一張含有 文字, 圓形, 符號, 標誌 的圖片

自動產生的描述

**國立中山大學資訊管理系碩士班**

**社群媒體分析**

**第三次讀書會報告**

**第十組成員：**

**N094020030 陳詠琳**

**N114020004 蔡志強**

**N114020006 黃銘輝**

**N114020007 陳嘉忻**

**N1140200012 黃延平**

**11221828030 范瑞洋**

**11221828031 王上豪**

**M121020012 凃宥安**

**指導教授：黃三益 教授**

**助教：蔡易航、蔡睿澤、張宸瑜、呂育真**

**中華民國 113 年 05 月**

**目錄**

[一、分析議題說明 3](#_Toc165922104)

[二、工作流程設計 3](#_Toc165922105)

[三、爬蟲、分群彙總、資料清理與DTM彙整 4](#_Toc165922106)

[四、LDA主題模型 11](#_Toc165922107)

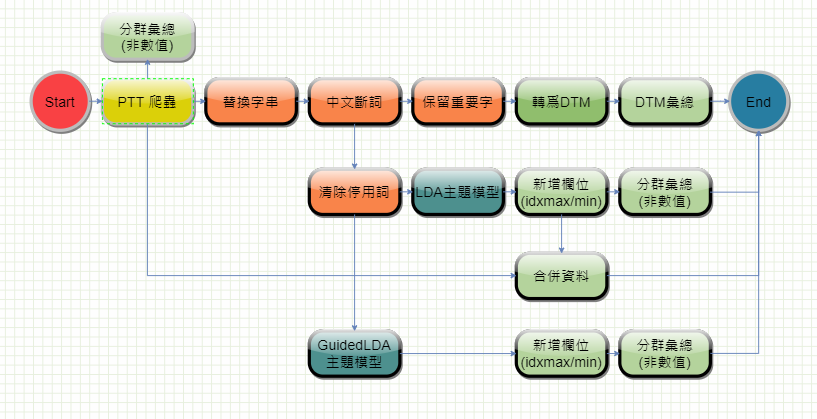
[五、GuideLDA主題模型 18](#_Toc165922108)

[六、結論 23](#_Toc165922109)

**一、分析議題說明**

* 主題：在2024年2月~4月PTT中，「**手機通訊**」、「**耳機**」、「**信用卡**」三個看板的文章分類與主題模型建立
* 使用平台：文字探勘工作流程設計平台

**二、工作流程設計**

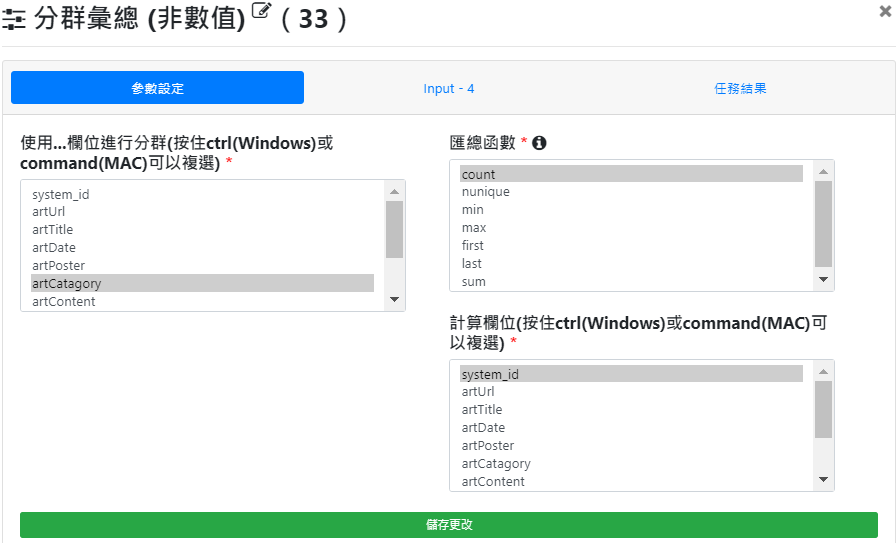
* 工作流程：
* 工作流程名稱：**讀書會3**
* 資料來源：PTT中「手機通訊」、「耳機」、「信用卡」等3個看板
* 分析期間：113/02/01 ~ 113/04/30
* 流程概述：
  1. 不指定特定關鍵字，爬取PTT今年2~4月手機通訊、耳機、信用卡等三個看板所有文章，共3,209筆資料。
  2. 以「替換字串」進行資料清理。
  3. 以「中文斷詞」將新聞內容分解成字詞單位。
  4. 使用「保留重要字」，透過自行建置的字典「0505-1.csv」，將三個看板的關鍵字留下，其餘去除。
  5. 將資料詞彙保留前200個，並轉為DTM。
  6. 以自行建置的字典「0505-1.csv」進行DTM彙整，以了解各篇文章可能的類別為何。
  7. 使用「清除停用詞」將不必要的符號、單字元去除。
  8. 使用「LDA主題模型」，並將主題數設定為3，
  9. 透過新增欄位，了解各篇文章可能分類為何。
  10. 計算在使用LDA主題模型後，各類文章數量。
  11. 使用「GuidedLDA主題模型」，將主題數設定為3，並給予個主題種子字。
  12. 透過新增欄位，了解各篇文章可能分類為何。
  13. 計算在使用GuidedLDA主題模型後，各類文章數量。

**三、爬蟲、****分群彙總、資料清理與DTM彙整**

1. PTT爬蟲

不設定關鍵字，爬取113/02/01 ~ 113/04/30期間，PTT中手機通訊、耳機、信用卡等3個看板之新聞資料，共3,209筆資料。

1. 分群彙總(非數值)

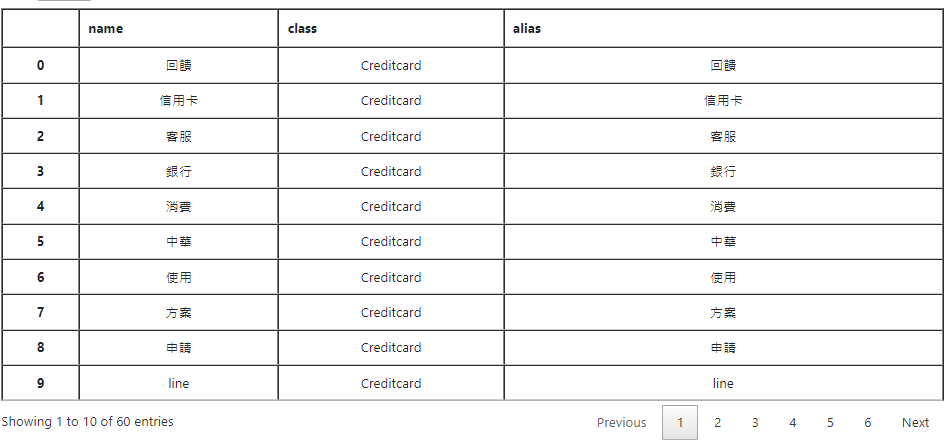
使用artCategory欄位進行分群匯總，彙總函數選擇count，計算欄位選擇system\_id，得出在資料期間內，信用卡版有566篇文章、耳機版有928篇文章、手機通訊版有1,715篇文章。 

1. 資料清理

* 替換字串
  1. \n>>，
  2. \n\n>>。
  3. Sent from JPTT on my .\*>>
  4. https:\/\/[a-zA-Z0-9-\.]+(?::\d+)?(?:\/[^\s]\*)?>>
  5. http:\/\/[a-zA-Z0-9-\.]+(?::\d+)?(?:\/[^\s]\*)?>>



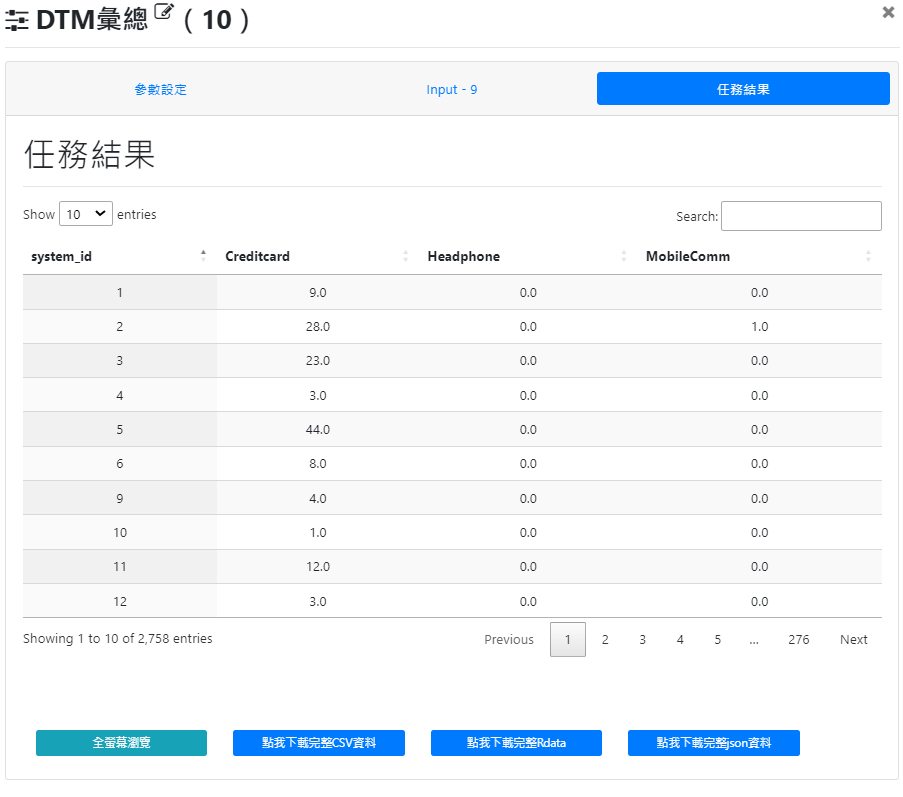
* 中文斷詞
* 保留重要字

透過自建的字典「0505-1.csv」，將三類文章關鍵字進行分類

* 轉為DTM

將篩選詞彙數量設定為200，並將上述結果轉為DTM

* DTM彙整

透過自建的字典「0505-1.csv」，將DTM以三類文章進行彙整，即可看到各篇文章中，三類文章關鍵字數量，可以以此判斷各篇文章可能所屬的文章分類為何，以第一篇文章為例，信用卡相關關鍵字出現9次、耳機及手機通訊關鍵字出現0次 ➔ 該篇文章應為信用卡版文章。 

**四、LDA主題模型**

* 清除停用字

將不必要的符號、單字元去除，如：分享、直接、最近、一下…。

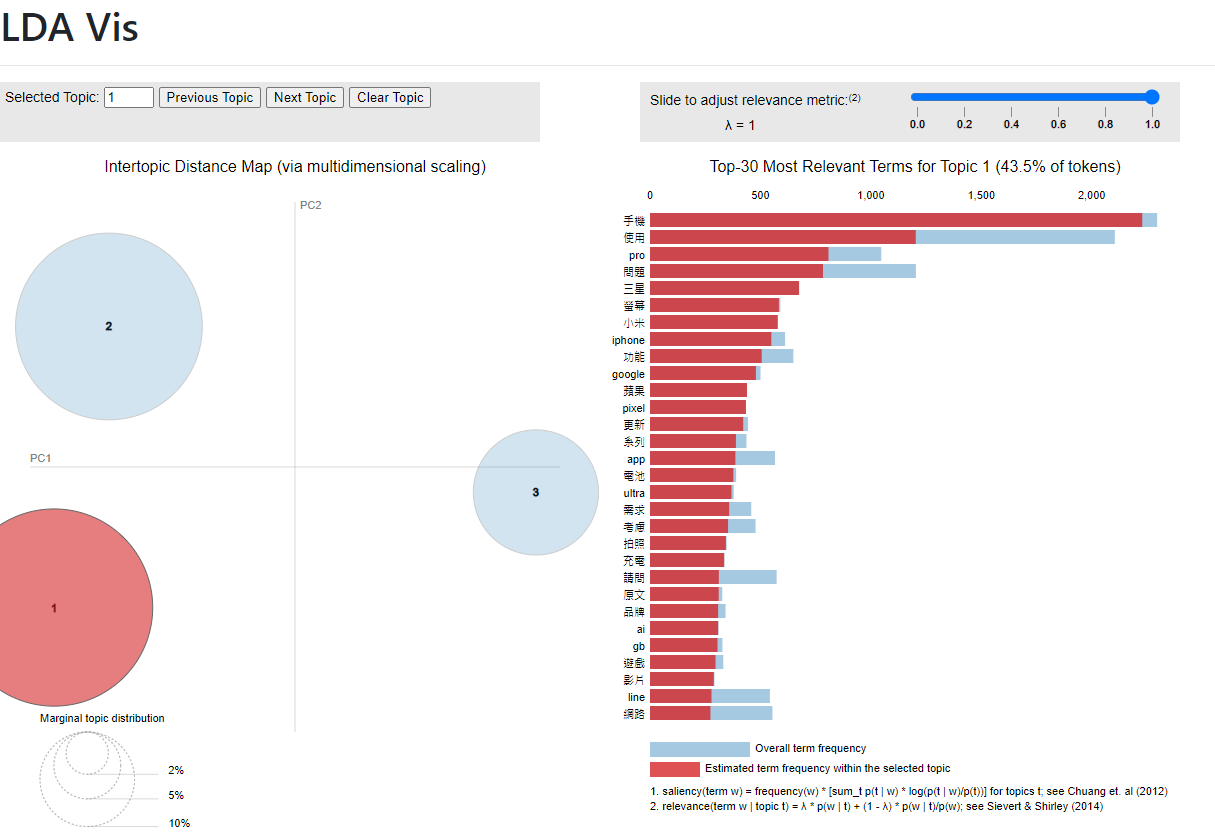


* LDA主題模型

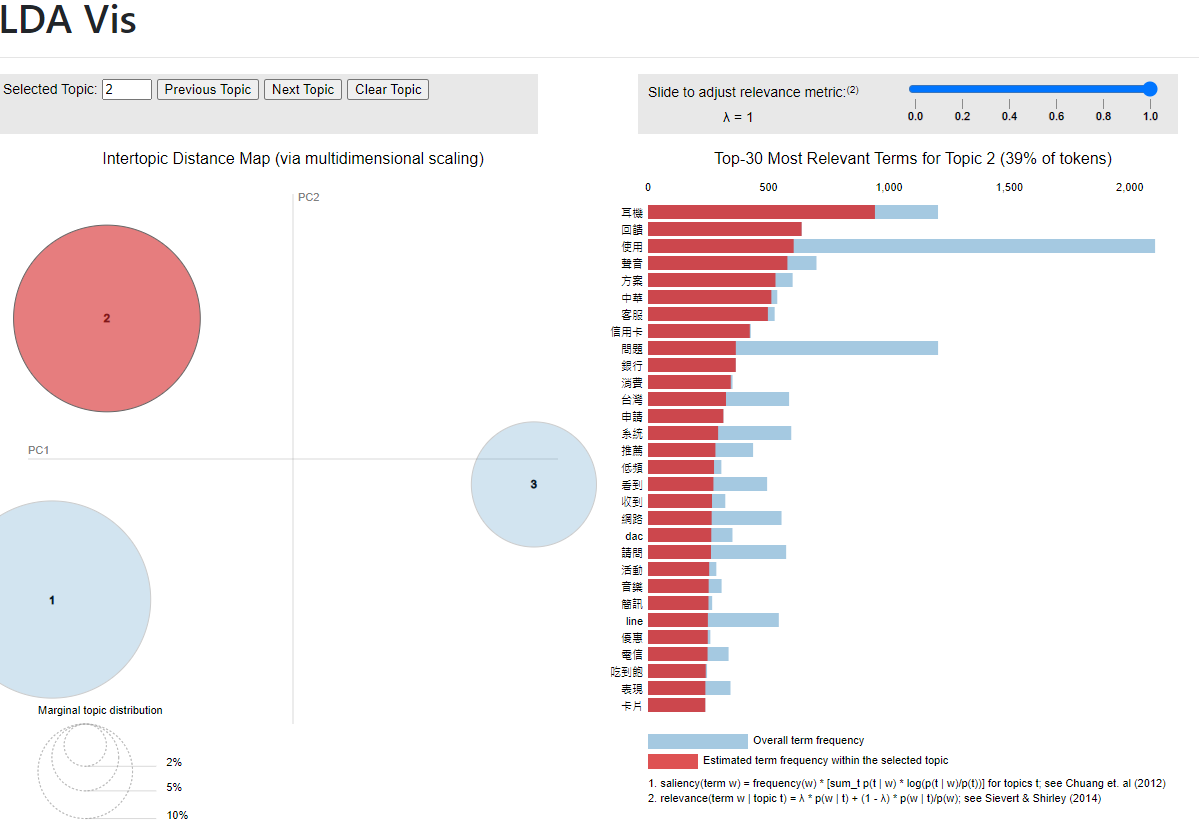
設定主題數 = 3，並透過迭代50次方式，產出結果



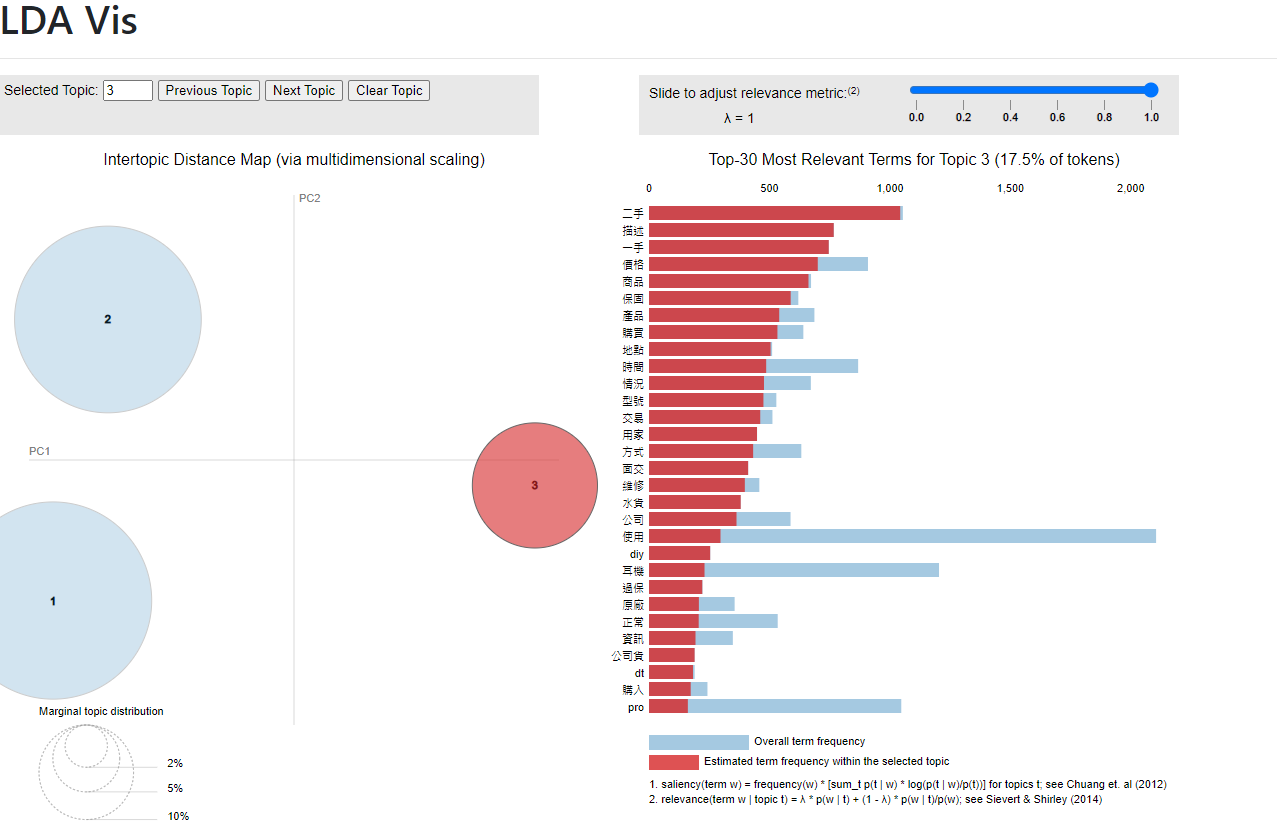
* 1. 首先看第1群，可以看到主要出現「手機」、「三星」、「小米」、「iphone」等關鍵字，與手機通訊較為相關之關鍵字。



* 1. 再來觀察第2群，可以看到主要出現「耳機」、「回饋」、「聲音」、「方案」、「信用卡」等關鍵字，與耳機或信用卡較為相關之關鍵字，此分類效果並不理想。



* 1. 最後，觀查第3群，可以看到主要出現「二手」、「商品」、「價格」、「保固」、「型號」、「耳機」等關鍵字，與耳機使用較為相關之關鍵字。



* 新增欄位(idxmax/min)

透過新增欄位，以max作為彙整方式，了解各篇文章可能分類為何。



* 分群彙整(非數值)

計算在使用LDA主題模型後，各類文章數量。

可以觀察到各群文章數量如下：

1. 第１群(手機通訊)：493篇
2. 第２群(耳機或信用卡)：1,279篇
3. 第３群(耳機)：1,437篇

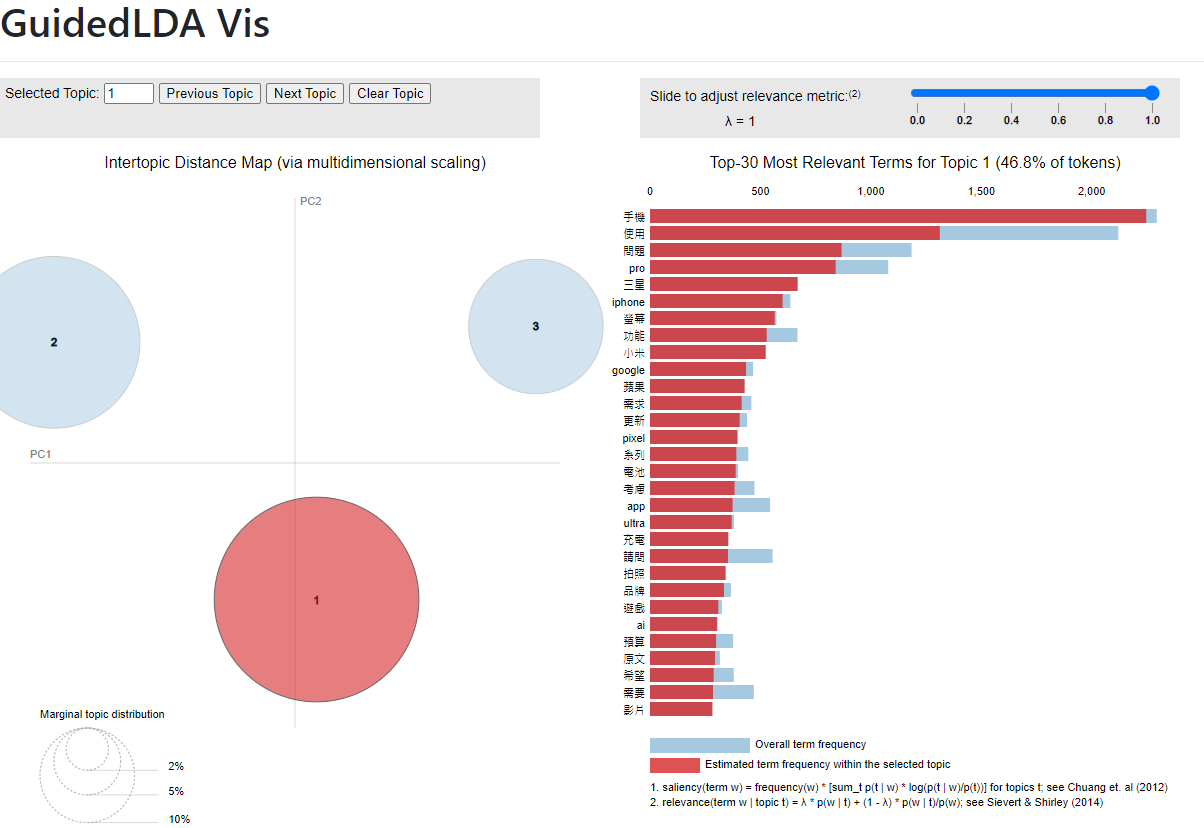


**五、GuideLDA主題模型**

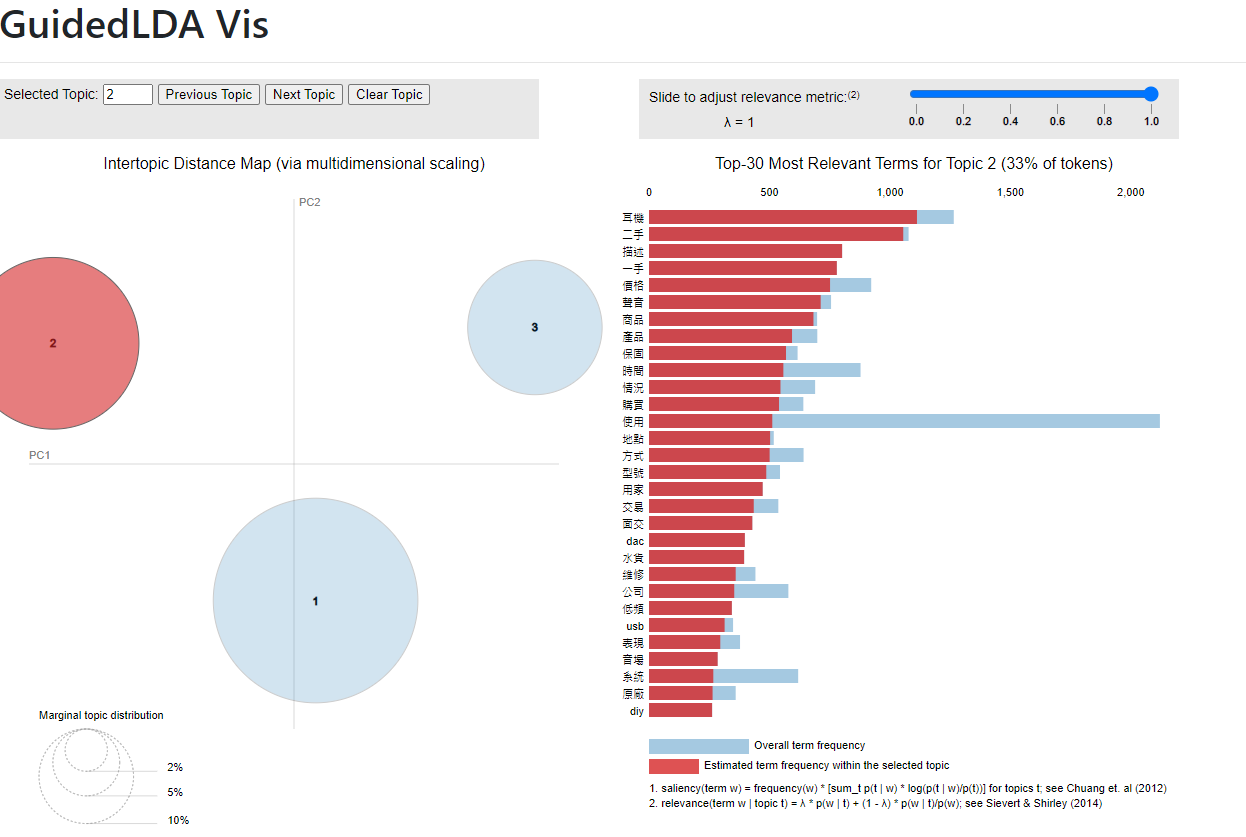
* GuideLDA主題模型

設定主題數 = 3，給予主題關鍵字「信用卡,回饋」、「耳機,耳罩,入耳」、「手機,電池,Apple,中華」，透過迭代50次方式，產出結果

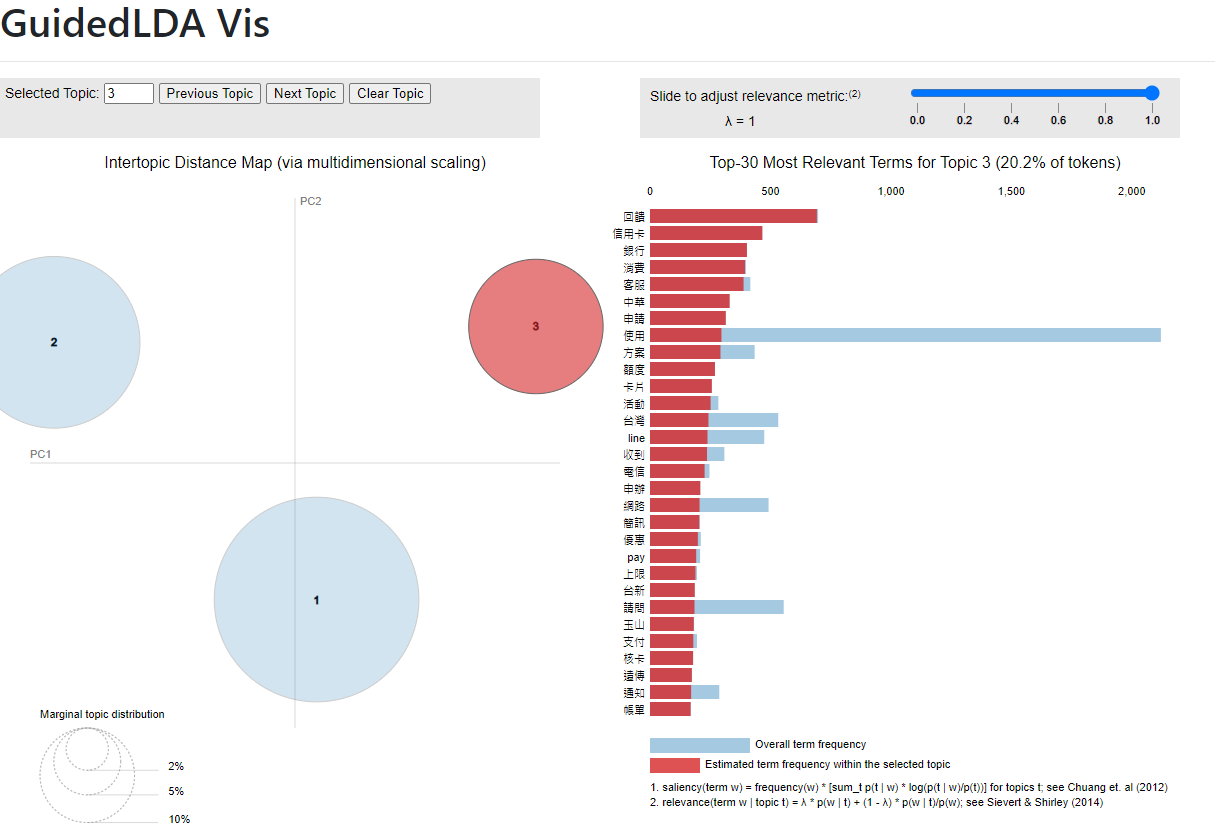
1. 首先看第1群，可以看到主要出現「手機」、「三星」、「使用」、「iphone」等關鍵字，與手機通訊較為相關之關鍵字。



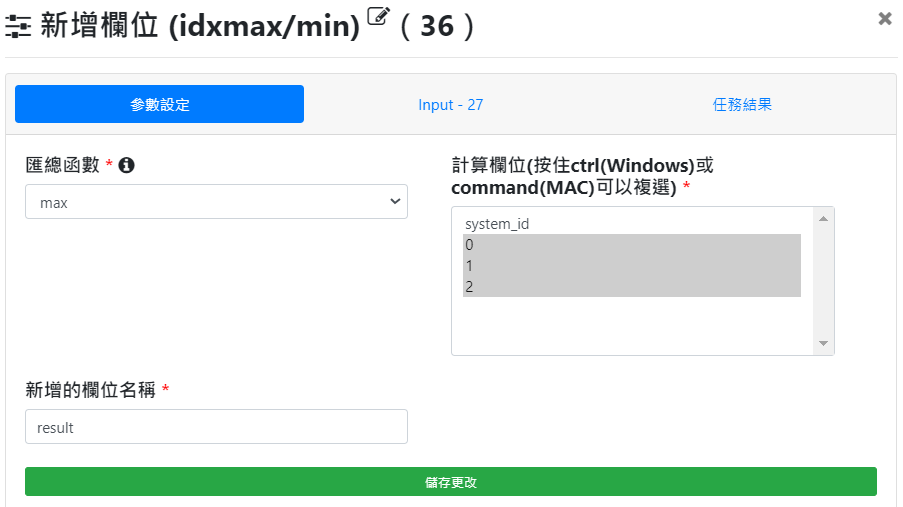
1. 再觀察第２群，可以看到主要出現「耳機」、「二手」、「保固」等關鍵字，與耳機較為相關之關鍵字。



1. 最後觀察第３群，可以看到主要出現「回饋」、「信用卡」、「銀行」等關鍵字，與信用卡較為相關之關鍵字。



* 新增欄位(idxmax/min)

透過新增欄位，以max作為彙整方式，了解各篇文章可能分類為何。 

* 分群彙整(非數值)

計算在使用GuidedLDA主題模型後，各類文章數量。

可以觀察到各群文章數量如下：

1. 第１群(手機通訊)：923篇
2. 第２群(耳機)：765篇
3. 第３群(信用卡)：1,521篇



# 六、結論

此次以PTT信用卡、耳機、手機通訊版為資料來源，使用不同方式來判斷文章類別，可以發現在使用LDA主題模型時，在三個類別上的分類效果並不如預期，而在改用GuideLDA主題模型後，因為給定相關關鍵字，針對目標文章的分類效果較貼近原始的三個分類。將結果比對原始文章的類別後發現，使用LDA及GuideLDA訓練後的模型在文章分類判斷上仍有一定誤差，可能原因在於關鍵字給予的不夠精準且數量不夠，以及在前處理的斷詞中，並未將一些有助於分類的關鍵字定義出來，導致判斷上沒有那麼準確。因本次tarflow組別得知需要再製作讀書會作業之時間較為緊迫，後續若有機會將再進行調整，以精進分析內容。