

**目錄**

[壹、 探討主題 2](#_Toc167739918)

[貳、 工作流程分析 3](#_Toc167739919)

[**一、 資料來源 : 3**](#_Toc167739920)

[**二、 Work2Vec 4**](#_Toc167739921)

[**三、 Bertopic主題模型 9**](#_Toc167739922)

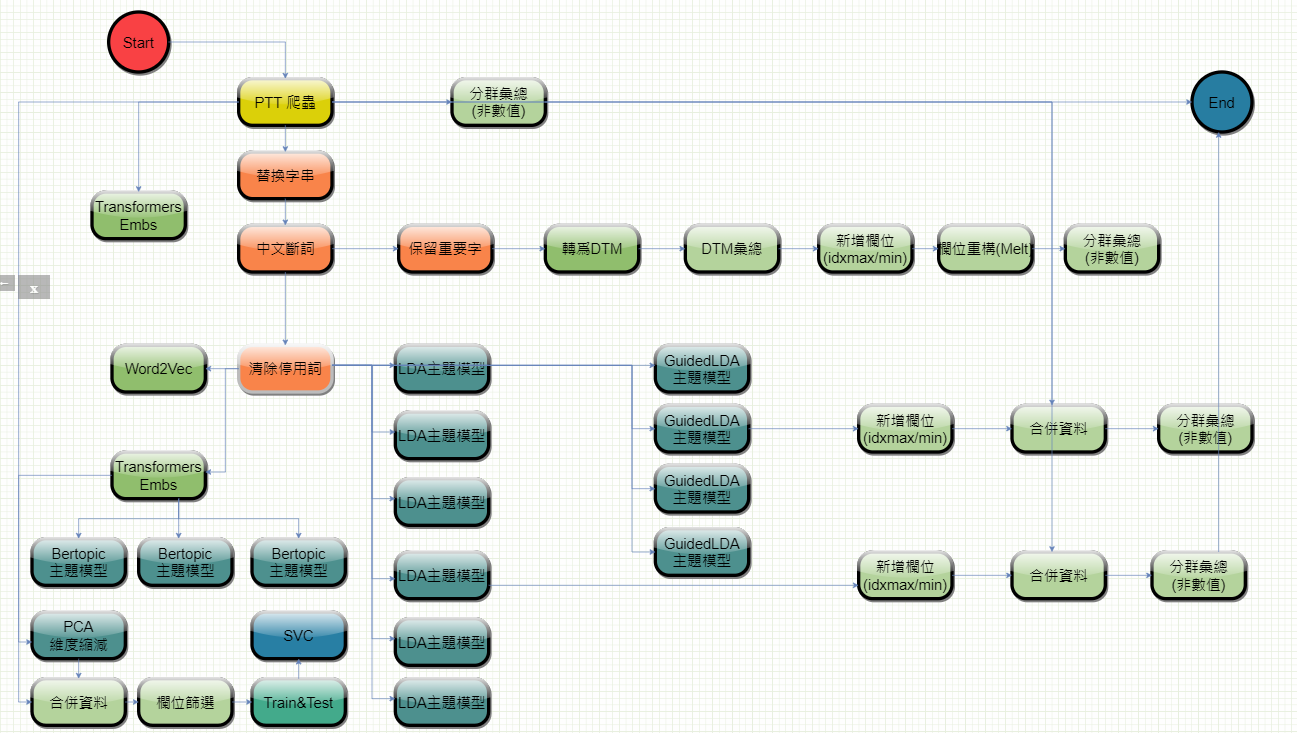
[**四、 Rag 16**](#_Toc167739923)

[**五、 社會網路 18**](#_Toc167739924)

[參、 結語 19](#_Toc167739925)

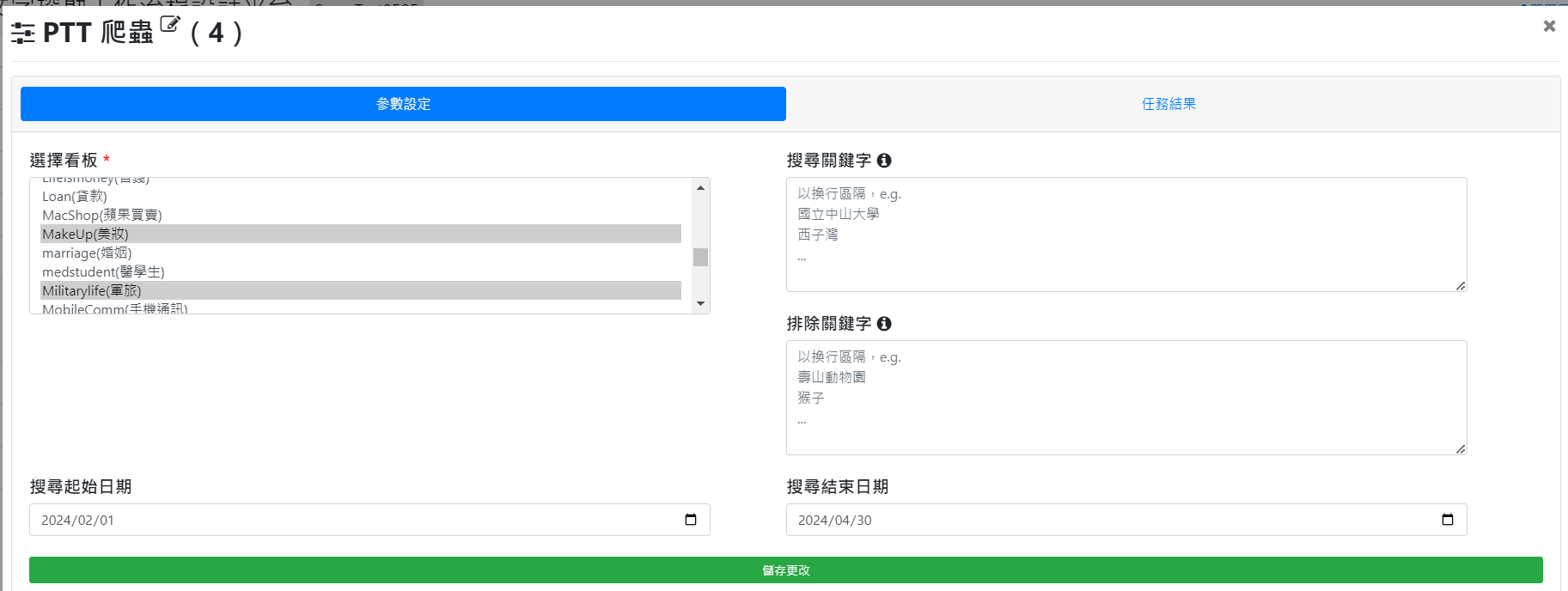
1. **探討主題**

我們透過PTT的媽寶版(媽咪寶貝版)、美妝版和軍旅板這三個版進行主題分析。主要希望深入了解**男性、女性以及小孩在不同生活階段的行為和認知變化**，像是媽咪寶貝版探討了小孩的成長及課業問題等，以及母親在處理家庭與職業間壓力，美妝版則展現了女性利用化妝品來表達自我的價值觀，而且越來越多男性加入美容領域參與討論，而軍旅板則聚焦於男性如何在嚴格的軍事環境中的心智成長和思維轉變。



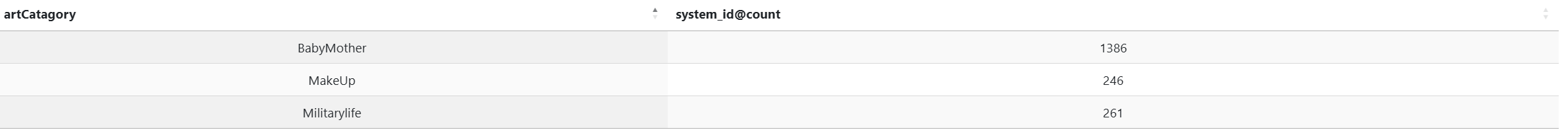
1. **工作流程分析**
   1. 資料來源 :

我們使用PTT的媽寶版、美妝版和軍旅版三個版，抓取資料時間為2024年2月1日至2024年4月30日的資料。



抓取結果如下:

計算了一下這三個版的文章量，分別為1,386則、246則及261則。



分別對這三個版設定了相關字典資料，並清除http相關字元。



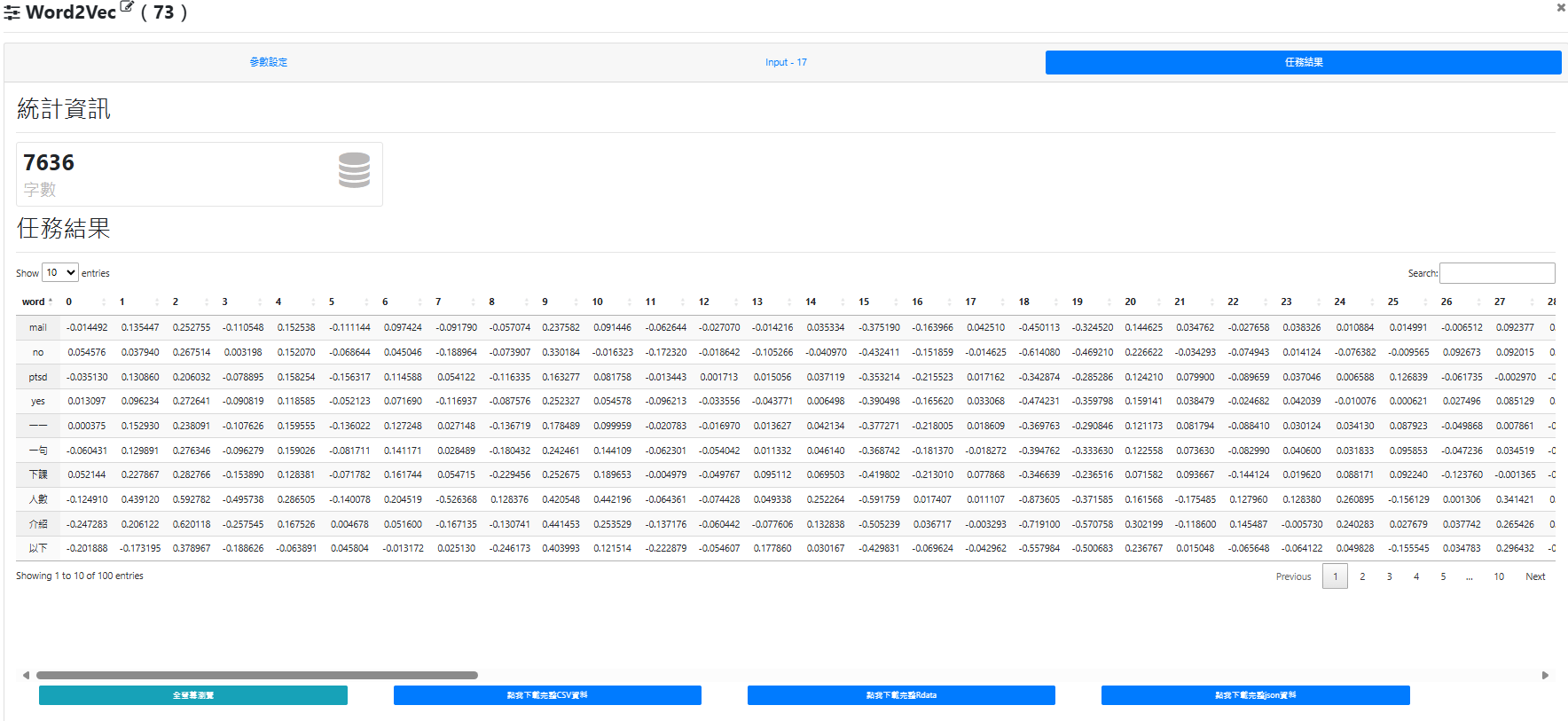
* 1. Work2Vec

我們沿用之前做主題模型分類的三個板分別為美妝/軍旅/媽寶來做這次的Word2Vec。

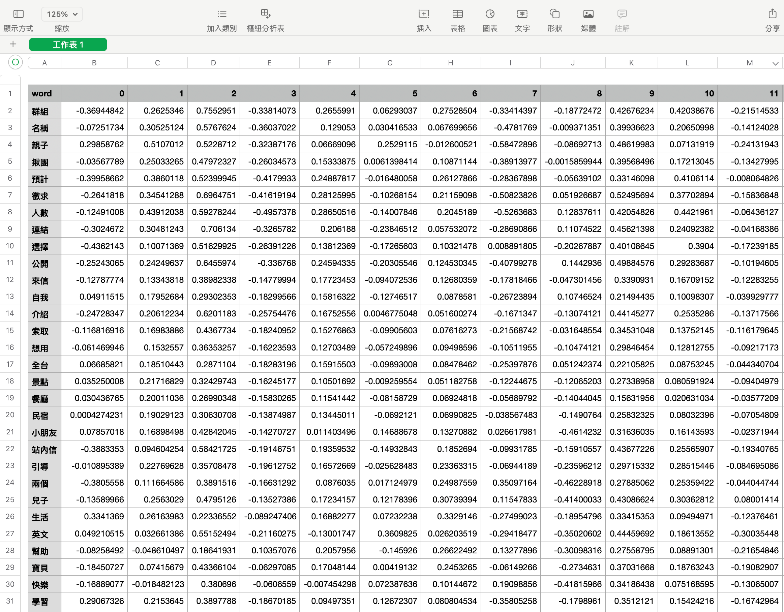
1. 設定參數空間維度為100。



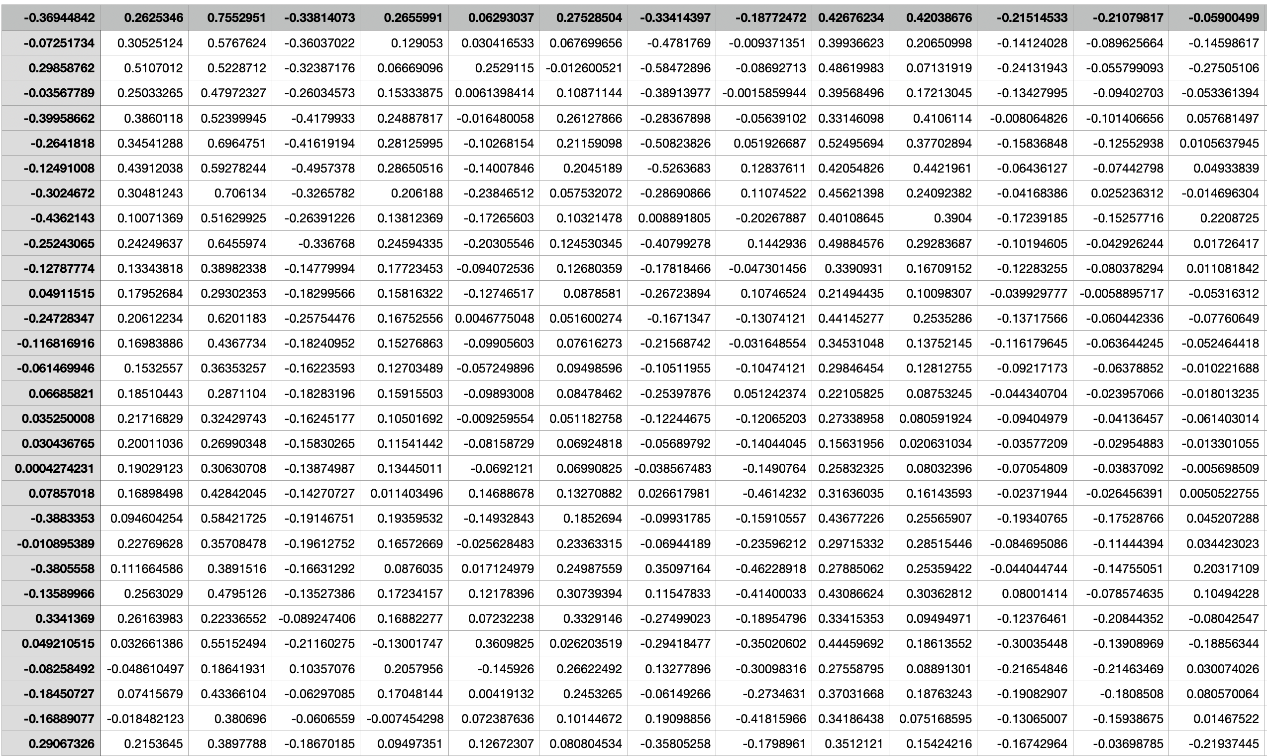
1. 轉為關鍵字詞word跟每一個詞之間的相似度並點選下載完整csv資料。



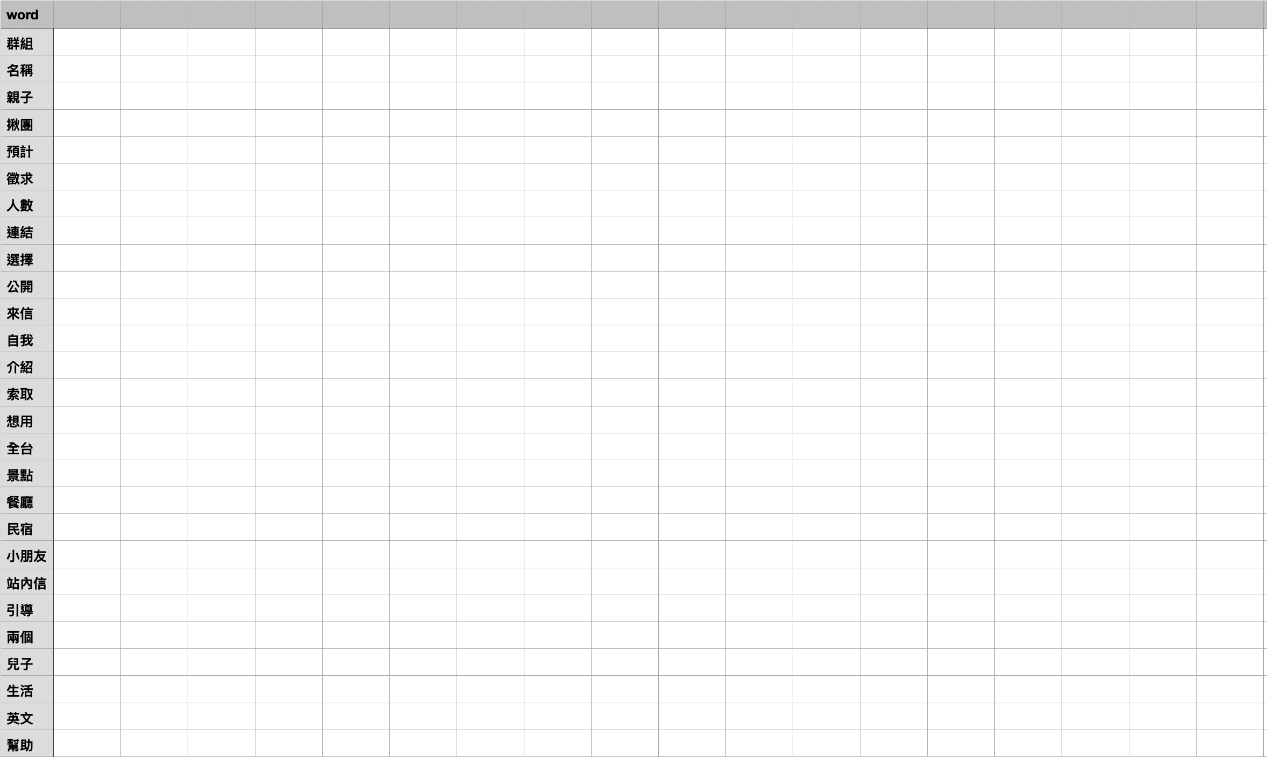
1. 下載後的csv資料轉分別轉為兩份TSV檔分別為Embedding及metadata。



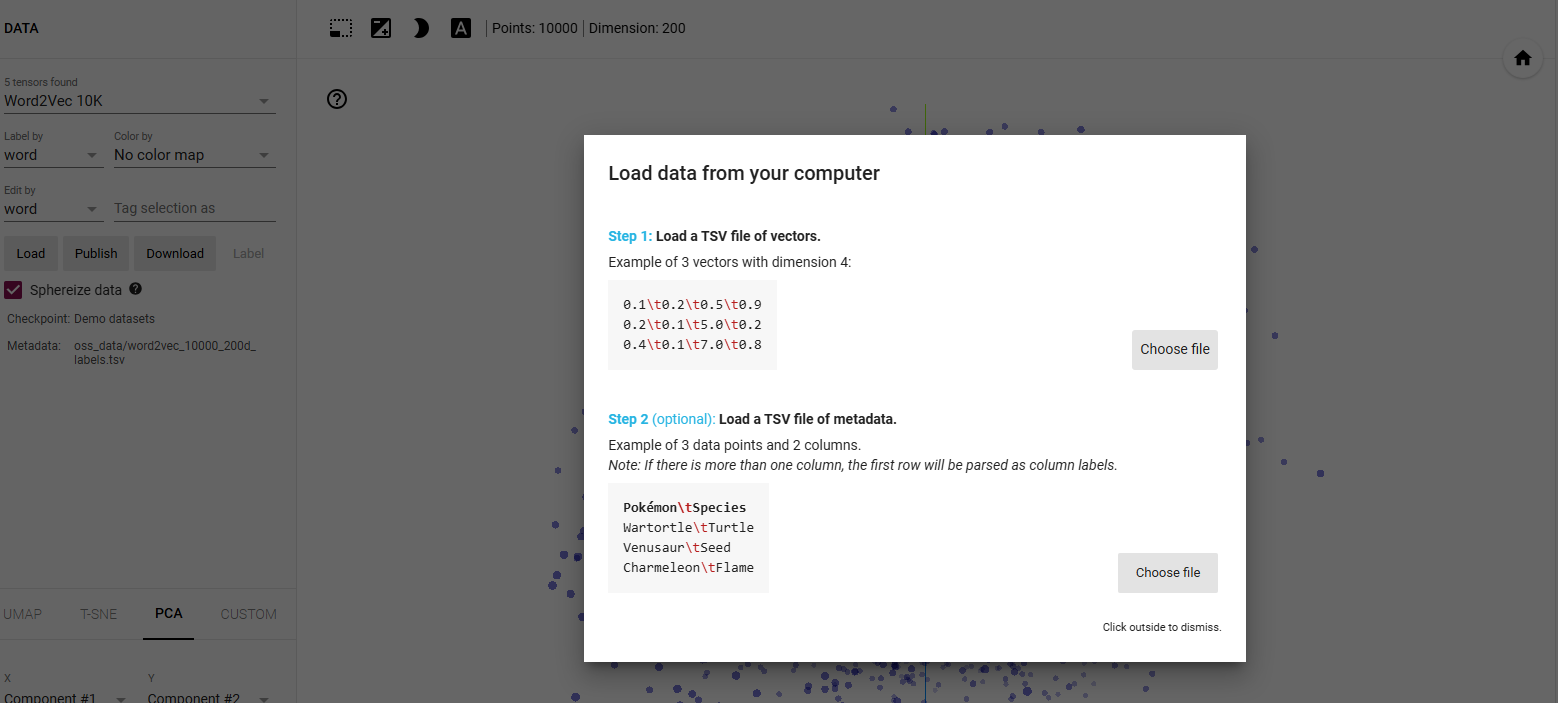
**Embedding**



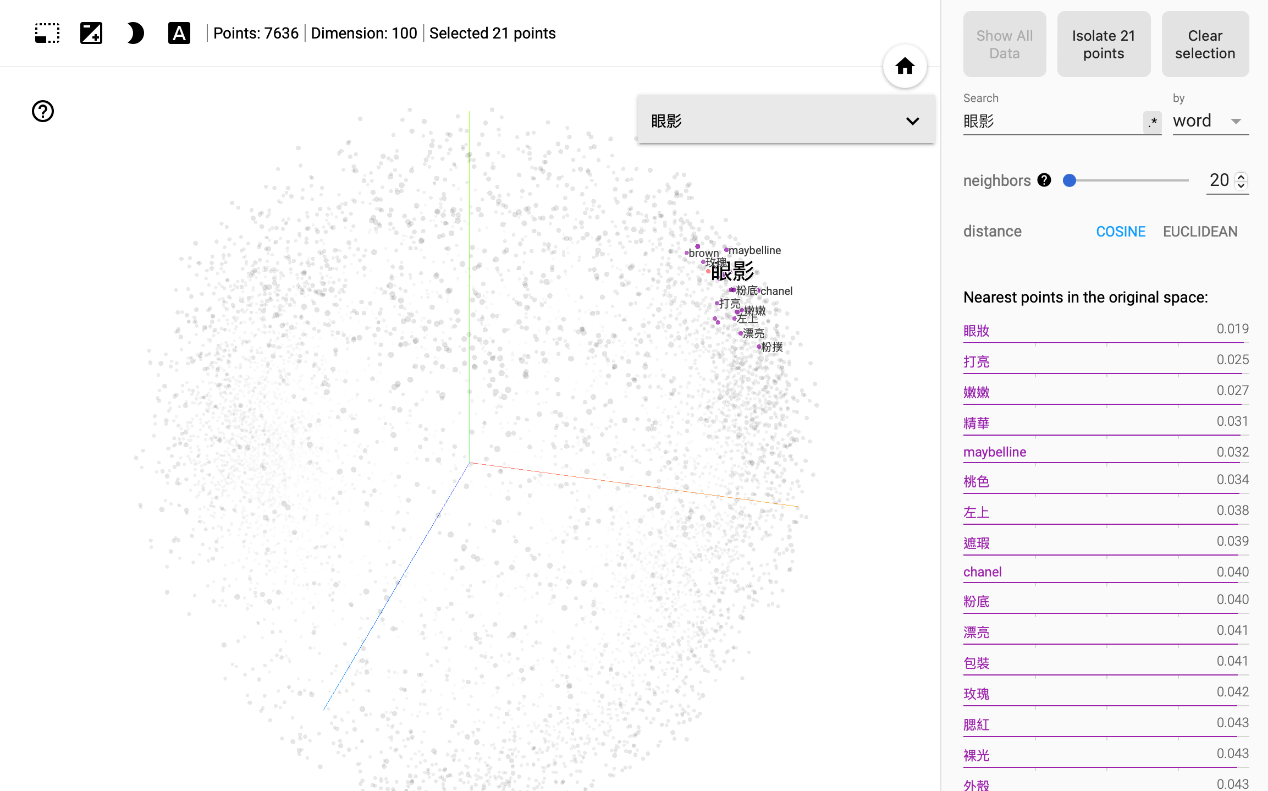
**Metadata**



1. 再到Embedding projector - visualization of high-dimensional data (tensorflow.org)，將剛剛兩個檔案分別匯入。

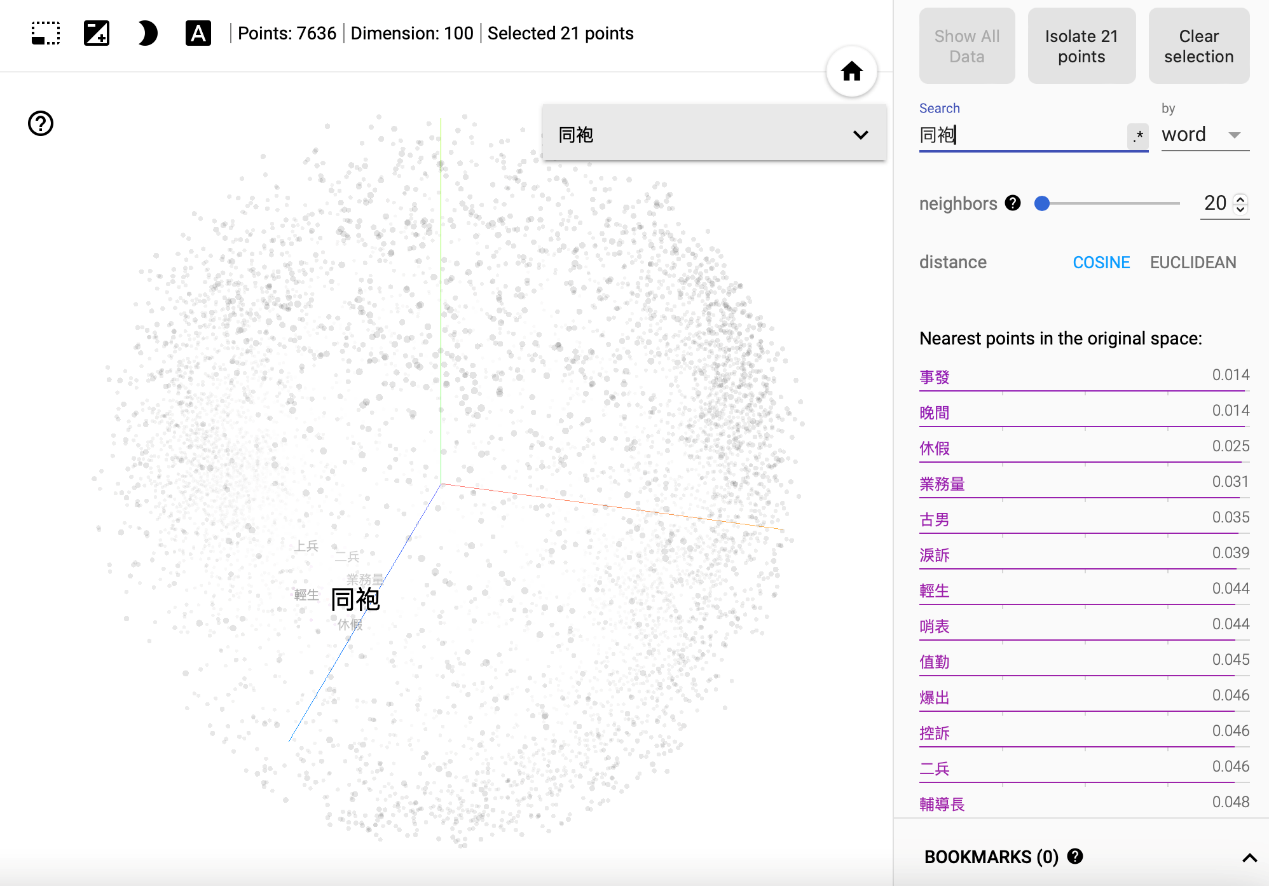


1. 在本次分析中，我們從美妝版的文章中提取了與「眼影」相關的關鍵詞，並使用主成分分析（PCA）技術來探討它們之間的關聯性。結果顯示，「眼妝」、「打亮」、「Maybelline」、「桃色」、「遮瑕」、「Chanel」、「粉底」和「漂亮」等詞彙與「眼影」關聯密切。這些詞彙涵蓋了眼妝技術、品牌名稱、產品顏色和化妝效果，反映了消費者在討論眼影時的多維度關注點。

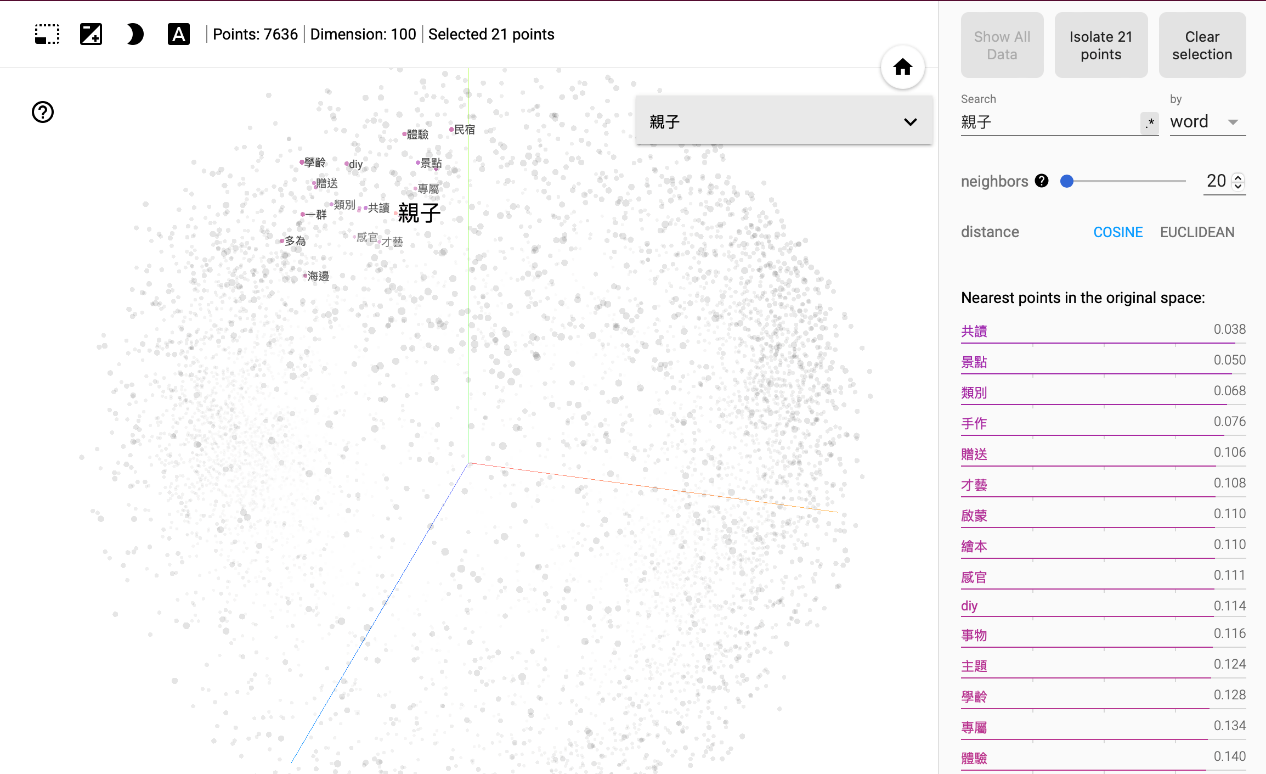


1. 再來選擇同袍來觀察結果發現與「同袍」相關的討論主要集中在事件發生、晚間值勤、休假安排、工作壓力、人際關係和心理健康等方面。

因為我們抓的時間是2024年2月1日至2024年4月30日，此期間發生憲兵211營上兵古姓男子3月1日執勤時持槍自戕身亡在總統府執勤衛哨時開槍自戕，頭部重創送醫不治，古姓上兵同袍控訴其業務繁重休息不足等資訊此事件引發了大量報導和檢討。



1. 與「親子」相關的討論主要集中在共讀、景點、手工製作、才藝培養、早期啟蒙、繪本閱讀、感官活動、DIY 手作、主題活動和學齡期教育等方面。幫助家長更好地理解和安排親子活動，促進孩子的全面發展。



* 1. Bertopic主題模型

我們使用Bertopic 主題模型來測試是否可分析出適當的主題，設定3種不同的參數與主題數的情境。文件向量壓縮維度及主題數設定維以下三種情況:

**情境1**:文件向量設定為10及設定主題數5。



**情境2**:文件向量設定為20及設定主題數5。

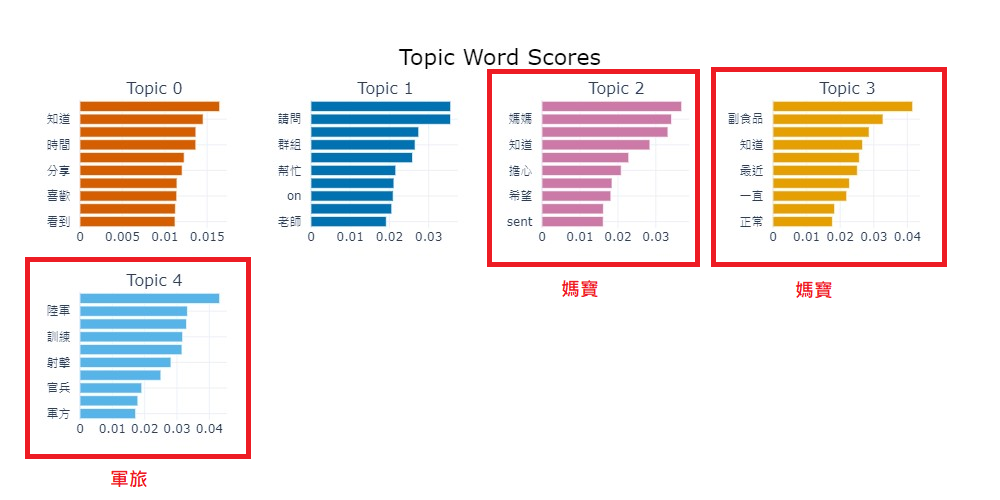


**情境3**:文件向量設定為200及設定主題數7。

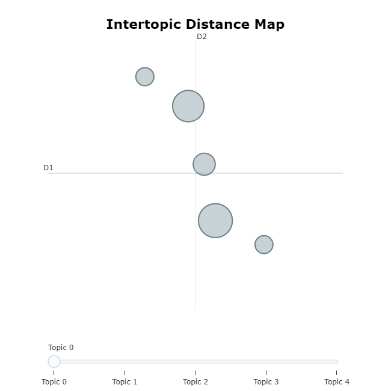


分析結果如下：

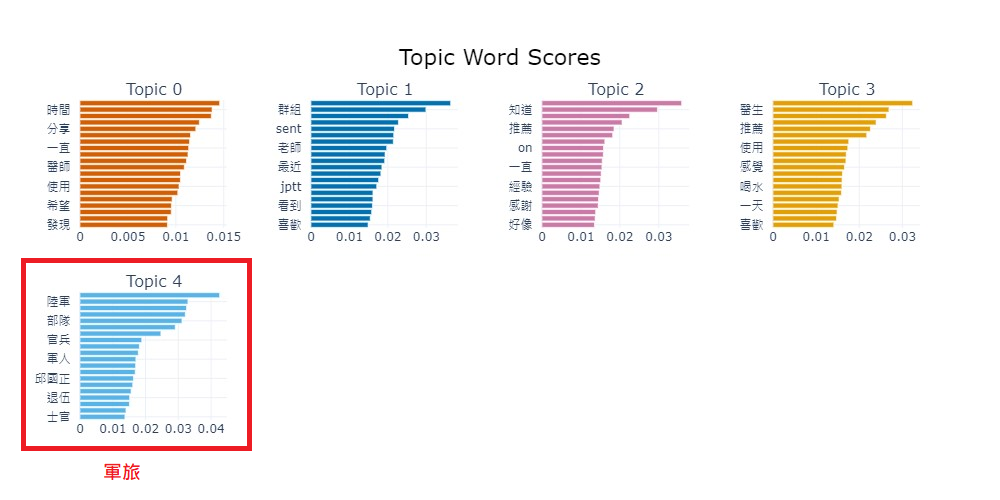
**情境1**:文件向量壓縮維度10 及主題數5，可以看到軍旅板與媽寶版主題有被分類出來，但美妝版的主題並未看到。



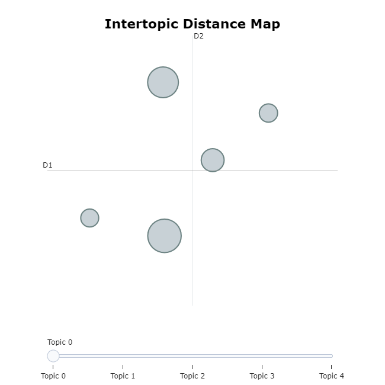
各個主題的**分散性**很**高**



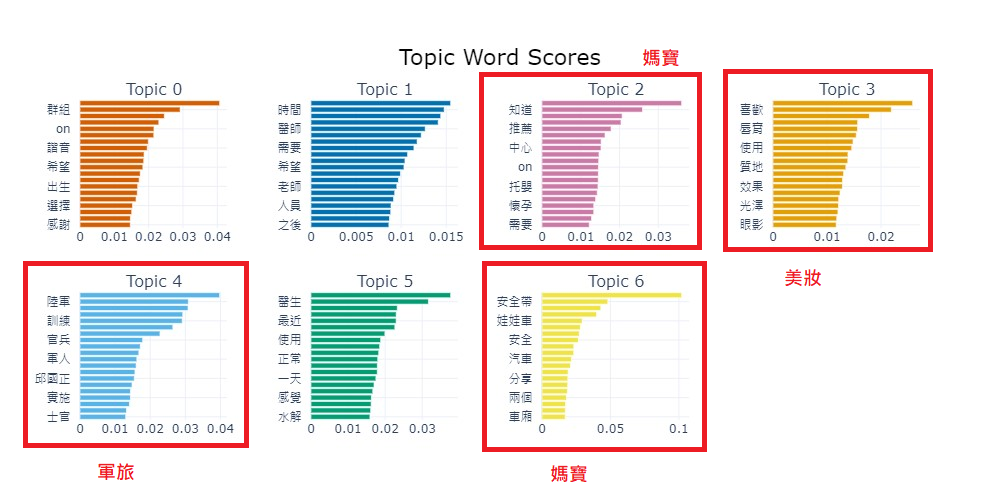
**情境2:** 將文件向量壓縮維度提升到20 及主題數5，分析結果只有軍旅版被分類出來，媽寶跟美妝的文章都沒有出現。



主題的**分散程度**也很**高**

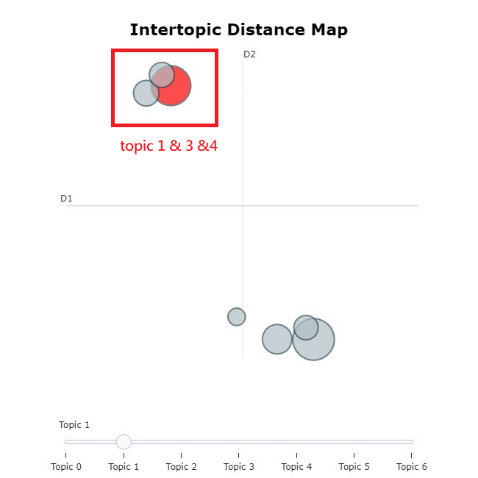


**情境3**: 將文件向量壓縮維度提升到200及主題數7，分析結果可以看出軍旅版仍然也有被正確分類之外，媽寶和美妝的主題也有被分類出來。

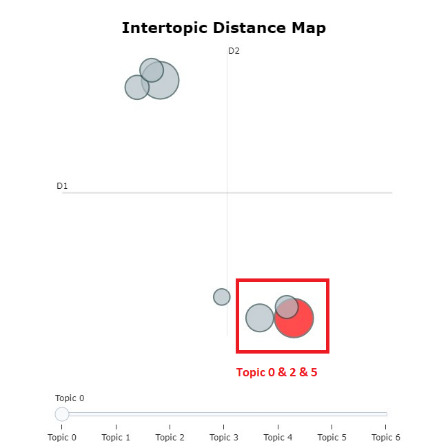


主題分散性主要分為兩類

1. 主題1 & 3 & 4

2. 主題0 & 2 & 5

Bertopic 主題模型embedding的效果比較不好，需要設定較高的參數才有辦法跑出符合的結果，與GuidedLDA 相比，GuidedLDA的分類效果較好。

最後我們嘗試PCA 維度縮減後再跑SVC，我們的核心使用rbf 、Linear、poly準確率先設為1，結果分析出的準確率僅有0.277。



我們將準確率微幅提升到10-1000時，準確率會提到0.285，沒有明顯變化



將核心改為sigmoid ,準確度更大幅下降到0.261，或許是因為媽寶版的話題非常廣泛，從醫療、教育、資源利用、社群活動、心理諮詢、心情舒壓，職涯規劃等，因為其他美妝也與醫療有相關、軍旅與職業規劃與政府相關政策有關係，使得媽寶的主題相當複雜而造成準確率低下。

* 1. Rag

A screen shot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

先透過Python將資料做初步的處理，如清除停用字、選擇特定分類(美妝版)等，之後刪除不會用到的欄位，僅存序列以及文章內容再以txt檔輸出。

A screenshot of a computer screen

Description automatically generated

之後透過網路上的工具，將txt檔轉換成pdf檔，

A screenshot of a computer

Description automatically generated

匯入進去GPT4ALL，之後對其做embedding的動作，做RAG的準備。

A screenshot of a computer

Description automatically generated

一開始先純粹採用gpt3.5的模型，來觀察他對於問題的解答情形，可以看到在沒有外在資料的輔佐之下，他已經回答得很完整，

A screenshot of a computer

Description automatically generated

之後，再gpt3.5之下，再搭配前述上傳外部ppt美妝版資料。可以發現整體來說所回答的內容差異不大，但由於有多一個外部參考資料的原因，因此，針對前兩個題目的回答，內容都有所增加，也與外部資料內容相符合。

A close-up of a text

Description automatically generated



因為場所的部分在文章內容中被提及，將其匯入模型中，有豐富了gpt3.5可以回答的內容。





文章中特別提到防曬係數的部分，在第二次的回答中，gpt3.5也將防曬的特點新增進去產品的特點當中。

* 1. 社會網路





我們用留言萃取分析媽寶、美妝、軍旅三個版的留言，經過分群彙總用社會網路分析後，發現留言的集中性很高，留言互動比較集中在少數的文章，所以可以從圖表中發現有許多圈圈對應到一則文章，呈現scale-free意見領袖的模式。

1. **結語**

本次研究透過對 PTT 媽寶版、美妝版和軍旅版的主題分析，得出了幾項重要結論。

* 1. 研究顯示**不同性別和年齡段**在各自生活階段中展現了**顯著的行為和認知差異**。

媽寶版主要關注於育兒、教育和家庭壓力，美妝版則聚焦於美容產品的使用和自我表達，而軍旅版則探討了軍事生活對男性心智和思維的影響。

* 1. 透過資料清理、Word2Vec 和主題模型（如 Bertopic 和 GuidedLDA）的應用，我們能揭示討論中的關鍵詞和主題，特別是 Bertopic 模型在設定高參數時能夠更準確地分類出具體主題，但 GuidedLDA 的分類效果相對較佳。
  2. 在 Bertopic 主題模型中，美妝版的分類結果**不如預期**，需要**調整較高的參數設置**才能獲得較為清晰的分類效果，這可能與美妝版文章中與媽寶版**重疊的詞彙**較多有關。
  3. 在使用 GPT-3.5 模型進行問答分析時，加入外部資料能夠豐富模型的回答內容，提升回答的完整性和準確性。
  4. 社會網路分析顯示，留言的集中性較高，互動主要集中在少數文章上，呈現出 scale-free 意見領袖模式，顯示在這些討論版中，少數意見領袖的觀點對於整體討論有著重要的影響力。

總結:針對不同版面討論的主題和特徵，可了解各種資料處理和分析技術的應用效果。未來的研究可以調整主題模型的參數設置，並探索更多樣的資料來源，可提升分析的準確性和廣泛性。