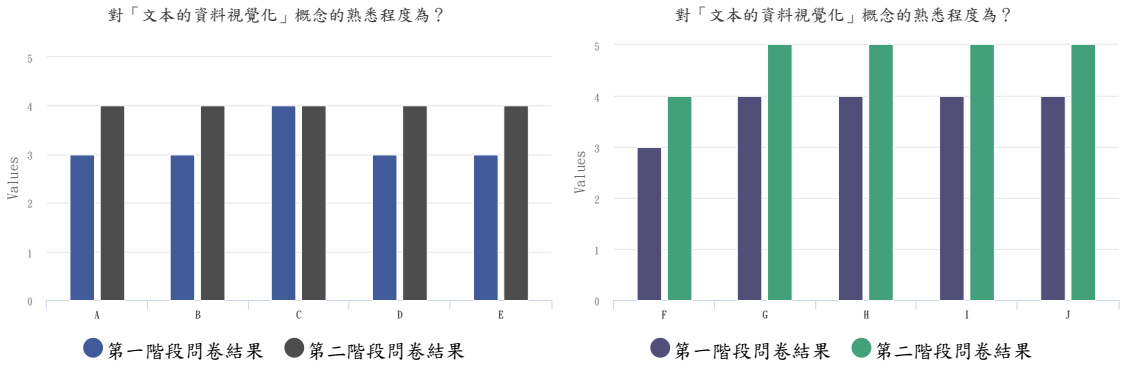
從問卷結果來看，無論是群組 1 還是群組 2，全數使用者均對主題模型的概念有所提升，可能是從完全不熟悉進步到稍微熟悉，或是從稍爲熟悉進步到普通熟悉。受測者 J 表示:「這種演算法好像用在英文文本效果會比較好耶，無論是斷詞出來的效果還是主題的聚集等等。像是如果套用在作者王文興的剪翼史上[4]，這種完全打破中文語意的文本規格，主題聚集的效果可能會很差」。從上述的言論可以看出受測者雖然沒有相關資訊背景，但對主題模型或是文本斷詞等文本分析方法的確已有了初步的認知。

[4] 王文興.《剪翼史》. 台北：洪範, 2016.

表 5.10: 實驗前測問卷結果分析－問題三

對「文本的資料視覺化」熟悉與瞭解程度爲？	群組 1 平均值	群組 1 標準差	群組 2 平均值	群組 2 標準差
第一階段結果	3.2	0.4	3.8	0.4
第二階段結果	4	0	4.8	0.4

※ 群組 1（受測者 A、B、C、D、E）爲具歷史背景領域之受測者，群組 2（受測者 F、G、H、I、J）爲非相關背景領域之受測者。



(a) 群組 1 前後受測結果　　(b) 群組 2 前後受測結果

圖 5.7: 受測者前測問卷前後結果分析－問題三

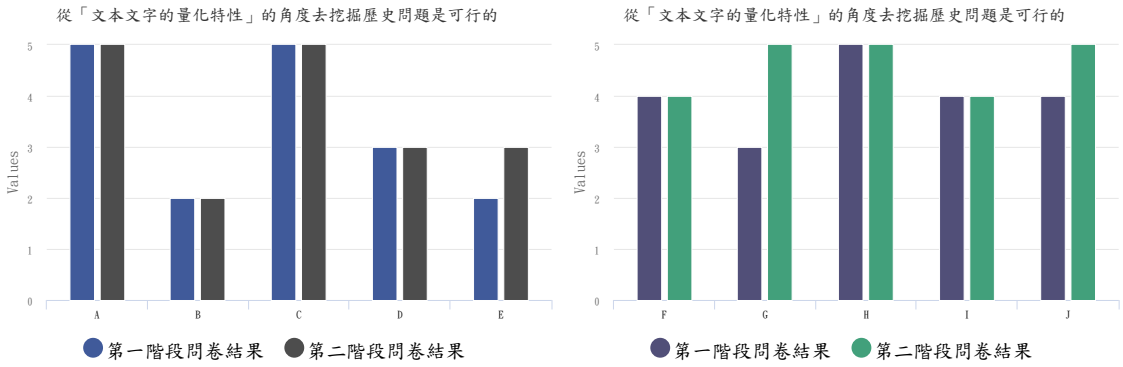
從問卷結果來看，大多數的使用者在實驗前就對文本資料視覺化有一些想像中的概念，可能是以圖表或是互動式的方式來呈現文本，例如受測者 B 表示，之前有在網路上看過 2016 總統就職演講稿的資料視覺化分析圖[5]。而在實驗結束之後，兩邊不同群組的人對於文本資料視覺化這個概念幾乎都能上升一分，多數使用者均表示看到了很多種不同呈現手法的文本資料視覺化，覺得又更加瞭解了此概念。

[5]http://www.chinatimes.com/realtimenews/20160520004170-260407

表 5.11: 實驗前測問卷結果分析－問題四

從「文本文字的量化特性」的角度去挖掘歷史問題是可行的。	群組 1 平均值	群組 1 標準差	群組 2 平均值	群組 2 標準差
第一階段結果	3.4	1.35	4	0.63
第二階段結果	3.6	1.2	4.6	0.49

※ 群組 1（受測者 A、B、C、D、E）爲具歷史背景領域之受測者，群組 2（受測者 F、G、H、I、J）爲非相關背景領域之受測者。



(a) 群組 1 前後受測結果　　(b) 群組 2 前後受測結果

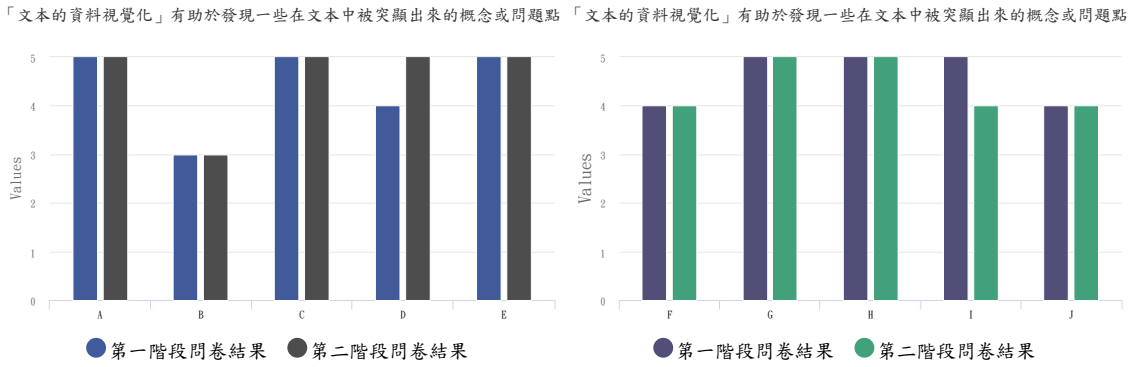
圖 5.8: 受測者前測問卷前後結果分析－問題四

這題是在所有問卷題目裡面分數差異相對其他問題差異最大的一組，在群組 1 中，有些人是同意這個觀點的，引述某位贊同此觀點的 A 同學的話:「簡單說，這種分析法的優點在於能提供很新的問題意識，實際要分析（文本）當然還是要研究者親自下去看去研究，但這個在我們碩一史學方法有教，另外，當同一個現象在文本頻繁出現，就有它的歷史意義」。而不贊同的其他三位則是不約而同表示用量化方式分析文本的做法有去脈絡化的嫌疑，並表示歷史文本不應該這樣研究，還是要深入到文本本身，並且人、事、時、地、物這五種最重要的歷史討論因素都要一併加進來考慮比較好。 而對於本身不熟悉自由中國文本的群組 2 來說，就完全沒有人提到和去脈絡化有關的字眼或疑問，同學 J 則表示這種做法（以量化方式研究文本的做法）如果早點發明出來，說不定當時的政府會從文本找到更多資訊，例如一些相關的敏感的人或地點，把更多人抓起來也說不定。整體而言在實驗後，群組 2 的受測者對於此量化方式分析文本的做法，抱持著持平態度或者約略上升其認同度。

表 5.12: 實驗前測問卷結果分析－問題五

文本視覺化有助於發現一些在文本中被突顯出來的概念或問題點。	群組 1 平均值	群組 1 標準差	群組 2 平均值	群組 2 標準差
第一階段結果	4.4	0.8	4.6	0.49
第二階段結果	4.6	0.8	4.4	0.49

※ 群組 1（受測者 A、B、C、D、E）爲具歷史背景領域之受測者，群組 2（受測者 F、G、H、I、J）爲非相關背景領域之受測者。



(a) 群組 1 前後受測結果　　(b) 群組 2 前後受測結果

圖 5.9: 受測者前測問卷前後結果分析－問題五

從問卷結果來看，群組 1 的使用者多數對於此問題是抱持著肯定的態度，除了受測 B 表示：「我覺得視覺化的結果並不見得可以凸顯到一些概念或問題點，被突顯出的或許是最多的，但不一定是最重要的概念或問題點」。而群組 2 的受測者在實驗前後也幾乎對於此問題是抱持的持平的態度，除了受測者 I 表示：「在實驗前我原本選擇 5（非常同意），但做完實驗之後才發現因爲我對文本不夠熟悉，單純用圖表去凸顯，或單純用詞頻去看的時候並不能去發現或瞭解每個圖表裡面想表達的重要意義，還是要深入文本研究才行，否則他對我來說只是一張圖而已，很難去挖掘其中的意義，因爲我不知道重點在甚麼地方」。而受測者 I 也是唯一在實驗完畢後對此題分數給予下降評分的受測者。

5.4.4 開放式問題訪談結果分析

本段落將針對實驗問卷的最後兩題開放式問題進行訪談後結果分析，其問題分別如下表（5.13）所示：

表 5.13: 開放式問題列表

問題 1	此視覺化平台有什麼特別吸引您的功能或特色？
問題 2	對此視覺化平台的建議？請敘述說明。

受測者在這兩題開放式問題當中都提供了許多非常寶貴的建議，以及未來視覺化平台可以改進的方向，在此將簡單歸納爲幾點來探討，詳細的開放式問卷答案內容請參閱附錄 B 。

在第一題中，受測者無論是群組 1 或是群組 2 ，全數都提到了整體網頁的介面設計、圖表的呈現還蠻漂亮的，在外觀上算是獲得了全數受測者的肯定，而對於群組 1（本身爲歷史背景的人）而言，有三位受測者皆提到對他們而言，最吸引他們的地方爲主題專題頁的主題年代機率分佈和各年代底下關鍵字，因爲可以和自己的所學相互驗證，是本視覺化平台最吸引她們的地方，例如受測者 A 表示：「我覺得主題年代分佈長條圖和底下立即呈現的關鍵字，這兩個功能是最吸引我的地方，因爲他能讓我的研究有所本，能證明一個事件（例如：出版法／表現自由）在每個年代的關鍵字和趨勢的走向，我覺得這就是人做不到的事情，也是在妳（此處指本作者）的研究裡最有價值的地方」。受測者 E 則表示看到圖表可以和所學做連結的感覺很好，可以讓我去回想當時發生了甚麼事，和所學做驗證。

而對於群組 2（非歷史背景的人）而言，平台中最吸引人的功能就比較分歧，每個人選擇的地方都不盡相同，例如受測者 F 表示：「我覺得泡泡圖和年代河流圖的呈現方式很一目瞭然，圓圈越大代表資料越多筆，河道以編年方式可以線性知道各時間區段的主題的變化幅度情況，這兩頁是我最喜歡的地方」。而受測者 H 則表示：「我覺得主題專題頁是最好用的一頁，因爲他有年代、有關鍵字，我覺得很好拿來研究一些議題，最討厭的則是年代河流圖，因爲我看不到每個主題底下的關鍵字，不太能懂每個主題的意涵」。受測者 I 則表示：「我最喜歡的功能是文本搜尋頁，感覺很方便，可以拿來校稿，如果能讓我上傳我自己寫的文本就太好了」。

從上述訪談結果可看出，本視覺化介面，對於文本熟悉程度不一的使用者而言，最吸引他們的地方都不太相同，對於本身已經熟悉自由中國的使用者而言，看到圖表和他們的所學有符合和一致的地方時，都會有相當程度的興奮感，對於電腦跑出來的結果能驗證他們的歷史知識，是本視覺化介面中最吸引他們的地方。而對於非歷史背景的使用者而言，雖然本介面最吸引他們的特點都不盡相同，但從訪談結果能得知本視覺化介面皆可以讓使用者自行挖掘出有興趣的地方

加以探索，以及針對個別覺得有用的地方去發掘出更多未來能應用的不同層面。

而關於第二題（對此視覺化平台的建議），兩邊的不同組別也有各別分歧和相同的地方，對於組別 1 的受測者而言，收到的建議中不約而同的對於圖表的呈現方式做出了大致相同的評論，受測者 A、B、C、D 皆表示，在本視覺化介面中，圓餅圖和長條圖是在所有的文本資料視覺化呈現中，最能清楚表達圖表內容的圖表。其他的圖雖然好看，也有互動性，但是過於複雜，反而不知道該怎麼看。例如受測者 A 表示：「希望可以把總覽圖改掉（甚至整個拿掉也沒關係），我覺得首頁圓餅圖的呈現主題方式對我而言比較清楚」。受測者 B 則表示：「我覺得其實圓餅圖就很清楚了，不過其他圖表的互動性比較好，應該很適合對高中生或是對歷史文本不熟的人來推廣，讓他們比較有興趣。如果要呈現主題模型，我認爲主題列表頁（List）長條圖會是最好的呈現，最清楚。整體上來說我覺得主題專題頁（Topic）這頁的概念最好，但整頁實在是呈現太多資訊了，內容、圖表也太多，使用者用起來會覺得很茫然」。

而對於組別 2 而言，圖表的複雜性或互動性相反的就沒有上述的意見評論，比較多針對的建議點是在於顏色和介面的部分，特別是主題年代河流圖的顏色，有三位受測者皆表示這個顏色有點讓他們無法接受，受測者 I 表示：「每次看到這張圖就覺得很像紫色高麗菜」，而受測者 J 則表示：「有些河流的寬度實在是太細了，在操作上並不是非常容易」。

而無論是組別 1 還是組別 2，都有兩至三位的受測者表示，文字索引頁（Word）可不可以不要把全部的字放在一起，覺得這樣難以閱讀，受測者 E 表示：「希望文字索引頁可以按照主題分群，或是依文本出現年代分群等等」，而受測者 G 則表示：「個人覺得文字索引頁（Word）的搜尋如果可以反過來從字詞找主題，也能與前三個以主題找字詞的方式搭配起來，整體的交互使用會比較完整」。受測者 J 則表示：「希望文字索引頁（Word）能有一欄是專門放人名的，因爲我只對人名比較有興趣，希望可以以類別例如人名、地名等等，來分類所有字詞」。

我認爲上述的評論對於正在做資料視覺化研究的研究者而言，其實是很寶貴的建議，因爲有時候設計者會不小心的陷入想要做一個很酷、很炫的視覺化呈現的盲點，但卻忽略了對於一般的大眾非資訊背景使用者而言，有過多互動性的圖表並不是他們所需要的，他們只想清楚的立即瞭解資料的內容，最喜歡圓餅圖、長條圖這種立即、直接的呈現方式。 而針對顏色上面的建議，在設計過這次的視覺化介面之後，本人的確發現紅色系的漸層顏色是會對大眾容易具有壓迫感的一個色系，即使調整過了好幾十次，還是容易帶給人不舒服的感受，這在往後的設計上面是一個蠻值得注意的地方。 另外關於文字索引頁（Word）的部分，在實驗的過程中可以發現的確只有少數受測者，可以有耐心仔細的一個一個去查看本頁面裡面的文字，並找到有興趣的關鍵字加以查詢（只有兩位受測者在這頁停留超過了十分鐘），如何提高此頁面對受測者的有用性和易用性感受也是本視覺化介面的一個未來主要的改進方向。

第六章

結論與未來展望

6.1 研究結論

本研究試圖利用圖像資料視覺化在資訊表達上的優勢和效率，設計出合適的文本資料視覺化介面，直觀的呈現文本中的有效資訊。在本視覺化介面裡，使用者可以透過各種不同方式的視覺化圖案呈現來從不同的角度探索同一篇文本，包含主題總覽圖、主題遠近分佈泡泡圖、主題年代河流圖、以及主題年代比例分布圖等等。再結合開放式數位平台的文本檢索功能，有效幫助使用者回溯當時的文本情境，讓使用者能對文本中有興趣的地方，做更持續深入的挖掘與探索。 本研究首先設計了任務導向性的實驗讓每位受測者都能對視覺化介面的熟悉程度盡量趨近相同，再利用指定任務的方式讓受測者自由使用本介面並寫下所觀測到有興趣或有疑問的地方，實際的操作本視覺化介面。最後再藉由三種評估問卷（SUS 量表）、（系統可用性問卷）、（兩階段實驗前測問卷）來驗證本介面的易用性（Ease-of-use）和有用性（Usefulness），以及瞭解實驗前後受測者對於此種文本量化探勘方法以及視覺化呈現的態度肯定與否。

透過第五章的實驗結果，可以發現受測者大多數都寫下了讓他們覺得感興趣或挖掘出有疑問的地方並進行探索與文本搜尋，或是也利用本介面順利完成了指定問題的作答。而從實驗結果與受測者的回饋，以及 SUS 量表的總體分數（70.5分）上面來看，也驗證了本介面雖然還未及完美，但使用者在整體操作上並沒有太大的問題，對於整體介面的設計也有相當高的滿意程度。另外，從開放性問題的訪談結果來看，本介面在實務上的確能夠輔助相關領域的使用者來進行文本研究與開創嶄新的問題思維，另一方面，也能讓對文本熟悉程度不一的使用者在利用此介面探索文本的過程中，更快速地掌握大規模文本的事件全貌。

6.2 未來發展與改進

在本研究中所設計的視覺化介面，提供了以視覺化的分析結合文本搜尋系統，透過資訊科技用現有的數位典藏資源進行人文研究上的輔佐與文本資料視覺化的呈現。在前面的章節雖已證實本介面之有用性與可用性，但仍有需多未臻完

善之處，在此本研究提出以下幾點，希望未來可繼續朝以下幾個面向邁進：

1. 在視覺化介面與文本搜尋平台的結合上，希望能有更有效的且對使用者更友善的呈現方式，例如在文本搜尋頁上當使用者按下搜尋按鈕時，搜尋頁就能立即提供和文本相關的後設資料的視覺化，像是所查詢詞彙的年代分布次數一覽表，或是所查詢詞彙的相關作者比例圓餅圖等等，將視覺化呈現和文本之間做更緊密的動態結合，不用再開另一個新網頁去察看視覺化結果的圖，對使用者體驗必定能有更上一層的幫助，在使用上也會變得更加直觀。

2. 在應用層面上，目前的視覺化介面是限定爲使用者已自定好主題標簽的文本，以方便跑演算法並呈現主題模型的視覺化。但這樣的文本其實對使用者是有相當程度的實現門檻的，若是能夠轉變爲不同的使用目的，例如受測者之一所表示的想拿來當作自己的小說校稿平台，以視覺化的方式呈現有過多同樣詞彙出現的章節，以及讓使用者隨意上傳自己的文本而不用再另外自訂標簽，應該能大大提高使用者使用本視覺化介面與開放式數位平台的意願，也能符合更多使用者的不同需求。

3. 在視覺化介面的圖形整合性上，還有許多再加強的空間，有超過三位以上的受測者表示，主題總覽圖和主題遠近分布圖，其實在本質上都是屬於泡泡圖的一種，把兩者圖案合而爲一，直接把主題的關鍵字放進主題遠近分佈圖裡整併在一起，或許能讓整體的視覺化介面看起來更加清爽，有鑑於此，在整體視覺化圖案的設計考量上，的確是還有再進一步精鍊的空間。

4. 最後，是在詞彙－主題－文本，這三種由點至線而面的文本模型層面上，彼此之間還有許多在交互設計上的改進空間，例如是否能在主題專題頁上面，當使用者按下年代長條圖時，底下陳列的不只是該年度關鍵字，還有該年度關鍵字所出現的文章，讓使用者一目瞭然該年度文章裡所發生的事件，再進一步點進去文本檢索頁上做全文搜尋或其他後設查詢。或者是在文字索引頁上增加點進去後也可以看見同主題的其他詞彙的功能等等，和前三個以主題找詞彙的視覺化介面互相搭配，豐富使用者整體的使用感受與體驗，並更完整的呈現詞彙－主題－文本，三者彼此之間的交互關係。

參考文獻

[1] W. Albert and T. Tullis. *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. Newnes, 2013.

[2] A. Bangor, P. Kortum, and J. Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3):114–123, 2009.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[4] M. Bostock. D3. js. *Data Driven Documents*, 492, 2012.

[5] A. J.-B. Chaney and D. M. Blei. Visualizing topic models. 2012.

[6] J. Chuang, C. D. Manning, and J. Heer. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 74–77. ACM, 2012.

[7] D. Cochran. *Twitter Bootstrap Web Development How-To*. Packt Publishing Ltd, 2012.

[8] S. T. Dumais. Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230, 2004.

[9] S. Eppinger and K. Ulrich. *Product design and development*. McGraw-Hill Higher Education, 2015.

[10] K. Finstad. The system usability scale and non-native english speakers. *Journal of usability studies*, 1(4):185–188, 2006.

[11] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[12] P. Legris, J. Ingham, and P. Collerette. Why do people use information technology? a critical review of the technology acceptance model. *Information & management*, 40(3):191–204, 2003.

國立政治大學 National Chengchi University

[13] R. Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[14] J. Murdock and C. Allen. Visualization techniques for topic model checking. In *AAAI*, pages 4284–4285, 2015.

[15] J. Nielsen and T. K. Landauer. A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 206–213. ACM, 1993.

[16] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics, 2009.

[17] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press, 2004.

[18] I. J. Schoenberg. *Cardinal spline interpolation*. SIAM, 1973.

[19] B. H. Sheppard, J. Hartwick, and P. R. Warshaw. The theory of reasoned action: A meta-analysis of past research with recommendations for modifications and future research. *Journal of consumer research*, 15(3):325–343, 1988.

[20] C. Sievert and K. E. Shirley. Ldavis: A method for visualizing and interpreting topics. 2014.

[21] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. In *NIPS*, pages 1385–1392, 2004.

[22] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 2012.

[23] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[24] N. Yau. *Data points: Visualization that means something*. John Wiley & Sons, 2013.

[25] 賈西平, 彭宏, 鄭啓倫, 石時需, and 江焯林. 基于主題的文檔檢索模型. 華南理工大學學報 *(自然科學版)*, 36(9):37–42, 2008.

國立政治大學 National Chengchi University

附錄一 介面引導任務問題列表

國立政治大學
National Chengchi University

問題 1	觀察首頁右側的圓餅圖，找到「文藝類」且點擊後， 再次點選「文藝類」並點擊「經濟/財政」
問題 2	請找出「總統三連任問題／修憲／國大」的主題泡泡， 找到後點擊泡泡
問題 3	觀察在此主題（總統三連任問題／修憲／國大）泡泡底下出現的關鍵字， 並將滑鼠遊標指向各個關鍵字（勿點擊）
問題 4	觀察完畢後可點擊有興趣的關鍵字，將連接到文本搜尋系統頁面查詢
問題 5	請觀察圖表上的泡泡大小、分佈情形，並觀察哪幾個主題較爲相近
問題 6	請找出「外交／聯合國問題」之主題泡泡，並以滑鼠點擊查看， 查看完畢再繼續回來問卷頁面
問題 7	請在表格右上方搜尋欄上輸入「民主」，並觀察此關鍵字 出現在哪幾個主題裡面。查詢完畢後刪除關鍵字
問題 8	請找到表格最左邊上方的〝Show 10 entris〞（每頁顯示10列）， 點擊下拉式選單更改成每頁顯示〝25〞列
問題 9	請找到「Topic3 法治」橫列後滑鼠點擊
問題 10	請觀察左側長條圖，分別爲主題底下關鍵字的比例
問題 11	請觀察右側上方柱狀圖，分別爲自由中國文本裡 各年代「Topic3 法治」所占的比例。
問題 12	請分別點擊〝1953〞、〝1954〞、〝1955〞年之柱狀圖， 並仔細觀察下方隨年代改變之主題關鍵字
問題 13	請點擊〝1955〞年之柱狀圖後，點擊底下第一個比例最高的關鍵字 「中美共同防禦條約」
問題 14	請觀察剛剛點擊開啓的新視窗，此爲在文本搜尋頁查詢 「中美共同防禦條約」後之文本搜尋結果
問題 15	請點擊「檢視全文」以便觀察更多文本之本文內容。 觀察完畢後請回來主問卷頁面
問題 16	請將滑鼠移至各個主題河流圖分支（勿點擊），觀察每個主題 隨年代而改變的文本占比及比例多寡趨勢
問題 17	請點擊有興趣之主題河流進入前述點擊過之主題專題頁面
問題 18	請將滑鼠移至各個主題底下之關鍵字，查看各關鍵字 同時出現在哪些主題裡面
問題 19	請點擊關鍵字列表中，第五行第一項之「科學論文」， 查看前述已點擊過之自由中國文本搜尋系統頁面

國立政治大學 National Chengchi University

附錄二 受測者開放性問題回饋

受測者	此視覺化平台有什麼特別吸引您的功能或特色？	對此視覺化平台的建議？請敘述說明。
A	我覺得泡泡主題遠近分佈圖、主題年代分佈長條圖和底下立即呈現的關鍵字，這兩個功能是最吸引我的地方，因爲他能讓我的研究有所本，能證明一個事件（ex:出版法/表現自由）在每個年代的關鍵字和趨勢的走向，我覺得這就是人做不到的事情，也是我覺得妳研究裡最有價值的地方。	希望可以把總覽圖改掉（甚至整個拿掉），覺得首頁的圓餅圖呈現主題比較清楚。
B	我覺得其實圓餅圖就很清楚了，不過其他圖表的互動性比較好，應該很適合對高中生或是對歷史文本不熟的人來推廣。如果要呈現主題模型，我認爲（List）會是最好的呈現，最清楚。我覺得（Topic）這頁的概念很好，但整頁實在是呈現太多資訊了，內容也太多，使用者會覺得很茫然。	1. 我認爲文本用關鍵字分析會有盲點，例如我用「憲政」這個字想查自由中國文本，但我很容易就會忽略掉其他有「三連任」、「修憲」、「違憲」等等也有憲政概念但沒有憲政關鍵字的文章。2. 河流圖的顏色不是很好看。
C	我覺得總統三連任的河流圖還蠻有趣的，主題專題頁可以看到事件的趨勢，還不錯，像是地方選舉這個主題在 1959 年和 1960 年比較高，可能是和當時 1960 年的台南市長選舉有關，然候在 1959 年的時候就開始在討論的關係。	搜尋頁可以一次搜多一點嗎？這份文本（1949 - 1960）其實比較舊，不是每個研究歷史的人都熟悉，不知道能不能換成比較新一點的文本。

D	我覺得文字索引頁（Word）有些字還蠻有趣的，像是「汽車底盤」！？居然有討論汽車底盤的文章。文本搜尋頁的設計非常棒，如果我手邊沒有文本就可以直接上去查看，如果我當時寫碩論的時候有這個平台就好了。我覺得圖表可以和所學做連結是最吸引我的地方。	我覺得長條圖比河流圖好懂，再來第二好懂得是圓餅圖，其它都有點太複雜。我覺得關鍵字裡廣泛的字太多了，可以篩選挑掉嗎？我覺得年代河流圖（Year）太細了。希望文字索引頁（Word）可以按照主題分類排序文字。
E	特色是文本搜尋頁很棒！這樣不一定要看完一整本書，搜尋特定地方即可。然後整體介面很美。	希望在文字索引頁（Word）增加搜尋欄，希望文本搜尋頁裡面的「全文檢索」也能做關鍵字 Highlight 。有時候不確定那些可以點擊，希望能做點提示的效果。
F	1. 整體介面還蠻漂亮的。2. 以泡泡和河道的方式很一目瞭然，圓圈越大代表資料越多筆，河道以編年方式可以線性知道各時間區段的幅度情況。	個人覺得文字索引頁（Word）的搜尋反過來從字詞找主題，也能與前三個以主題找字詞的方式搭配起來，交互使用比較完整。建議各字詞的超連結點選可以變成手勢圖標，會讓人比較直覺點下去。
G	利用主題找字詞，視覺化就能讓人易上手並且直覺操作，個人認爲那樣的呈現方式有助於了解各主題的整體資料，很不錯 。網頁整體看起來還蠻美觀的。	有些字詞，比如「沒有」、「我們」、「這個」等等，偏屬中性詞彙，以我來講，查詢上需求較不大，對以「物品、人物或者事件」等名詞或專有名詞或許比較需要這樣的資料。再篩選掉這類中性詞彙或許也能使視覺更清爽。（Scaled）和（Grid）這兩者，都以泡泡圖呈現，因爲感覺上比較相像，不知是否能整併在一起？整合前者的四個象限屬性和後者的列出每個主題顯著字詞以及字詞和其他主題的關聯。
H	整體介面還蠻漂亮的。主題遠近泡泡圖還蠻有趣的。	我覺得每個主題底下的「所以」、「沒有」這種比較空泛的關鍵字太多了。希望主題專題頁（Topic）頁的版面內容可以不用滾動滑鼠就能一次完整呈現。我覺得圓餅圖還是比較清楚。
I	整個網頁看起來畫面做的還不錯。主題遠近泡泡圖還蠻有趣的，爲什麼「考試」和「外交」離最遠？	可以上傳其它文本嗎？我想拿來上傳自己的小說，拿來檢查重複的字（例如：我）會不會在同個章節裡出現太多次。關鍵字有點太廣泛了。河流圖看起來好像紫色高麗菜，不太容易閱讀。

國立政治大學 National Chengchi University

J	我覺得文字索引頁（Word）裡面有好多人名好有趣，會讓我想要去搜尋這些沒聽過的人名。我覺得主題專題頁（Topic）這頁因爲有年代，所以還蠻好用的。	希望主題河流頁能增加主題關鍵字在上面，不然對主題的概念其實有點模糊。希望文字索引頁（Word）能按造人名、地名等等專有名詞來分類，因爲我只對人名比較有興趣而已。

國立政治大學
National Chengchi University