

**目錄**

[壹、探討主題 2](#_Toc165979193)

[貳、工作流程分析 3](#_Toc165979194)

[一、資料來源 : 3](#_Toc165979195)

[二、DTM彙總 4](#_Toc165979196)

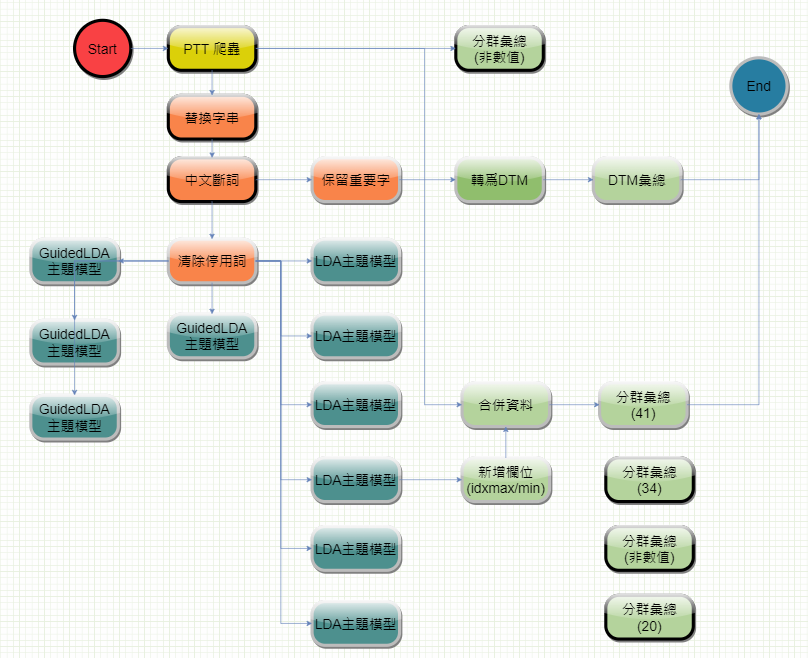
[三、潛在語意分析（LDA） 4](#_Toc165979197)

[四、GuidedLDA主題模型 13](#_Toc165979198)

[參、結語 23](#_Toc165979199)

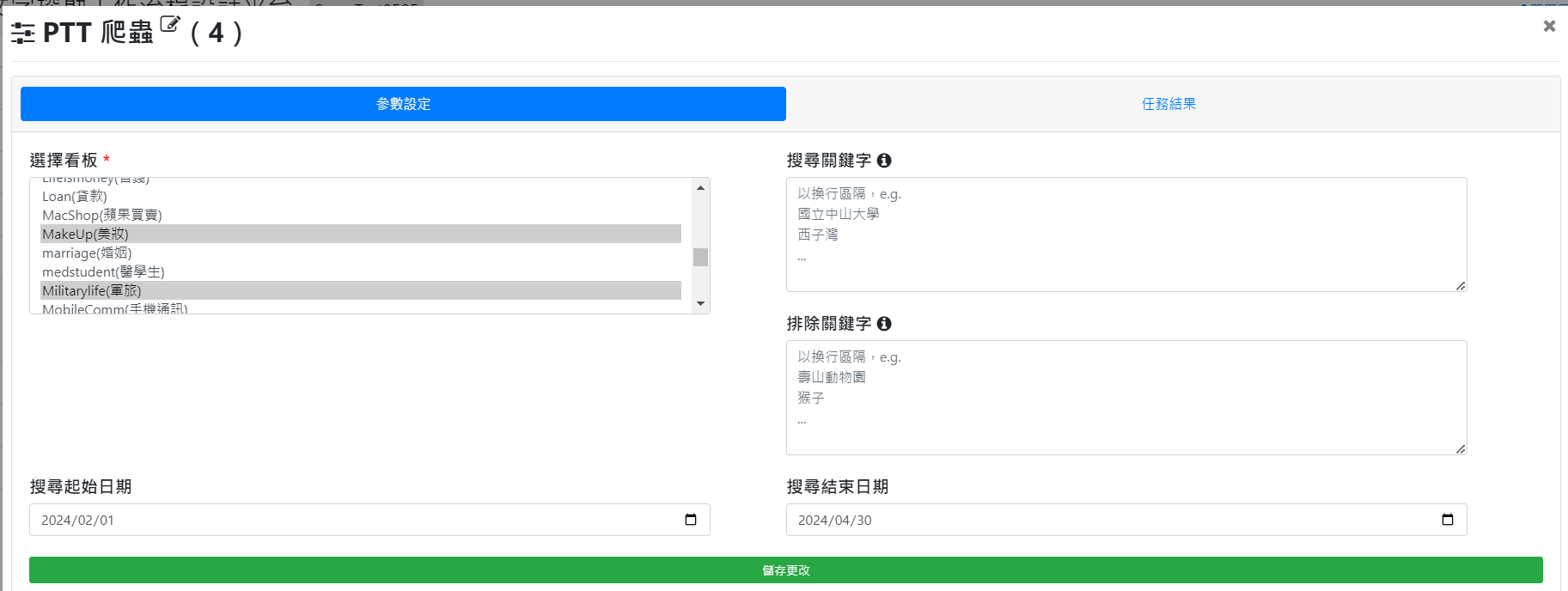
1. **探討主題**

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。



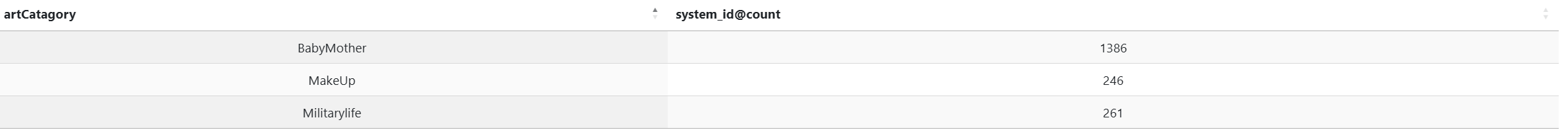
1. **工作流程分析**
   1. 資料來源 :

抓取PTT的媽寶版、美妝版和軍旅版三個版，時間為2024/2~2024/4月的資料。



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。

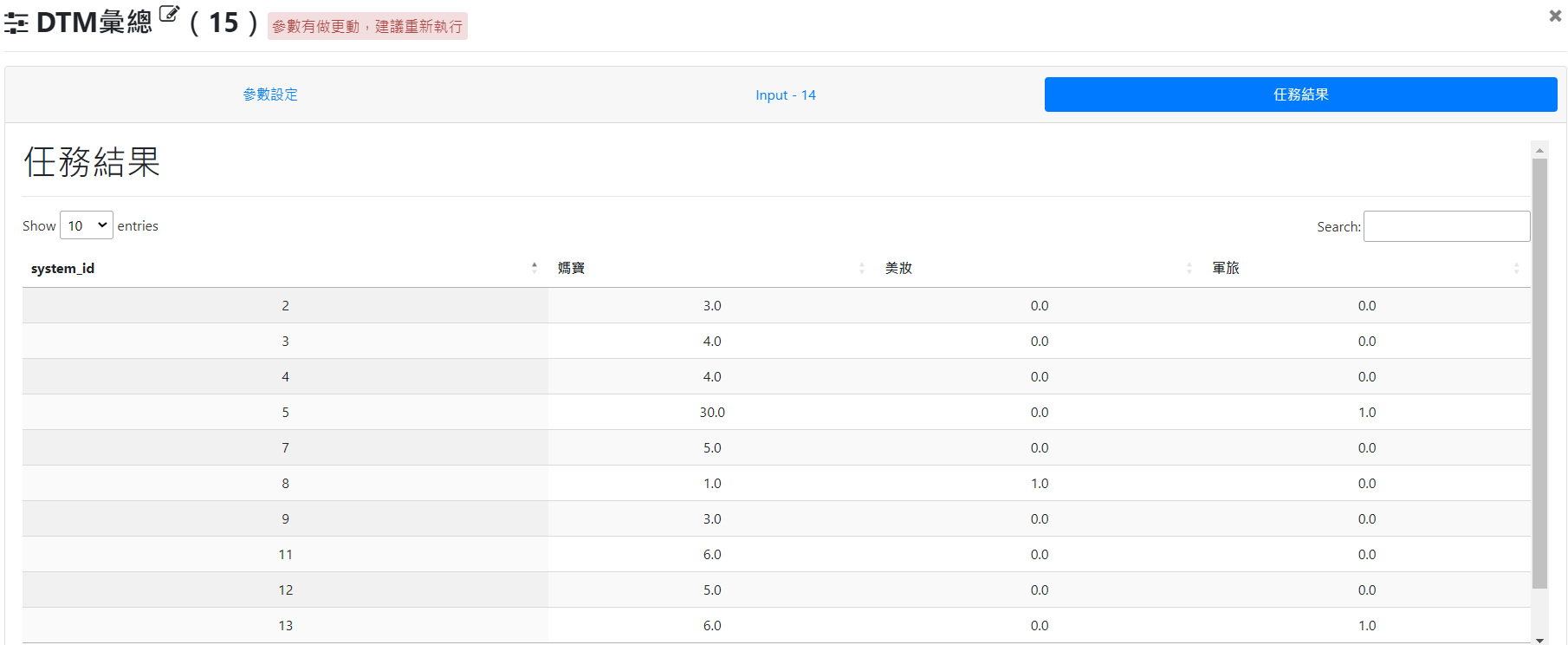


* 1. DTM彙總

根據我們字典所產生的DTM彙總結果，每篇文章都會給予相關的分數。

檢視任務結果大部分文章都可以正確的被分類出來。





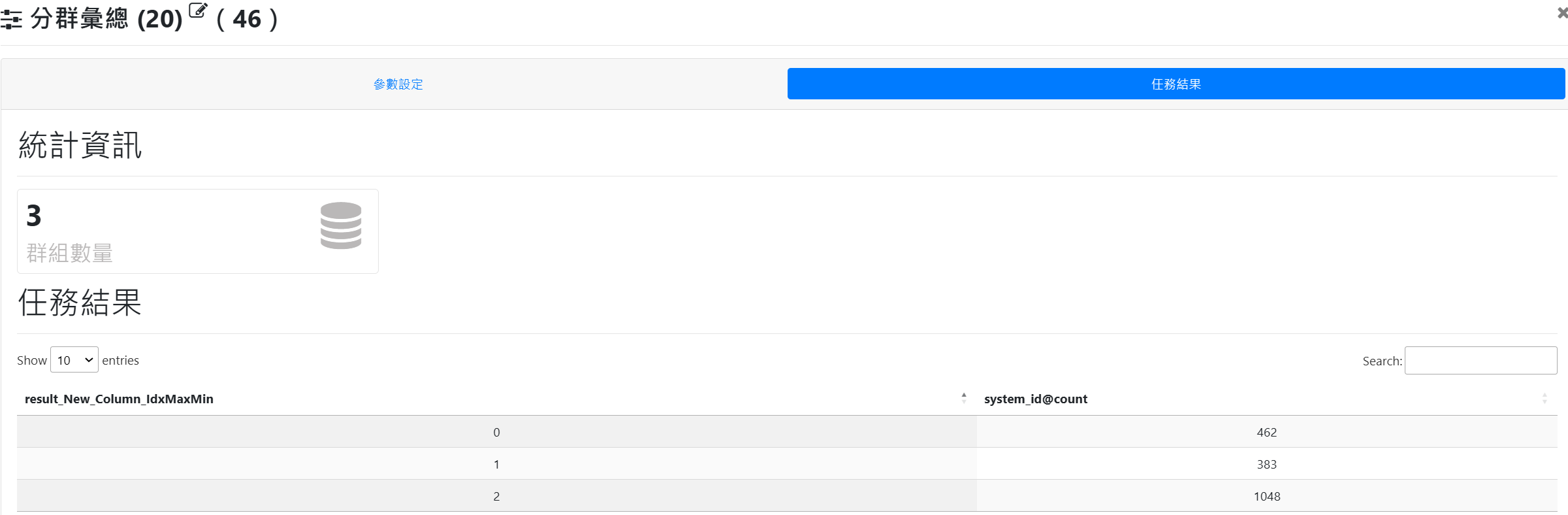
* 1. 潛在語意分析（LDA）

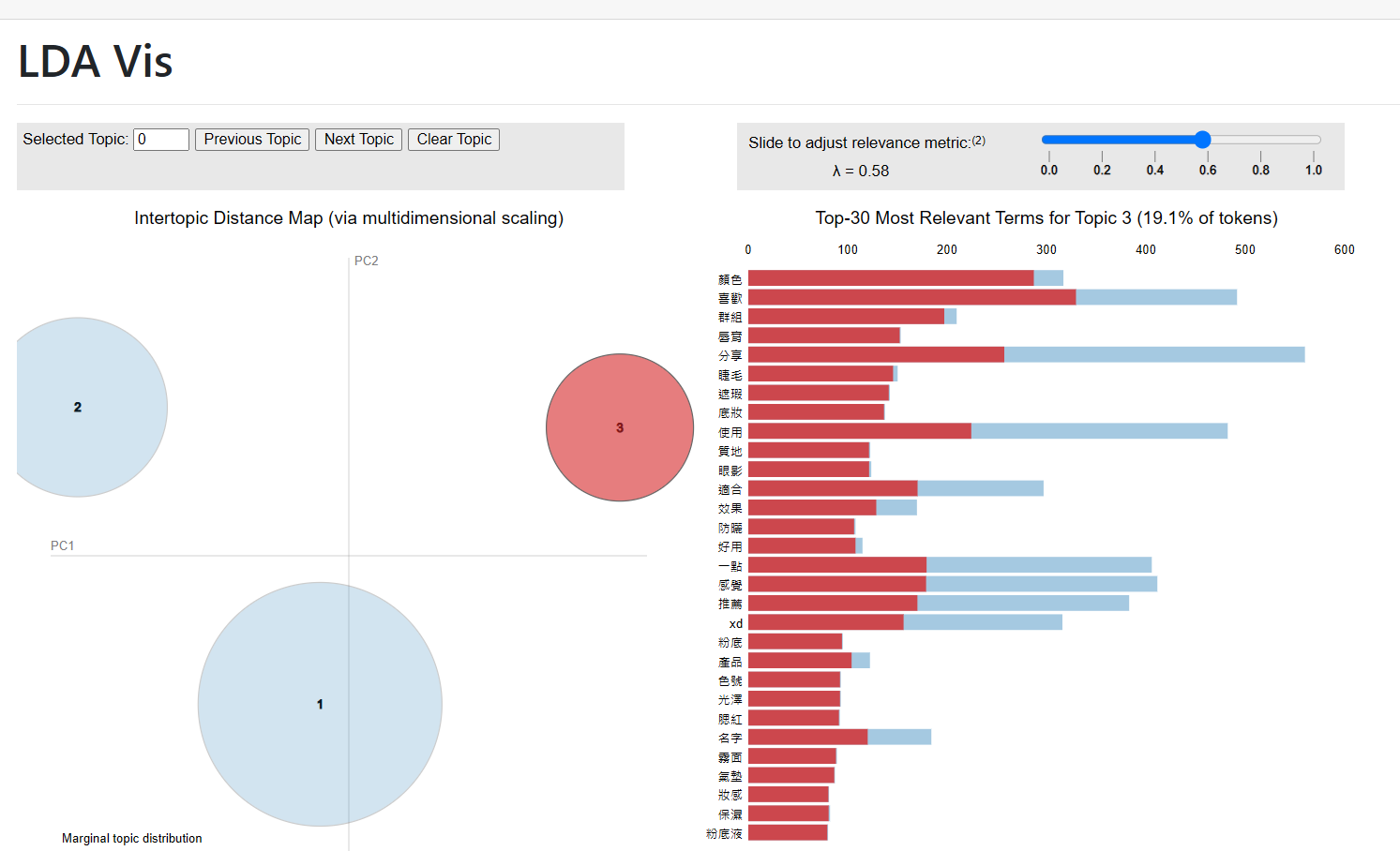
我們先設定主題數為3、詞彙頻率上限設為20，詞彙頻率下限設為0.7。跑出來的結果混淆度763、主題連貫度-2.4。實際觀察在美妝相關主題的分類中，有少許詞彙與其他主題類別交疊、軍旅與媽寶相關的主題表現出很好的獨立性和清晰度。

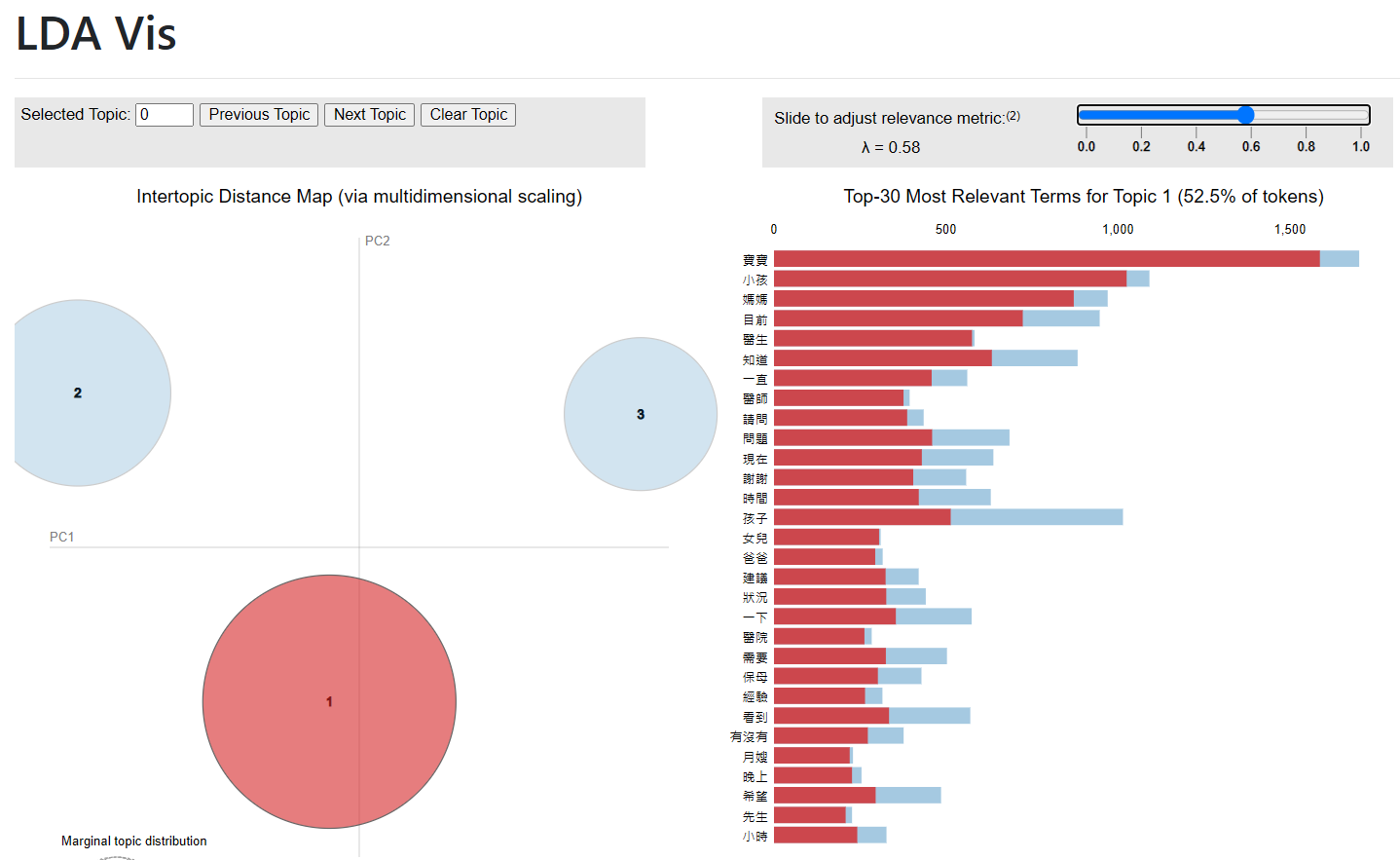
我們也可以從分群彙總的數字看出雖然文章分類有些誤判，但比例還是維持的不錯。

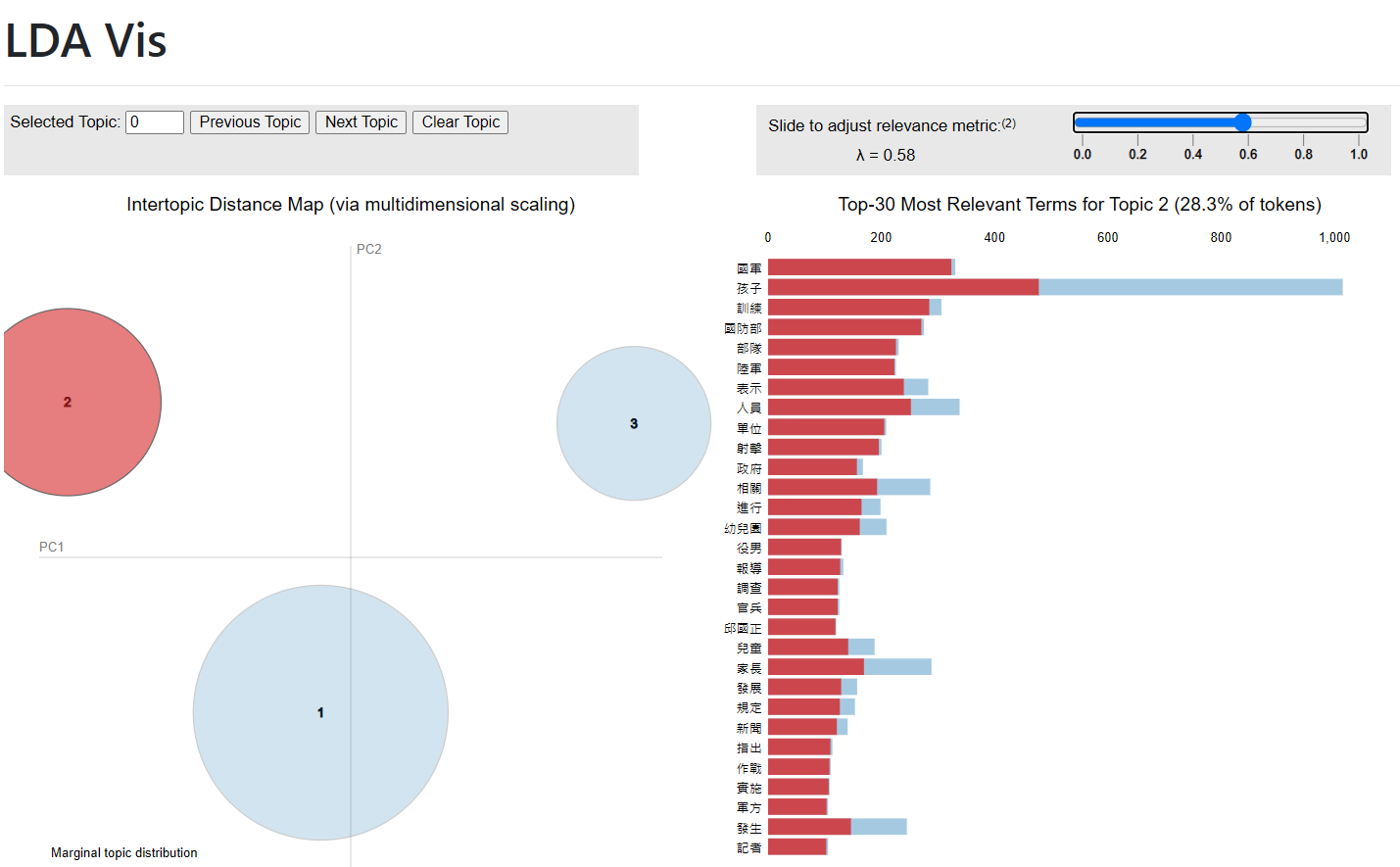




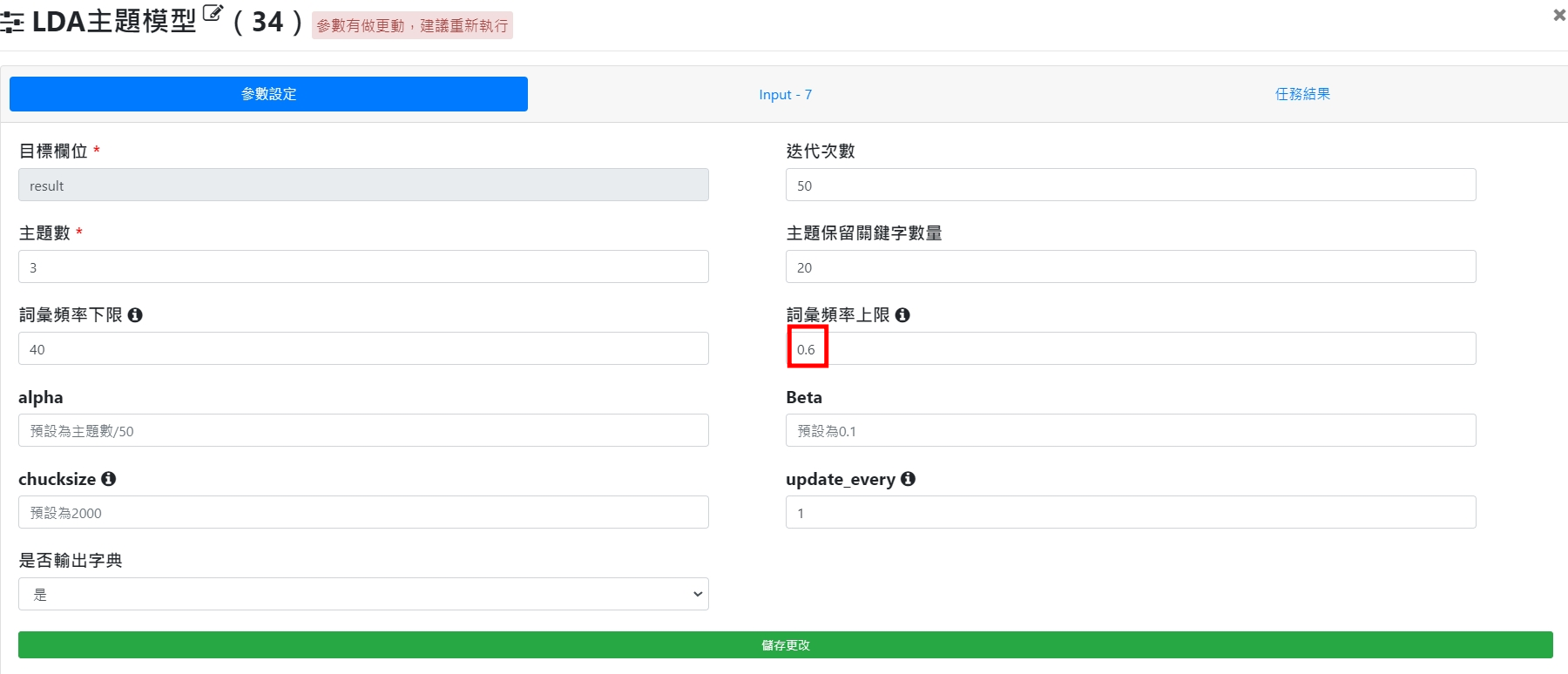




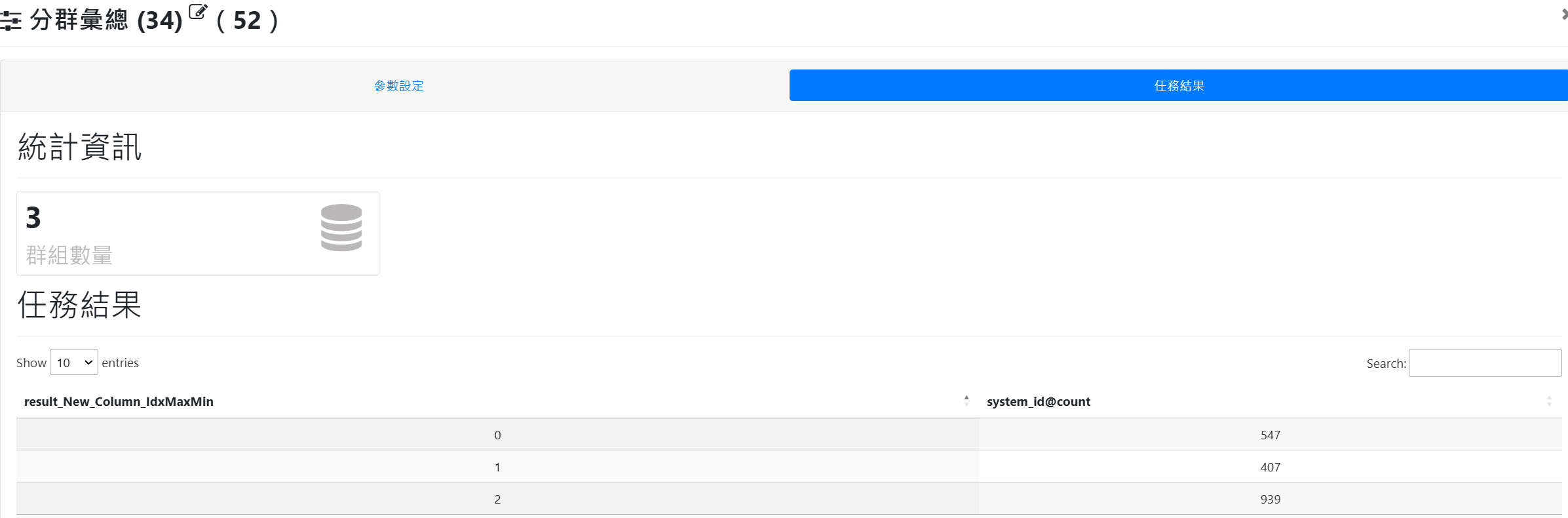




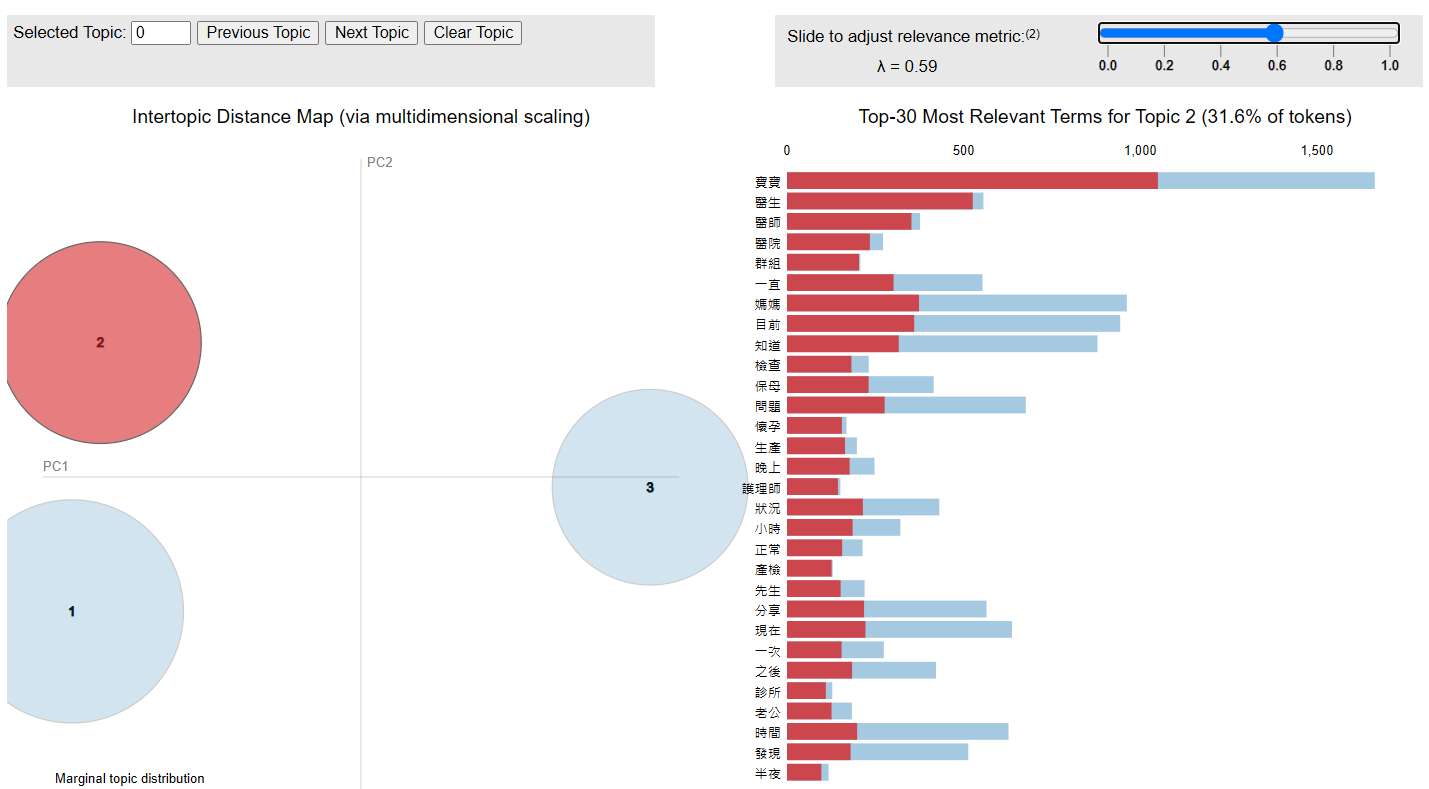
我們接下來調整詞彙頻率**下限設為40**，詞彙頻率**上限設為0.6**。跑出來的結果混淆度401、主題連貫度-2.0。雖然混淆度、主題連貫度數字表現變好了，但實際觀察軍旅主題還是有正確的抓出來，但美妝的主題已經消失，這可能是由於較高的詞彙上限和較低的頻率下限允許了更少的常見詞彙進入模型，進而降低了主題之間的區分度。我們也可以從分群彙總的數字看出有媽寶與美妝的文章被大量誤判了。













我們接下來調整詞彙頻率**下限設為10**，詞彙頻率**上限設為0.8**。跑出來的結果混淆度1553、主題連貫度-2.3。混淆度、主題連貫度數字表現變的更差，實際觀察這次則是美妝的主題還存在，但軍旅的主題就消失了。

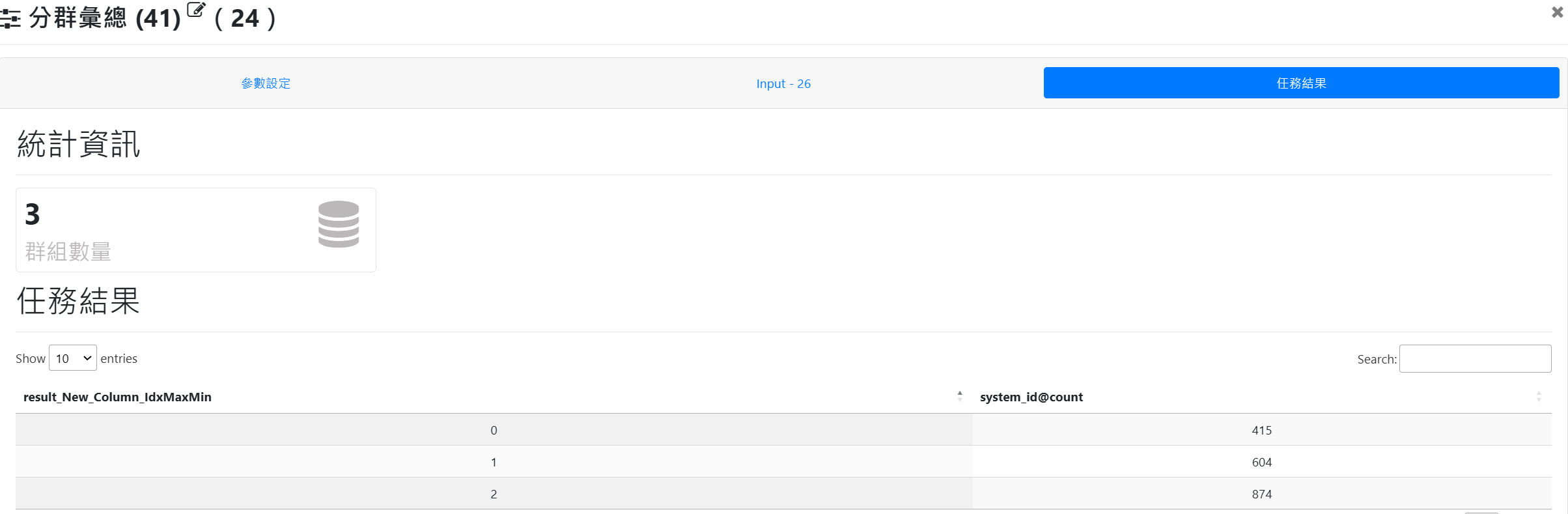
我們也可以從分群彙總的數字看出有媽寶與軍旅的文章被大量誤判了。

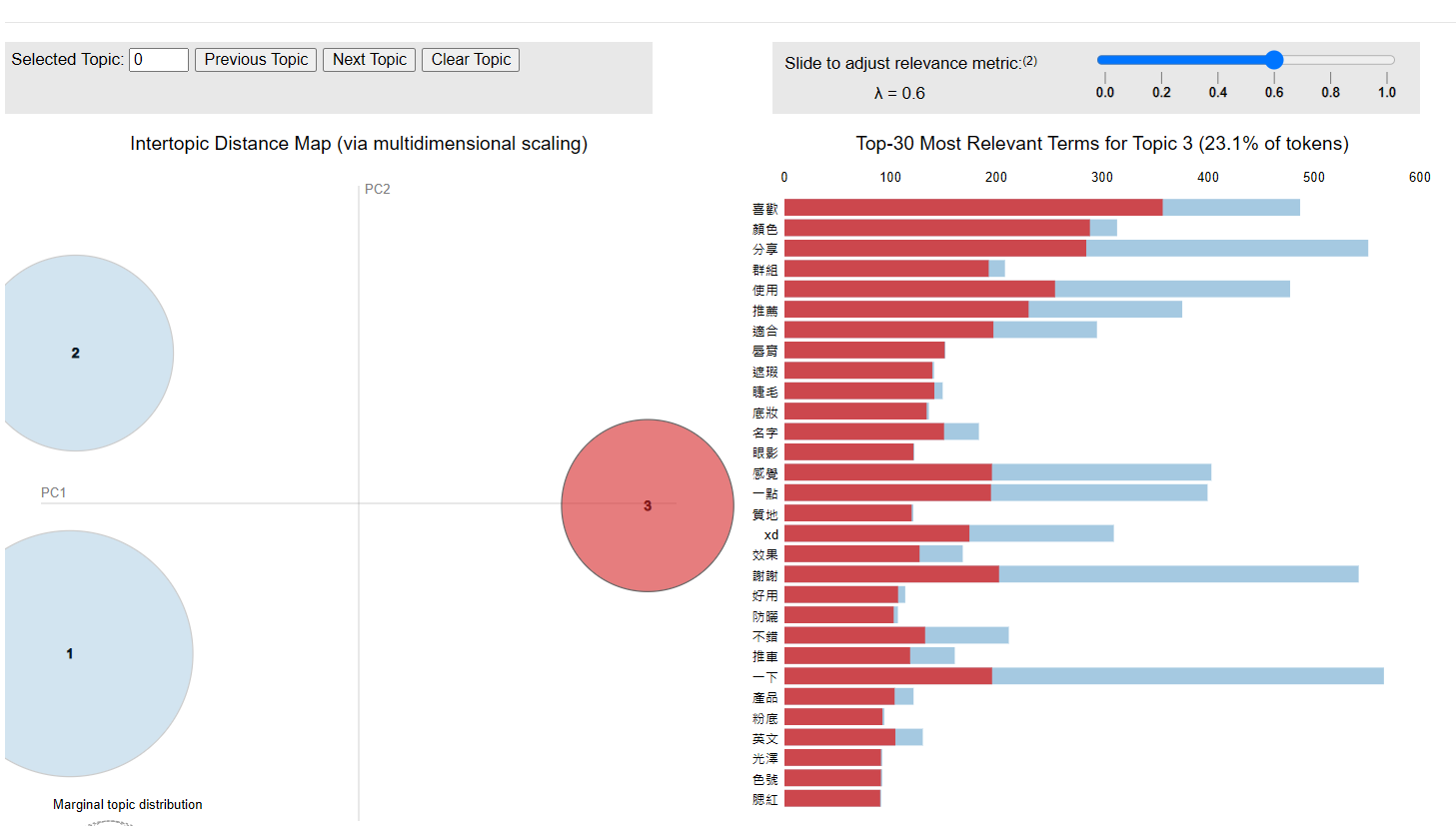
所以詞彙頻率的上下限不能一昧的求高或求低，也沒辦法從混淆度、主題連貫度的數字看出分類的好壞，實際在跑模型時還是需要多調參多看看實際分類的狀況。

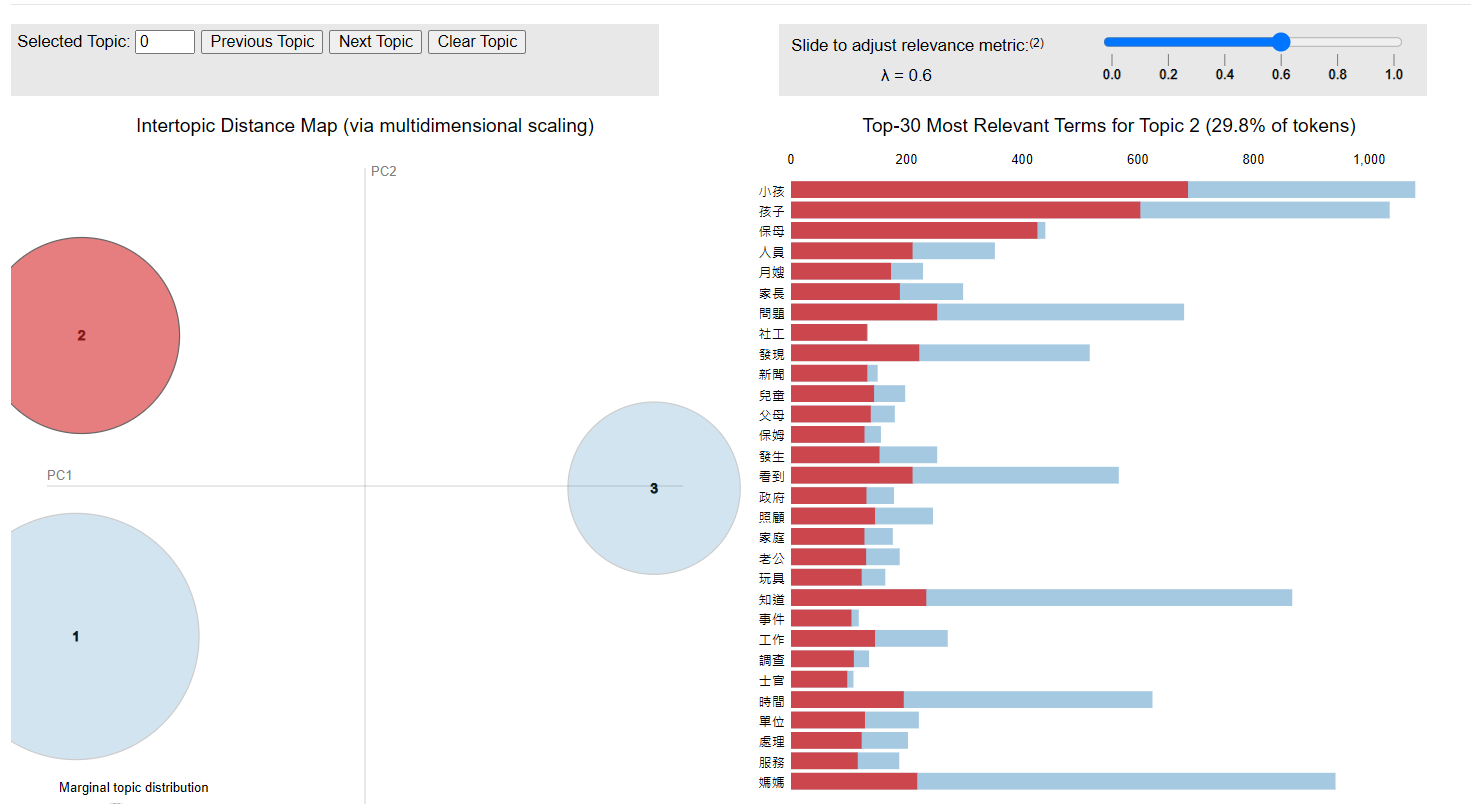
A screenshot of a computer

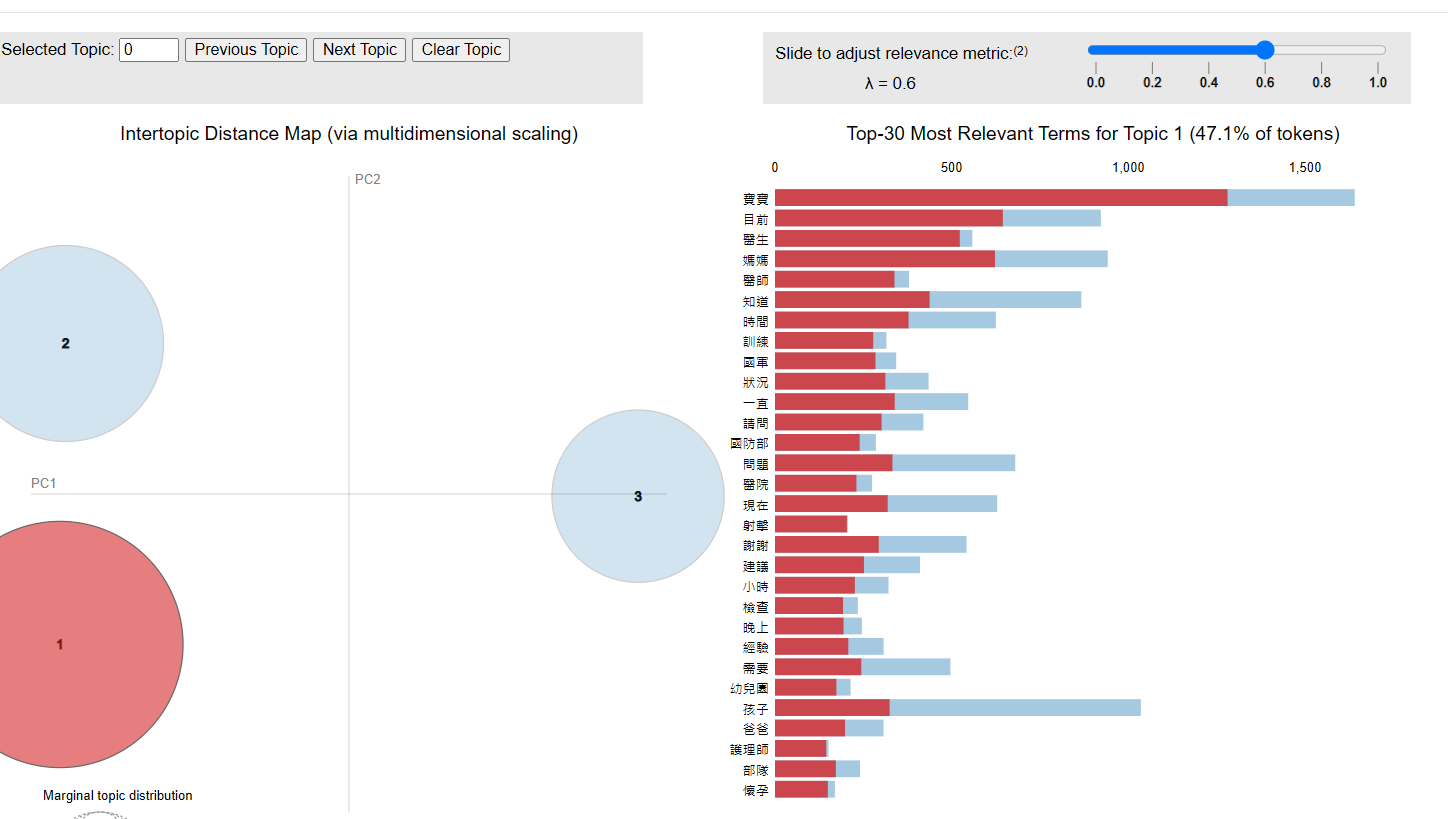
Description automatically generated









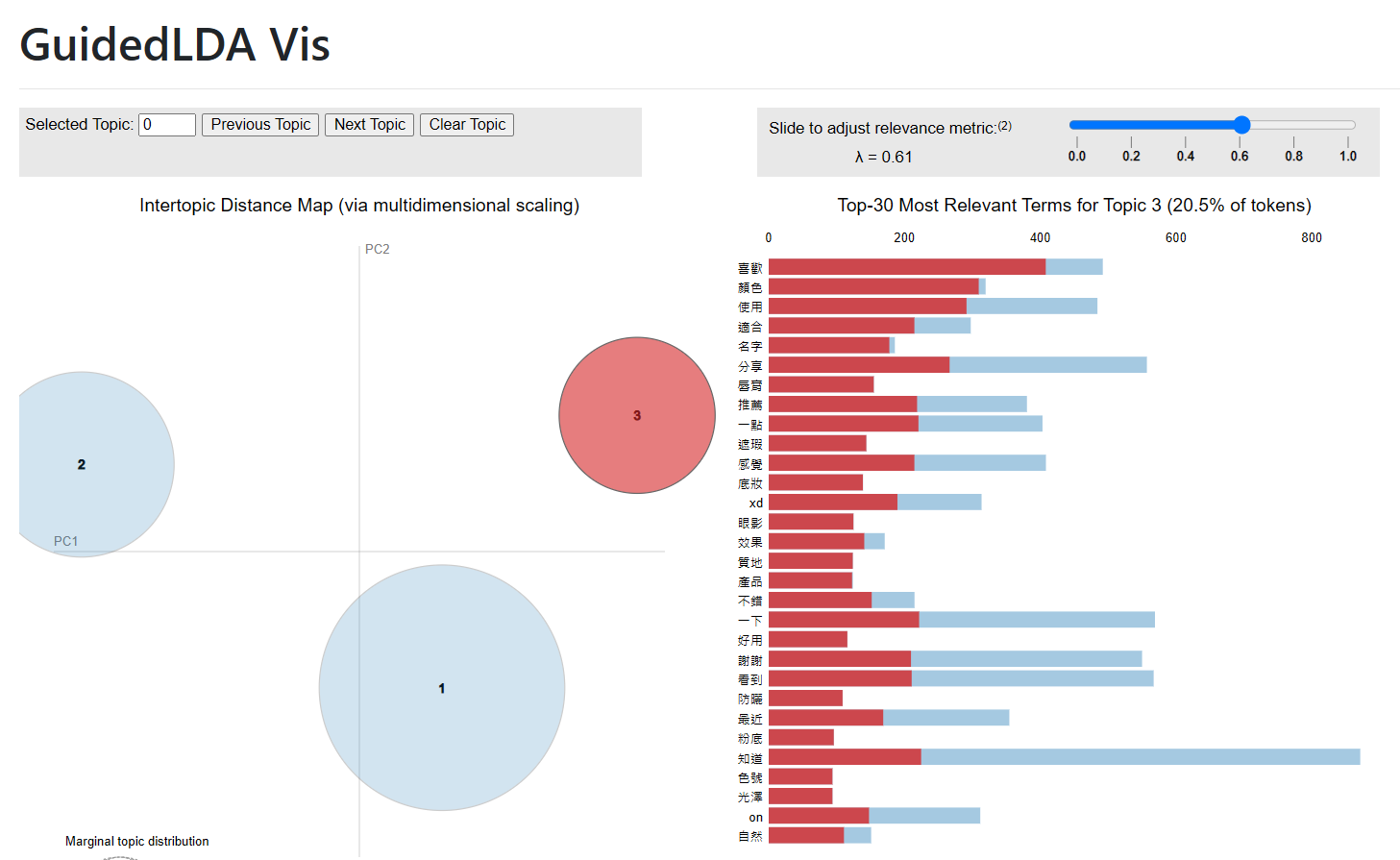


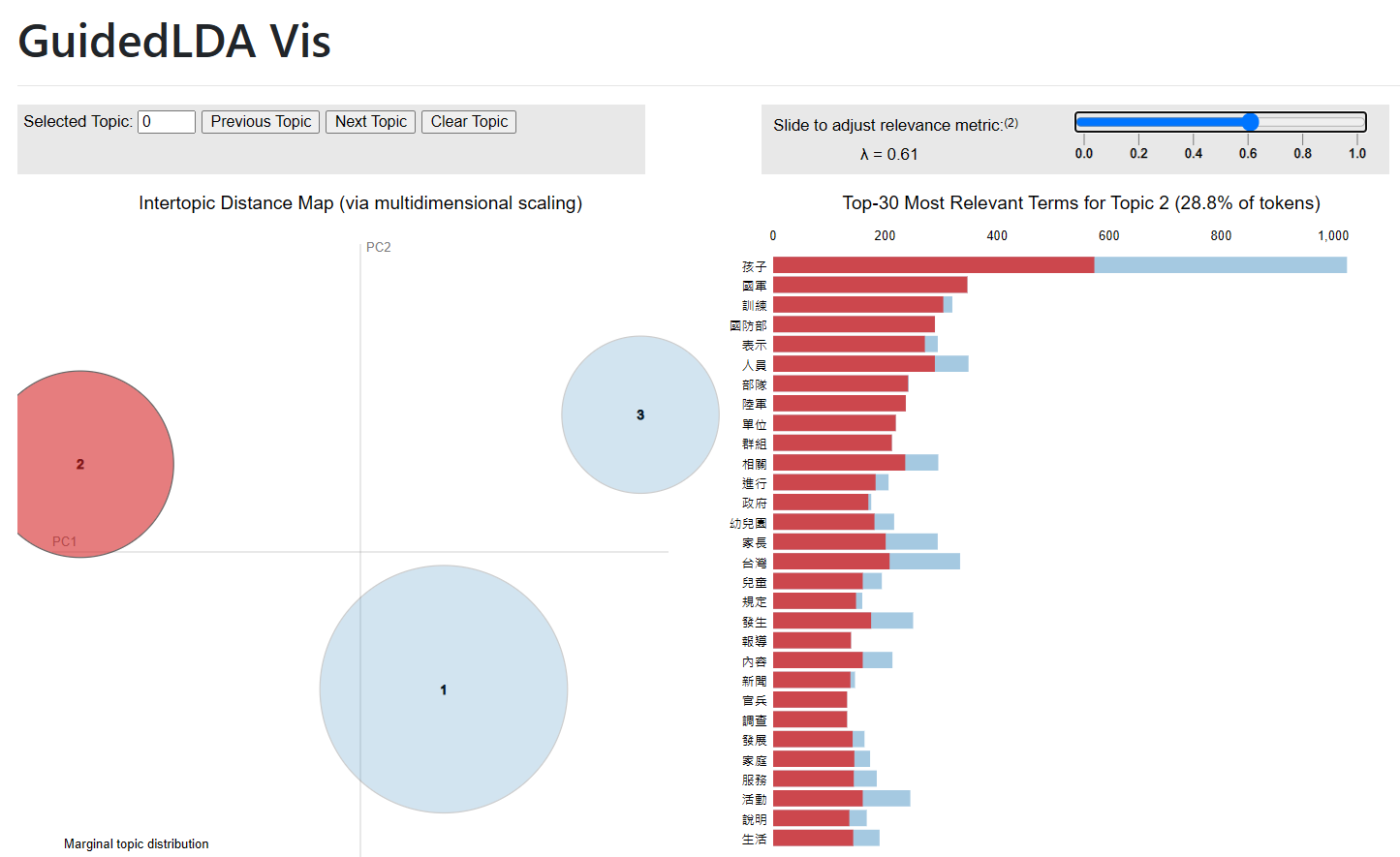
* 1. GuidedLDA主題模型

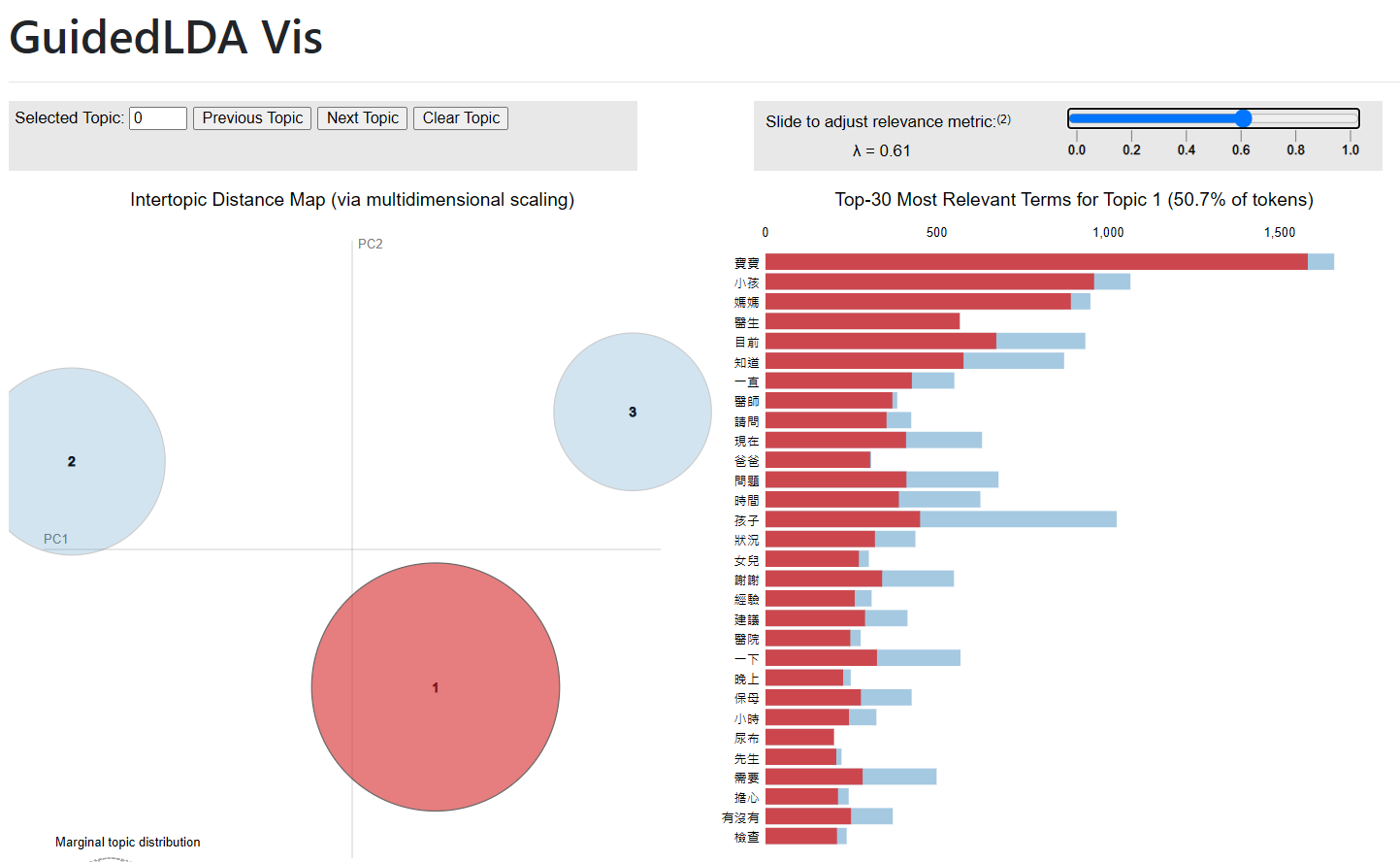
當缺乏適合的字典支持LDA模型時，我們可以採用GuidedLDA主題模型來實現自動分類。即使缺乏好的字典，在GuidedLDA主題模型中，我們透由測試不同的參數設定，促發主題的生成。我們可以先自行設定主題數量，然後讓系統根據文本的內在結構和特徵進行自動分類。這與LDA模型設定相同，在參數設定中，主題數量、詞彙頻率下限和上限這3個欄位都與LDA模型參數設定相同。透過GuidedLDA，我們可以更有效地利用文本資料，快速且準確地進行主題分類。這種方法有助於提高模型的性能和效率，並且適用於各種文本分類任務。



在主題數為3，詞彙頻率下限為40，詞彙頻率上限為0.6的設定下，比起同樣參數的LDA效果令人滿意，但還是有抓到其他主題的詞，例如在軍旅的主題中有孩子與幼稚園，這兩個應該出現在媽寶。

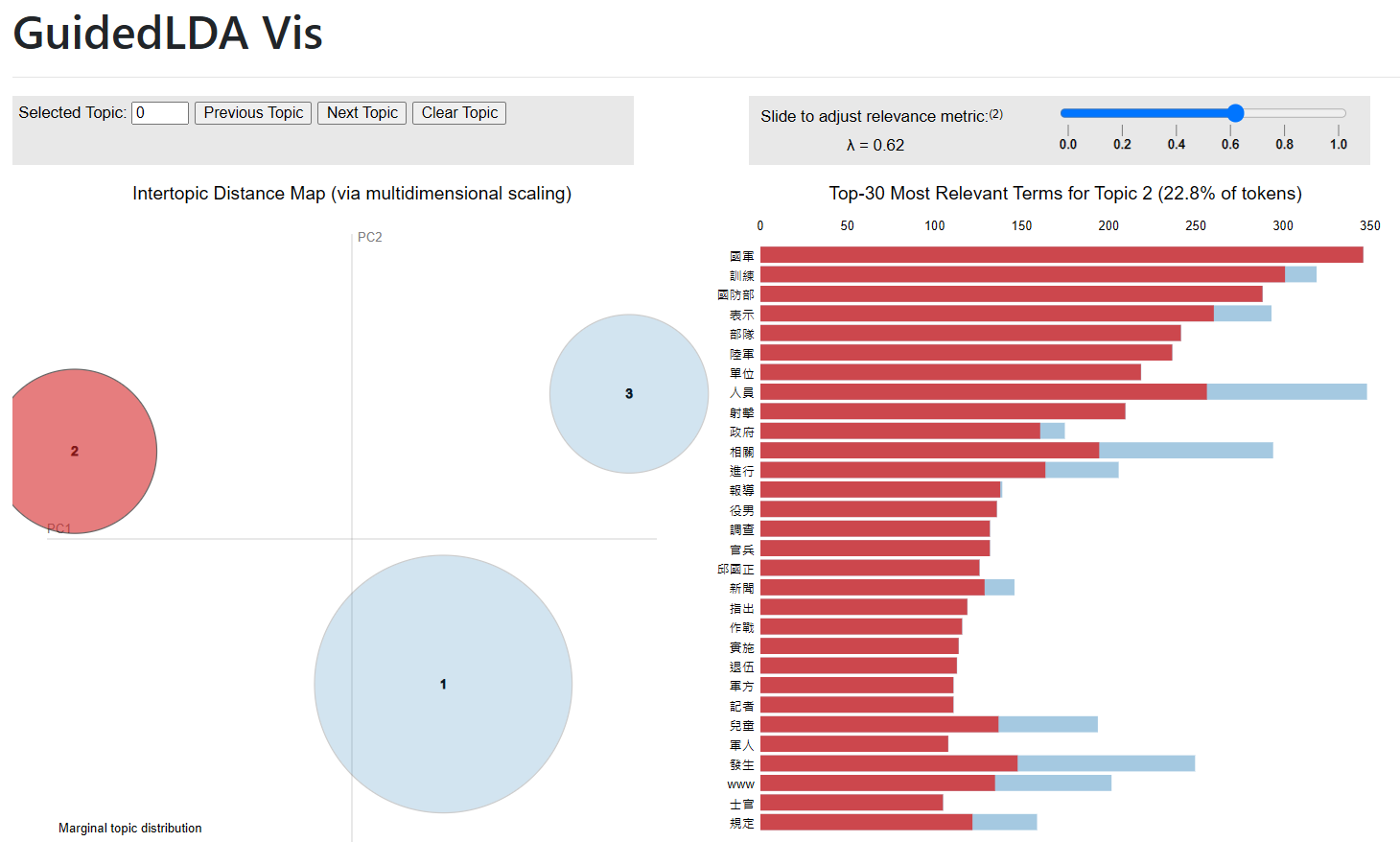






當主題數:3、詞彙頻率下限:20、詞彙頻率上限0.8:相同的設定，在LDA模型中是表現好的，在GuidedLDA主題模型軍旅與美妝的部分則會更為精確；上面軍旅版出現孩子與幼稚園的問題也沒有了。



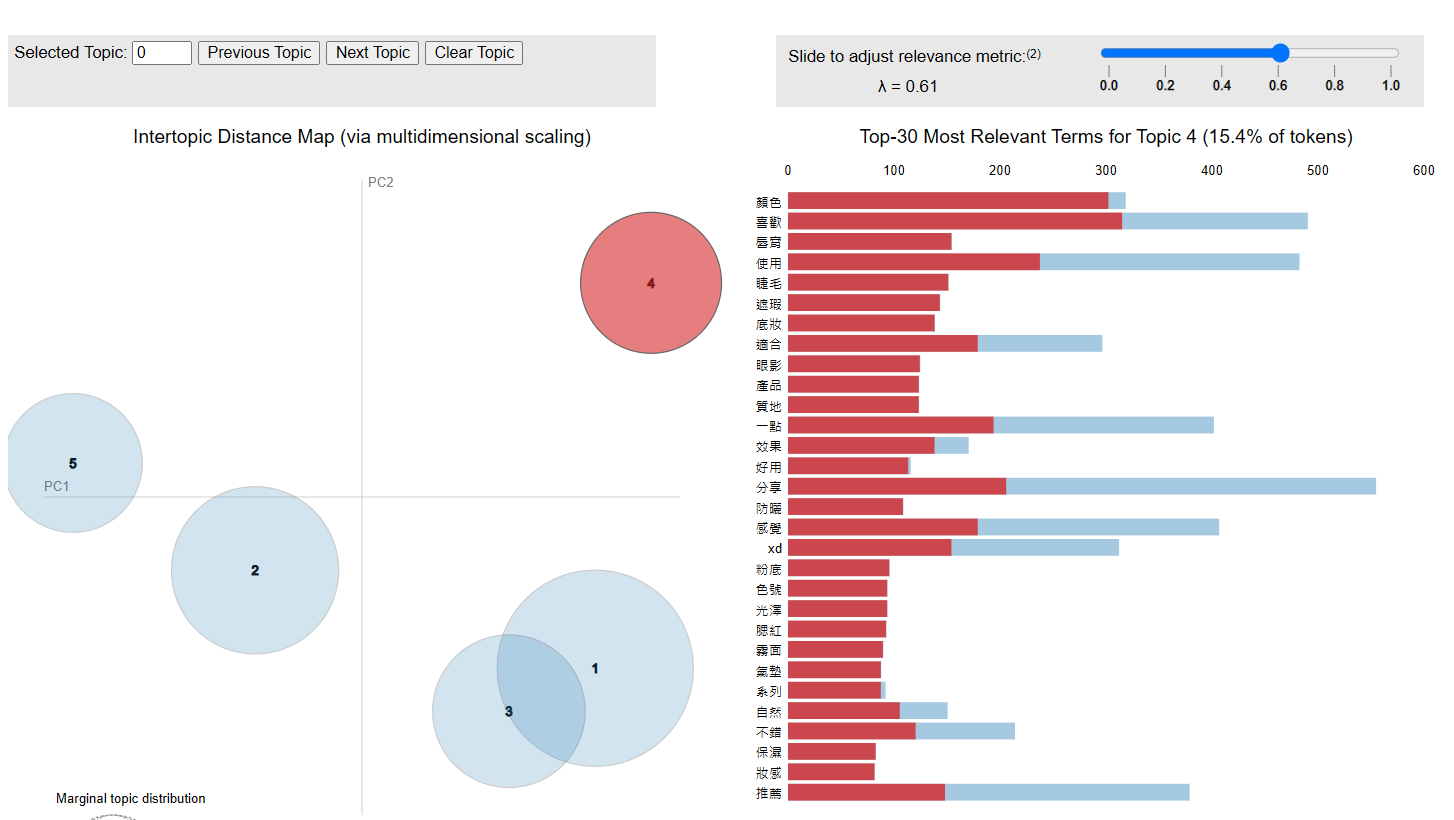


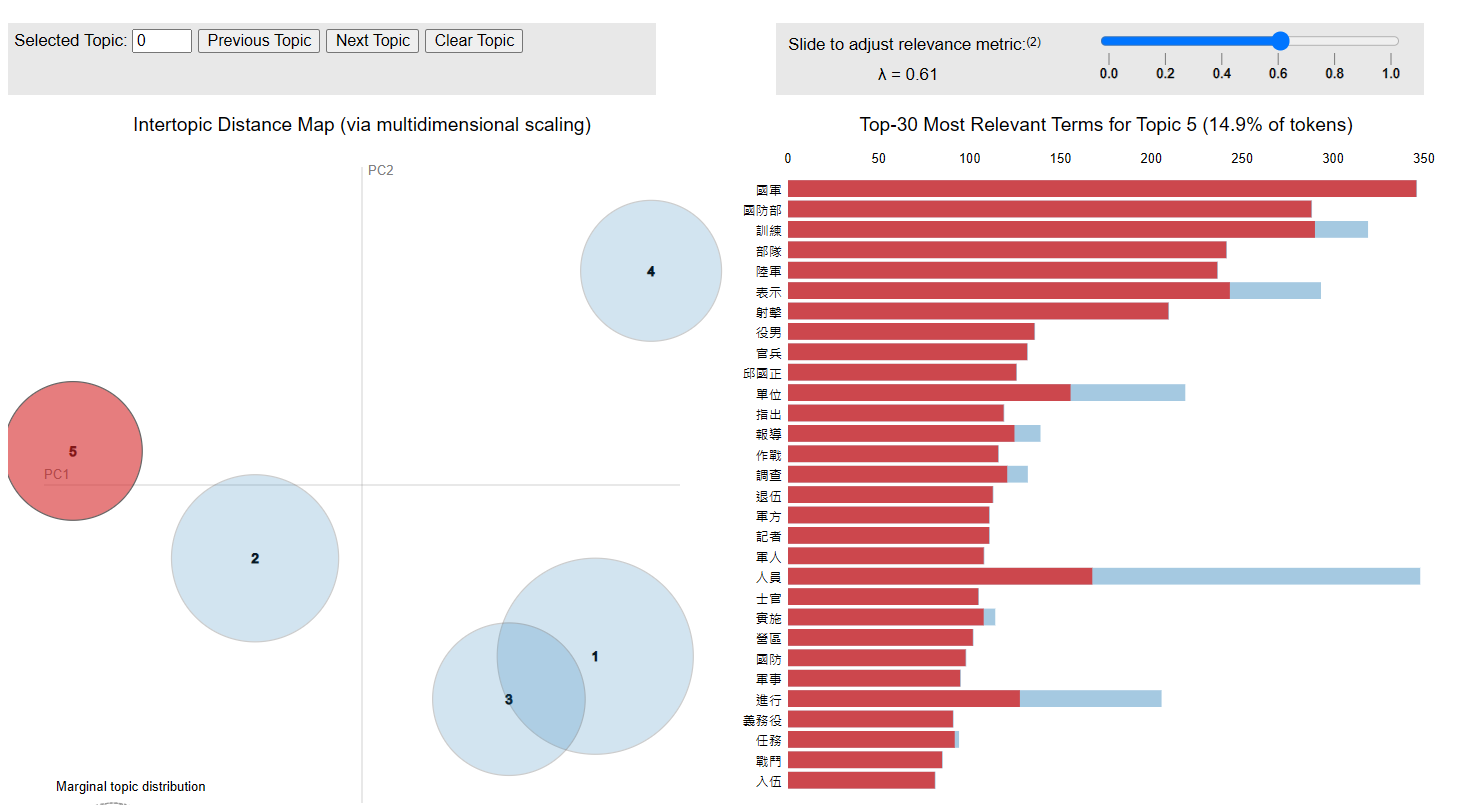
接著我們把主題數增加到5個，藉此觀察會有哪個主題被拆分，會被分成怎麼樣的主題，

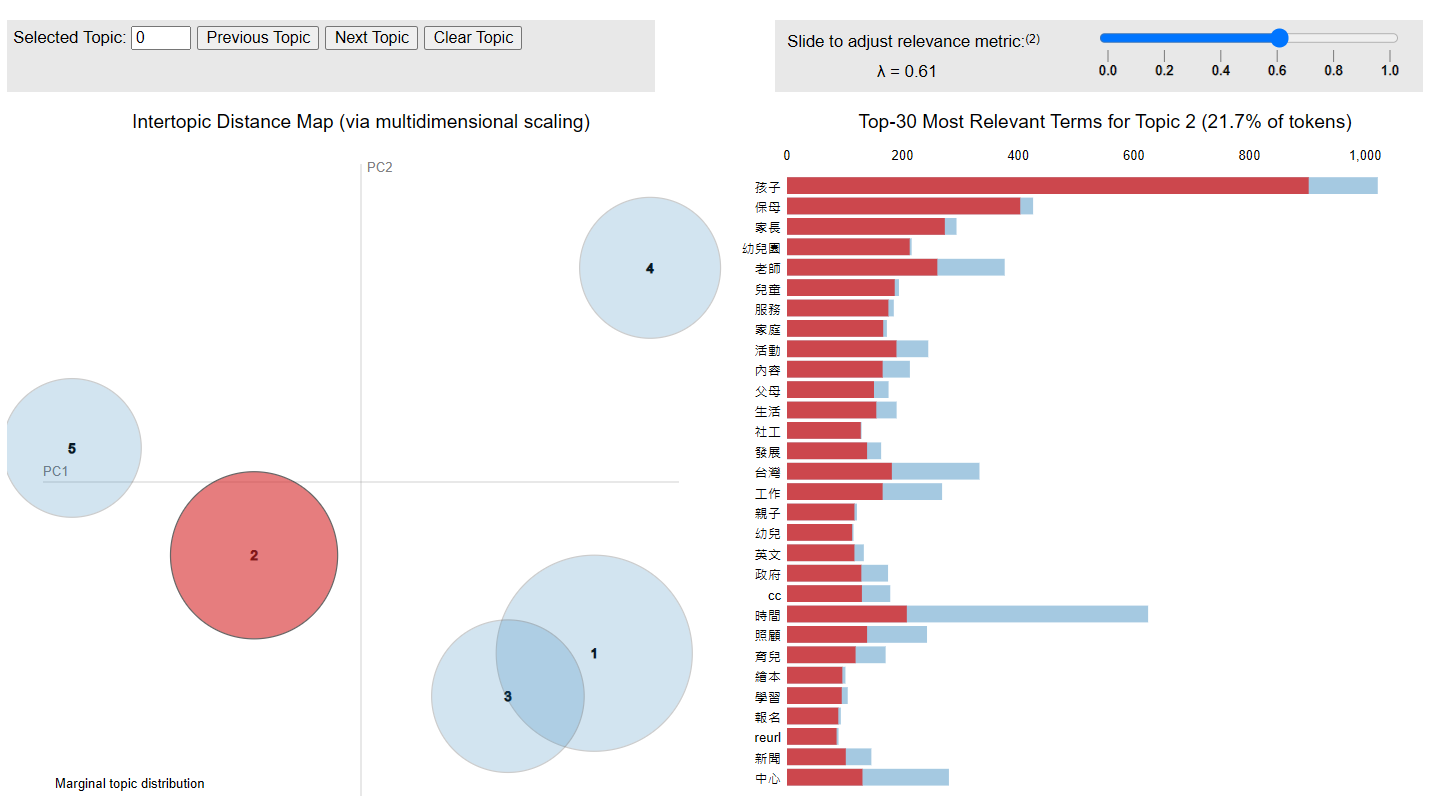
除了原本的軍旅及美妝之外，系統則將媽寶再分割為兒童教育、醫療、日常生活三塊，

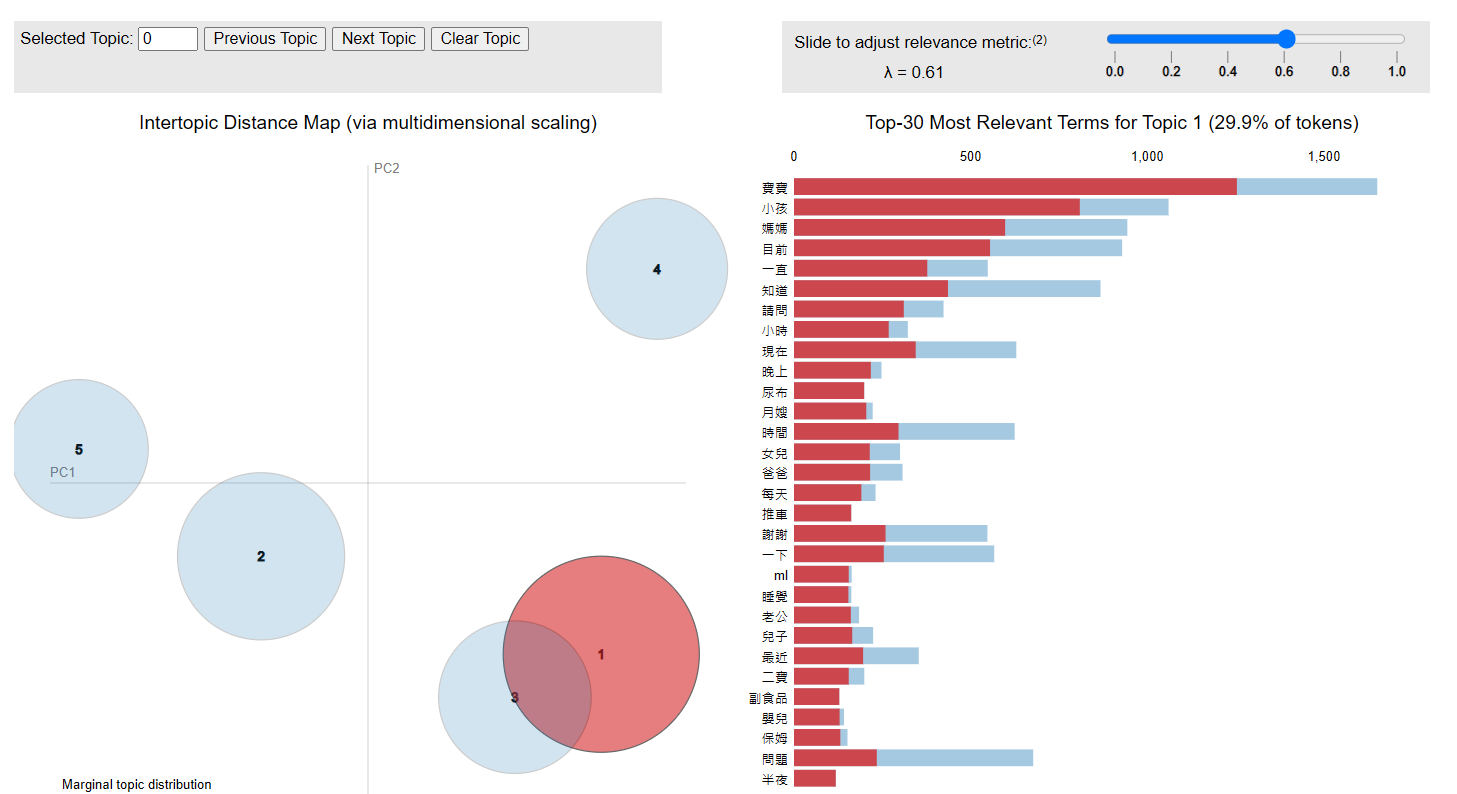
其中醫療、日常生活這兩塊有蠻大的重合。

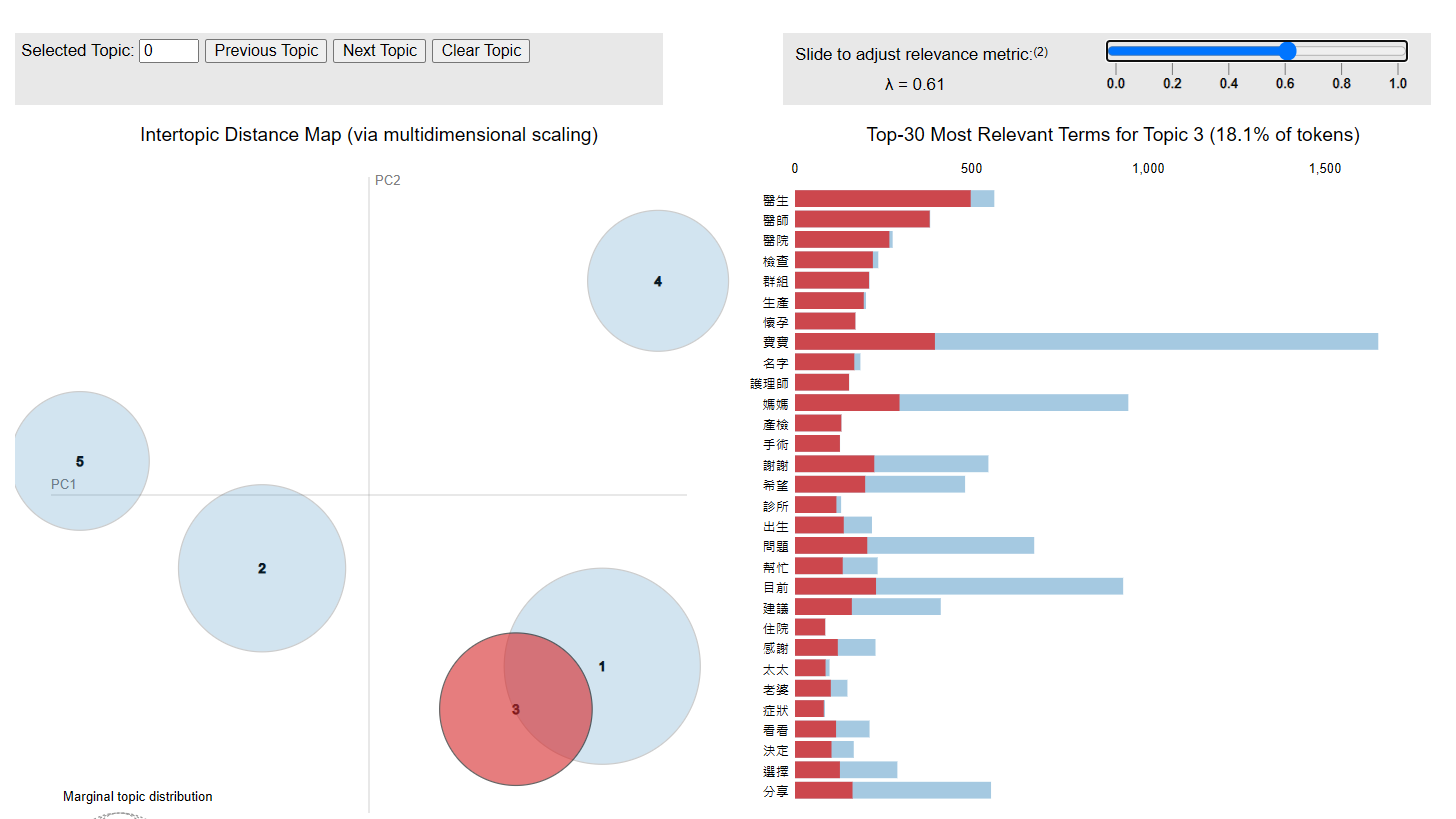






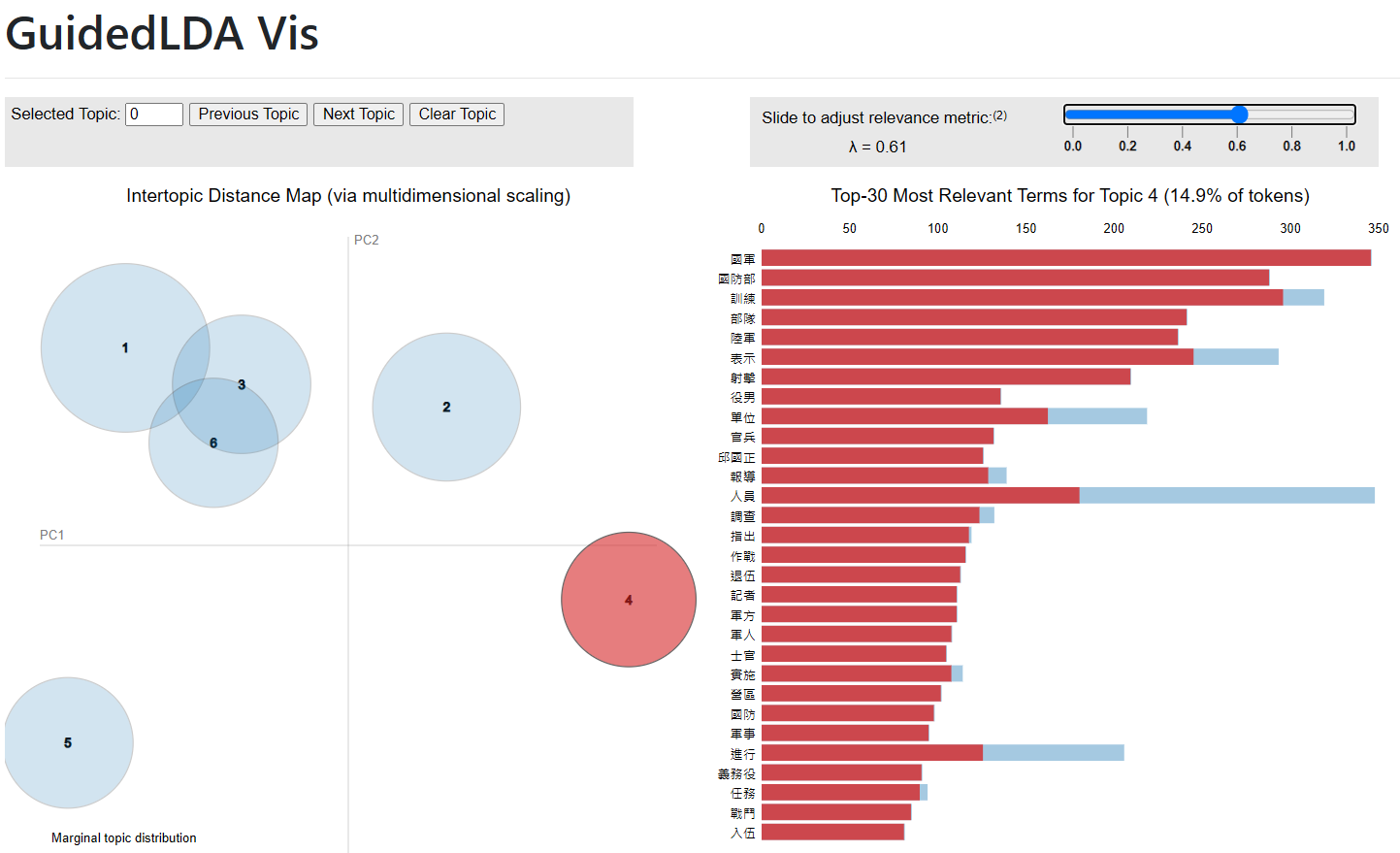


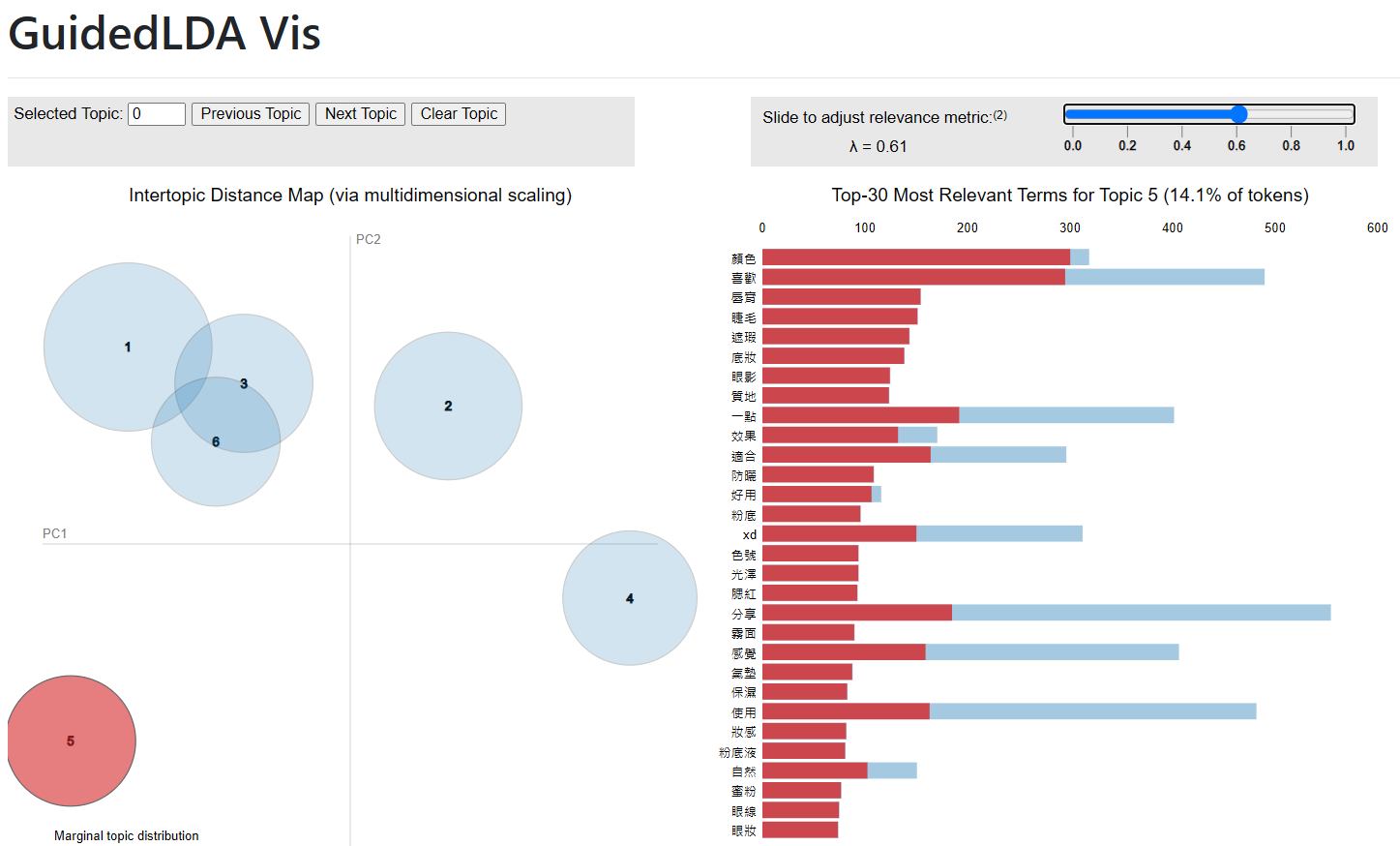


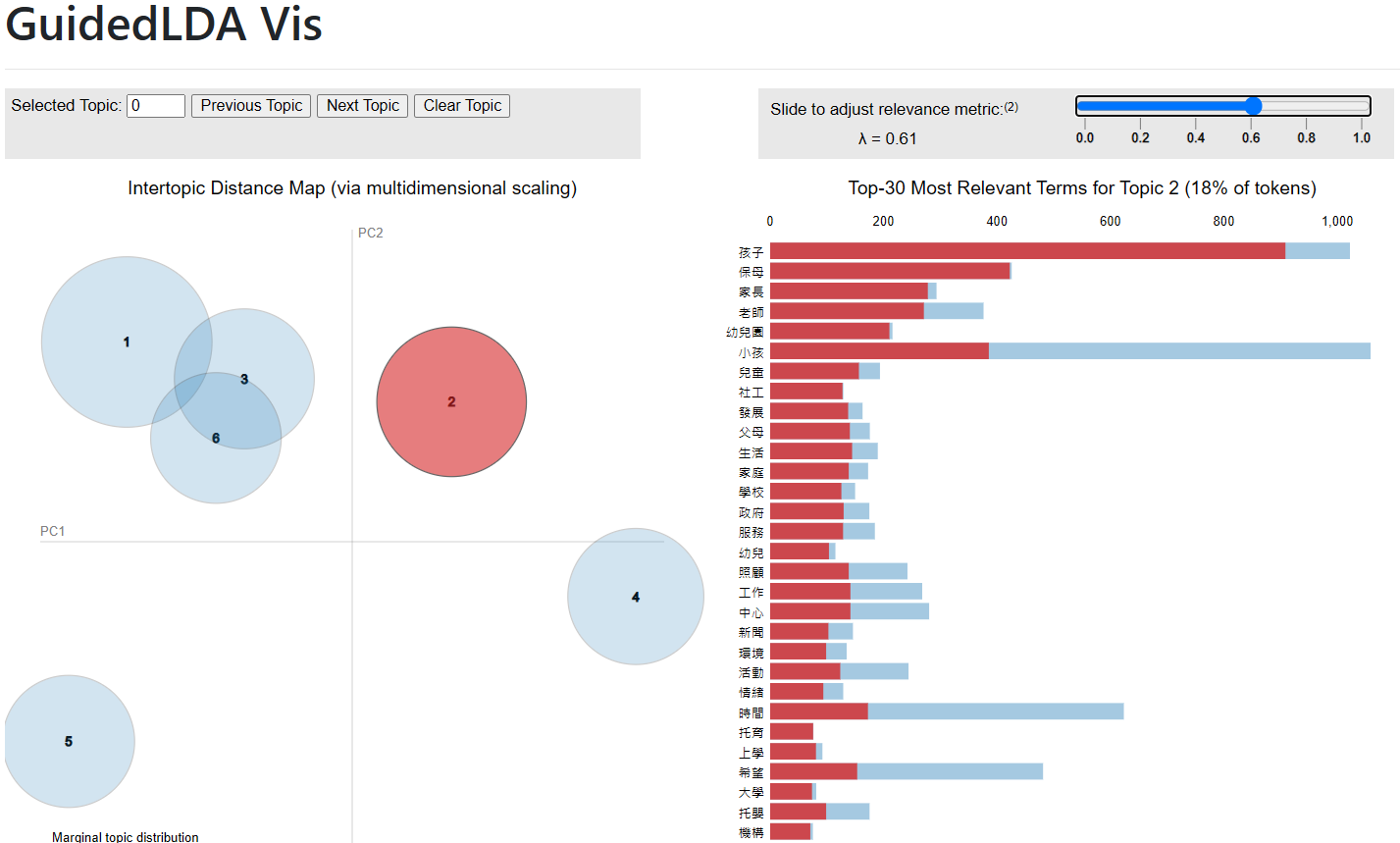


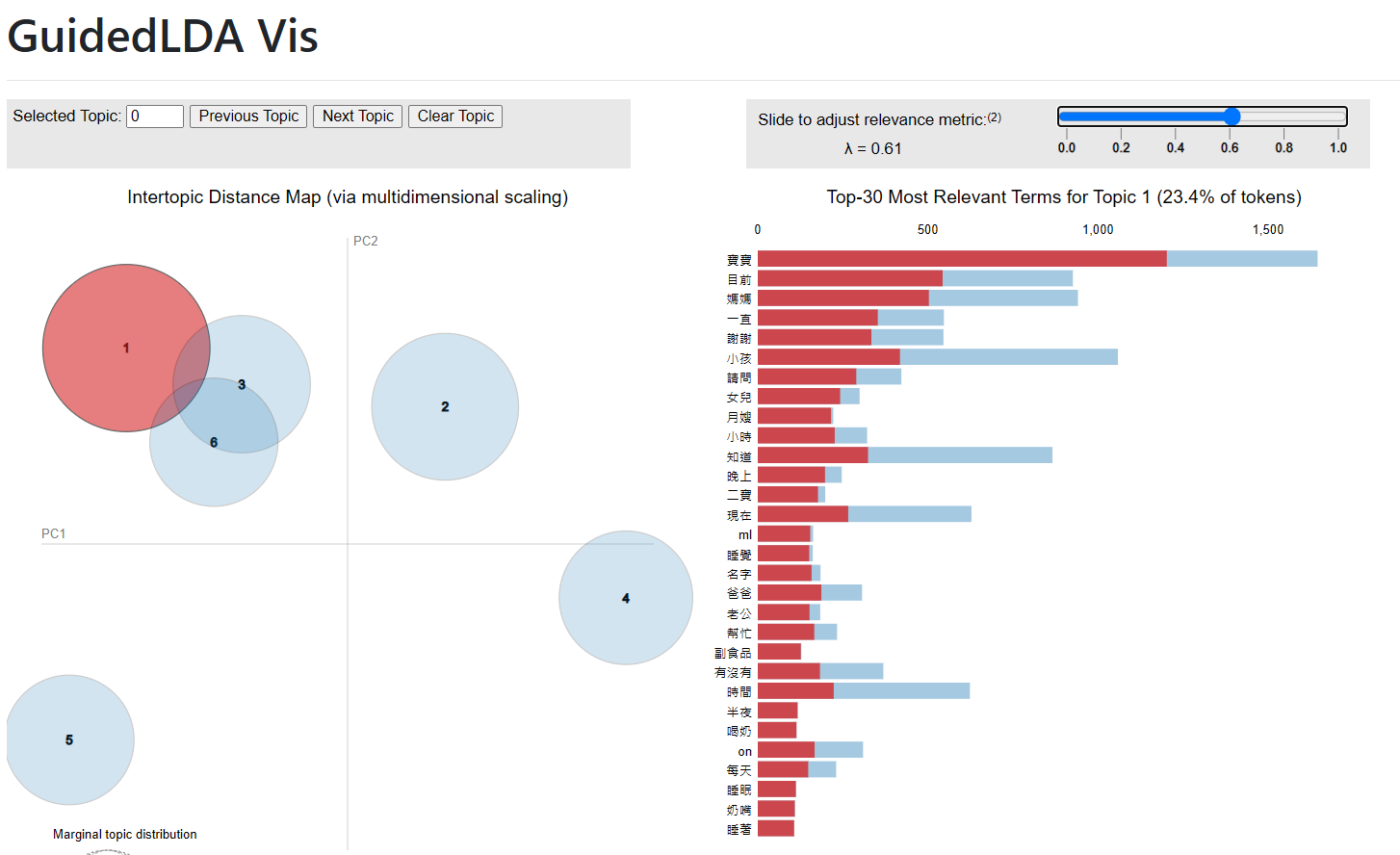
我們這次增加五種主題種子字(兒童教育、醫療、生活日常、軍旅、美妝)，並且把主題數增加為6個，想藉此觀察系統是否能正確的使用我們設定的主題種子字，以及另外一個主題會是甚麼樣的方向。

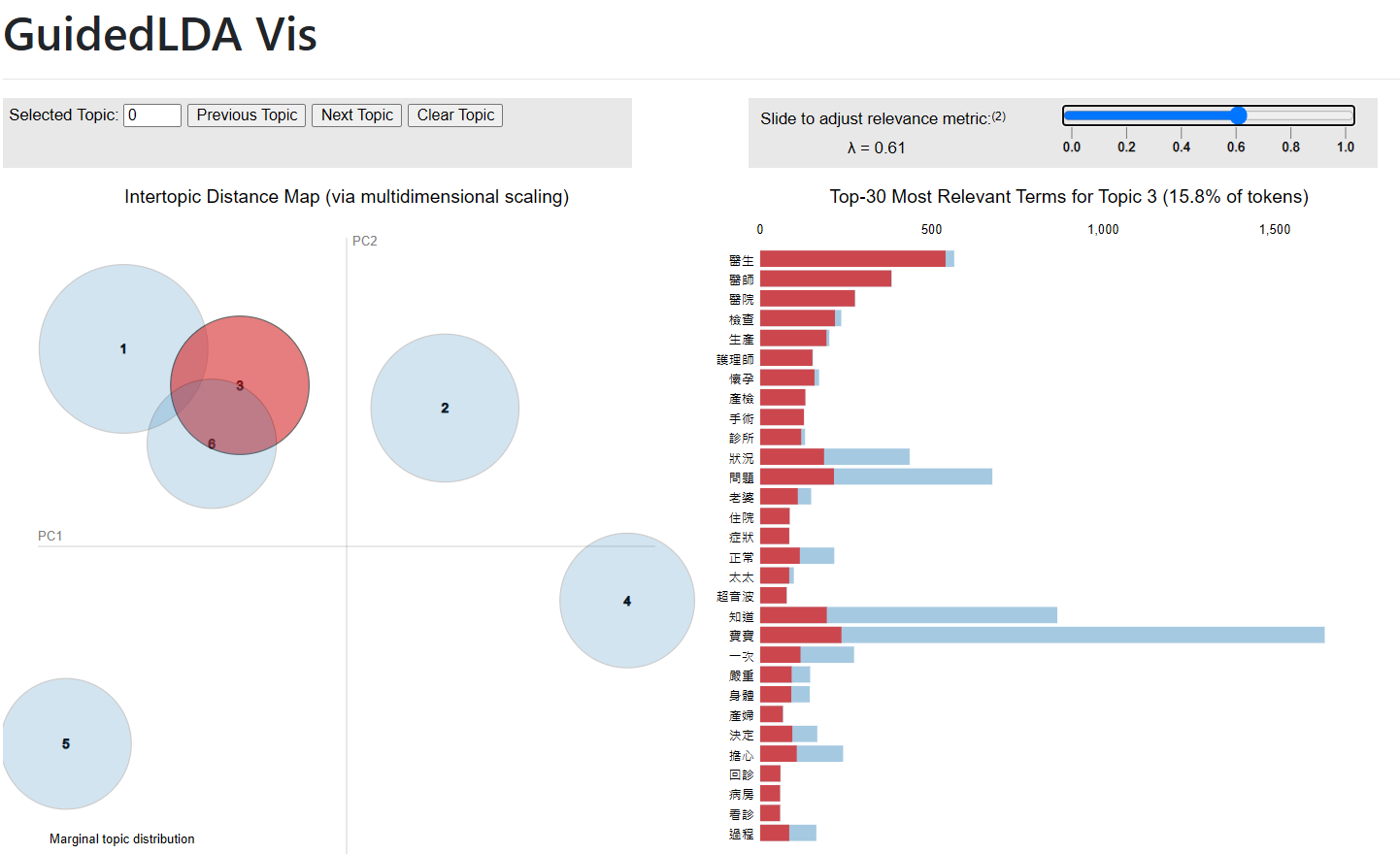


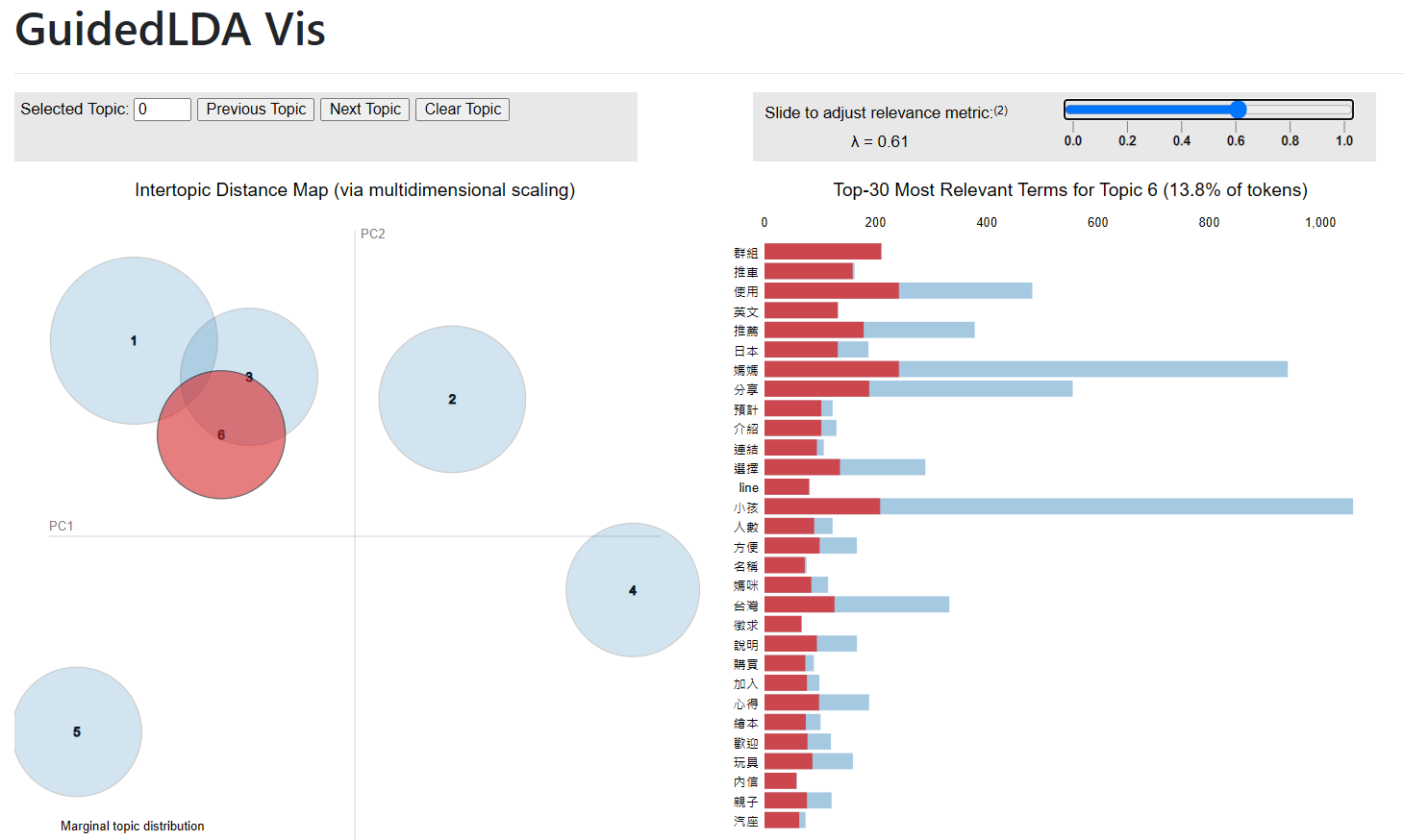












1. **結語**

原本設定的五個主題都有被正確的區分出來，剩下的主題我們可以由它的關鍵字:群組、分享、推薦、介紹、連結、推車、座椅..等將它分類為嬰幼兒團購。

我們可以發現原本從媽寶主題分出來的兒童教育、醫療、生活日常、嬰幼兒團購位置比較相近，其中兒童教育又與其他三個沒有交集，代表他們沒有共用的詞彙可以較好區分，另外醫療、嬰幼兒團購兩個重合相當多，應為它們相關的詞彙比較多，如寶寶、推薦、介紹…等。