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 冷宥安

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

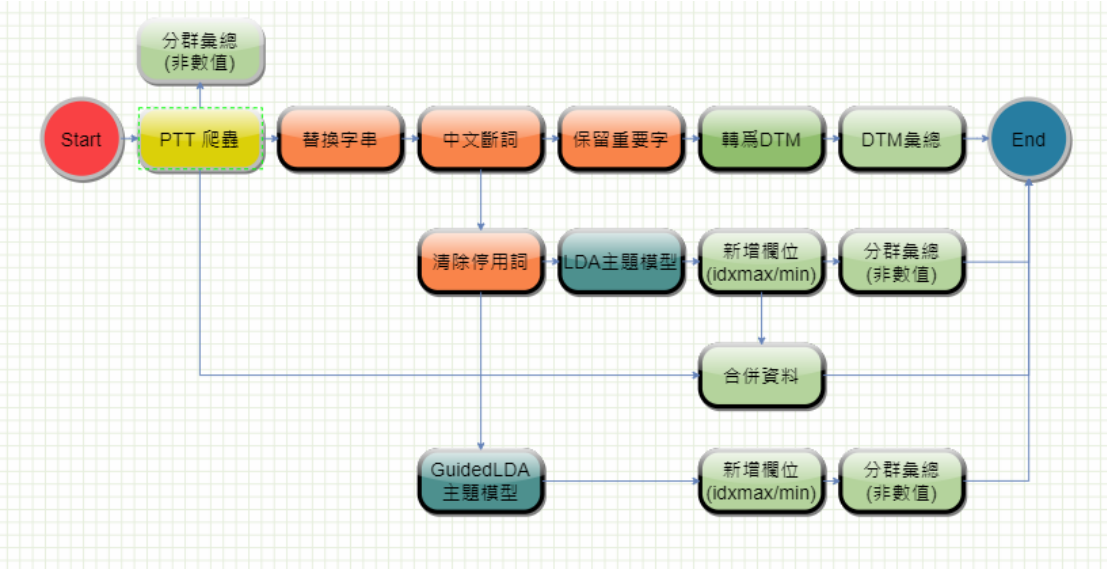
目錄

一、分析議題說明

- 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：

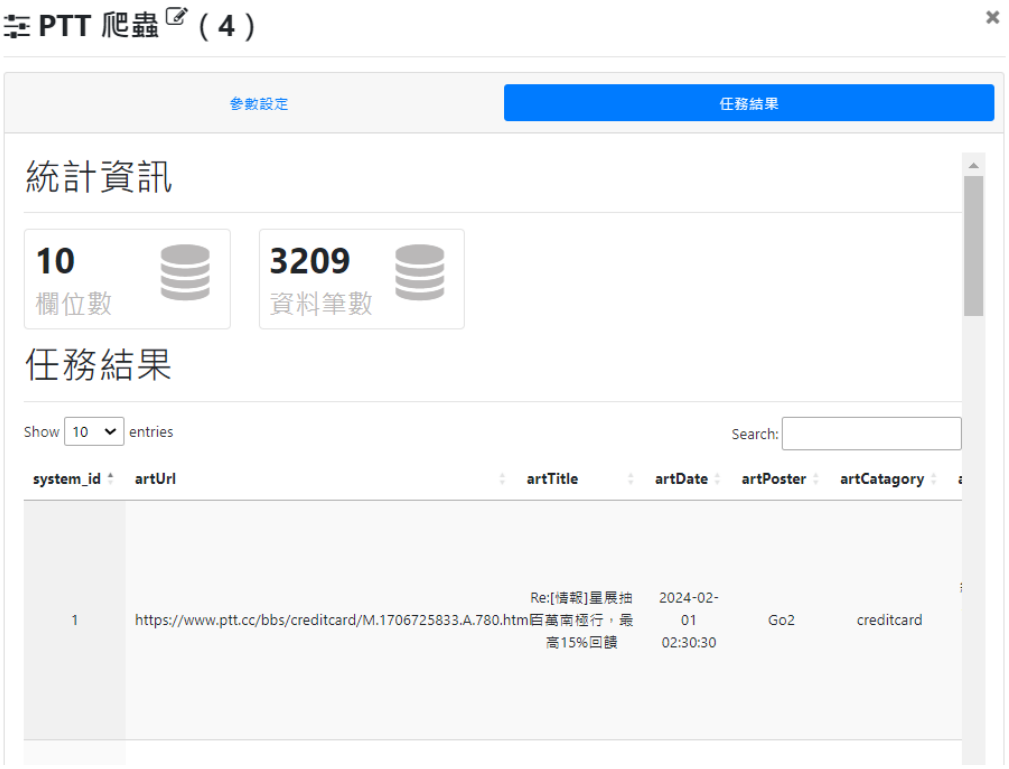
1. 不指定特定關鍵字，爬取 PTT 今年 2~4 月手機通訊、耳機、信用卡等三個看板所有文章，共 3,209 筆資料。
2. 以「替換字串」進行資料清理。
3. 以「中文斷詞」將新聞內容分解成字詞單位。

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。



2. 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。

3. 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


中文斷詞（7）
參數設定
Input - 6
任務結果
選擇處理欄位 *
result
定義詞彙
以換行符號區隔，e.g.
詞彙 權重
國立中山大學 1000
西子灣 500
...
選取字典
----------請選擇----------
儲存更改


中文斷詞（7）
參數設定
Input - 6
任務結果
統計資訊
14713
最大字數
677898
總字數
0
最小字數
211
平均字數
任務結果
Show 10 entries
Search:
system_id
result
1
[結果, 又, 有, 新, 檔期, 活動, 了, 但, 改爛, 很多, 主打, 海外, 實體, 且, 用, 消費, 順序, 比較, ，, ，, 避免, card, +, 又, 掛掉, (, 他們, 其實, 有, 獨立, 的, 動感, 活力, 網, 不用, 硬要, 用, 網銀, XD,), ，, ，, 活動, 網址, :, ，, ~, 2024, /, 5, /, 31, ，, ，, 登錄, 期間, :, 2024, /, 5, /, 27, 10, :, 00, ~, 2024, /, 5, /, 31, 23, :, 59, ，, ，, (, 登錄, 期間, 登錄, 即可, ，, 不用, 搶, 登錄, ！, 回饋, 以, 消費, 達標, 時間, 排序, ，, 盡快, 刷, 回饋, 等, 您, 拿, ！, ...]
2
[OPENPOINT, 回饋, ，, 每人, 每月, 回饋, 上限, 70, 點, 。, ，, (, 每月, 總, 回饋, 70, 萬點,), ，, ，, 富邦, 銀行, ，, *, 7, -, ELEVEN, 實體, 門市, ，, *, 統一, 集團, 指定, 通路, ，, 單筆, 消費, 滿, 200, 元享, 10%, OPENPOINT, 回饋, ，, 每人, 每月, 上限, 80, 點, 。, (, 每月, 總, 回饋, 45, 萬點,), ，, ，, 中信, 銀行, ，, *, 7, -, ELEVEN, 實體, 門市, ，, *, 統一, 集團, 指定, 通路, ，, 單筆, 消費, 滿, 200, 元享, 10%, OPENPOINT, 回饋, ，, 每人, 每月, 上限, 100, 點, 。, (, 每月, 總, 回饋, 14.5, 萬點,), ，, ，, ...]
3
[★, 職業, 類別, :, 工程師, ，, ，, ★, 年資, :, 3, 年, ，, ，, ★, 年齡, :, 28, ，, ，, ★, 年, 收入, :, 85, 萬, ，, ，, ★, 提供, 之, 財力, 證明, :, 6, 個, 月薪, 轉, +, 扣繳, 憑單, ，, ，, ★, 最近, 三個, 月, 有無, 申辦, 卡片, :, 華南, i, 網購, ，, ，, ★, 申辦, 過程, :, ，, ，, 近期, 因為, i, 網購, 太神, 有, 先辦, 了, i, 網購, ，, 另外, 因為, 有些, 銀行, 額度,
全螢幕瀏覽
點我下載完整CSV資料
點我下載完整Rdata
點我下載完整json資料

- 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

	name	class	alias
0	回饋	Creditcard	回饋
1	信用卡	Creditcard	信用卡
2	客服	Creditcard	客服
3	銀行	Creditcard	銀行
4	消費	Creditcard	消費
5	中華	Creditcard	中華
6	使用	Creditcard	使用
7	方案	Creditcard	方案
8	申請	Creditcard	申請
9	line	Creditcard	line

Showing 1 to 10 of 60 entries

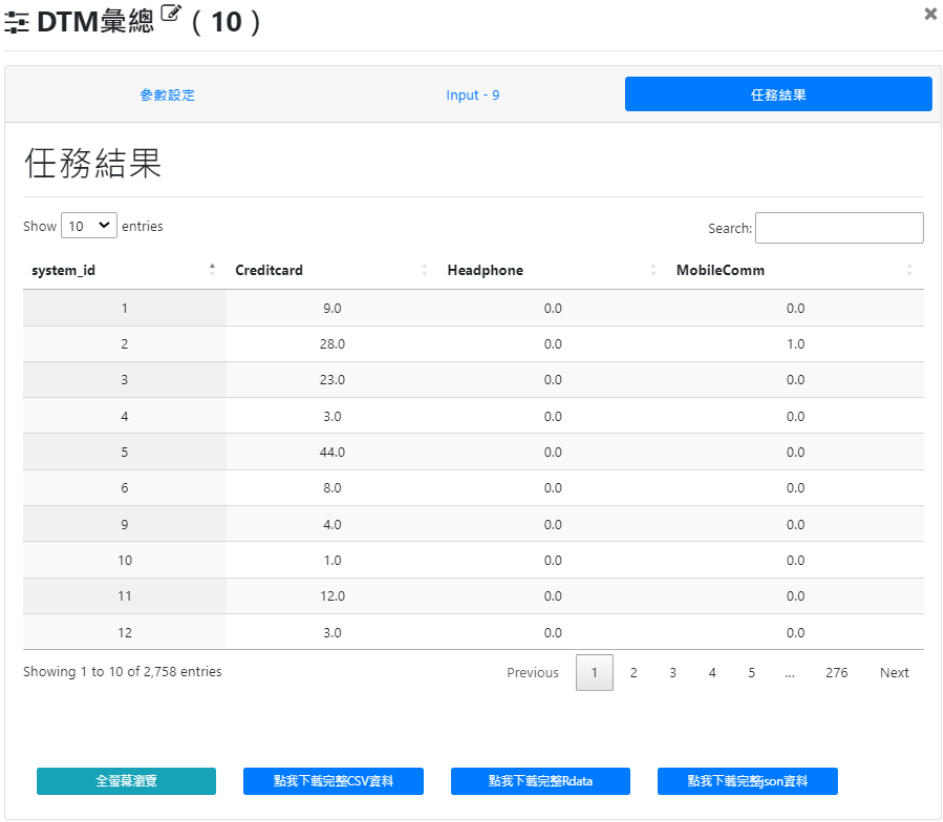
Previous 1 2 3 4 5 6 Next

- 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

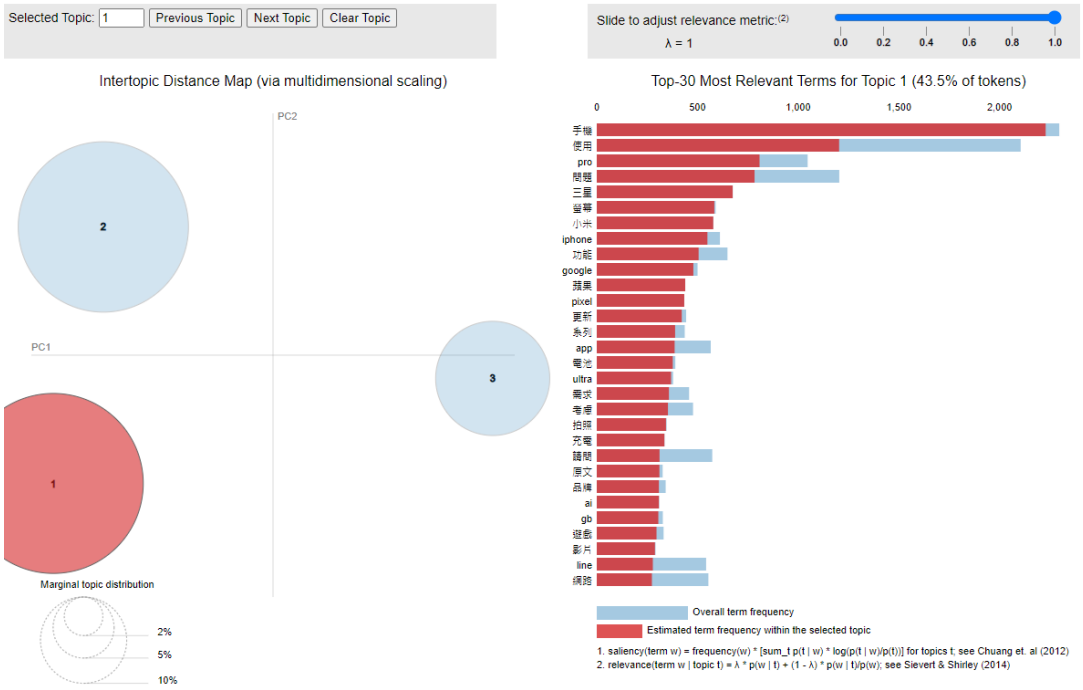
  將不必要的符號、單字元去除，如：分享、直接、最近、一下...。

- LDA 主題模型

  設定主題數 = 3，並透過迭代 50 次方式，產出結果

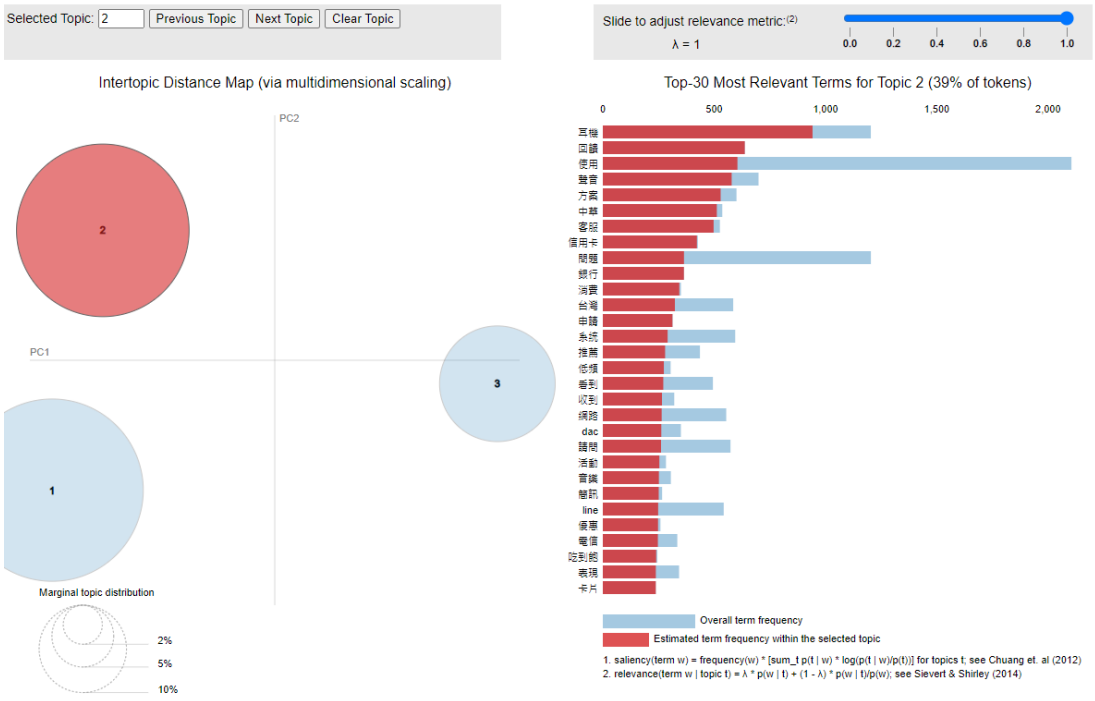
(6) 首先看第 1 群，可以看到主要出現「手機」、「三星」、「小米」、「iphone」等關鍵字，與手機通訊較為相關之關鍵字。

LDA Vis



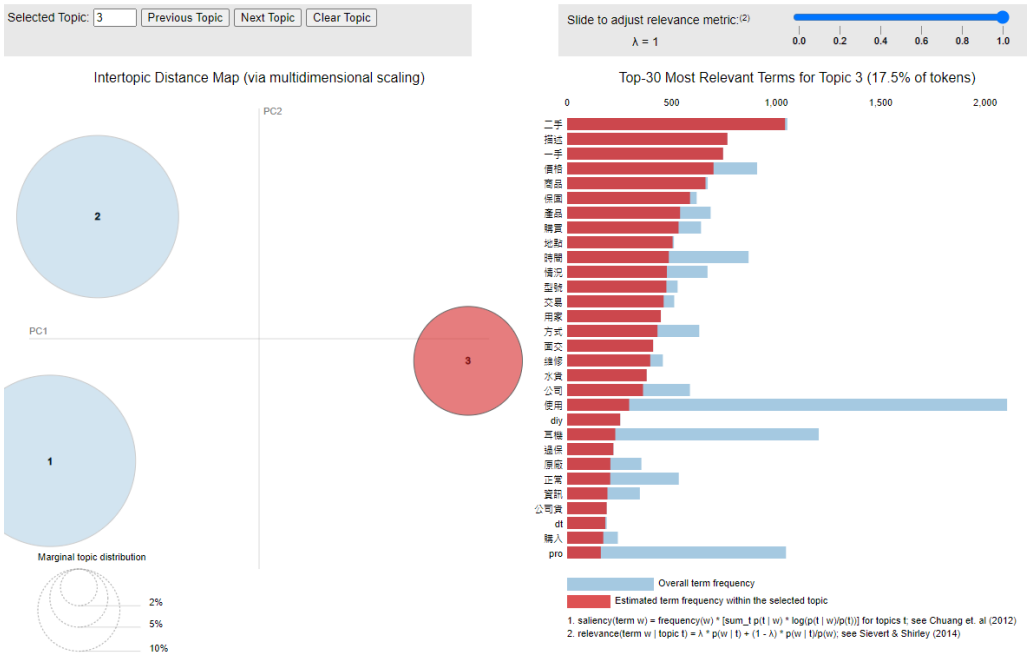
(7) 再來觀察第 2 群，可以看到主要出現「耳機」、「回饋」、「聲音」、「方案」、「信用卡」等關鍵字，與耳機或信用卡較為相關之關鍵字，此分類效果並不理想。

LDA Vis



(8) 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。

LDA Vis



- 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

分群彙總 (非數值) (20)

參數設定 | Input - 19 | 任務結果

統計資訊

3
群組數量

任務結果

Show 10 entries Search:

result	system_id@count
[]	493
[]	1279
[]	1437

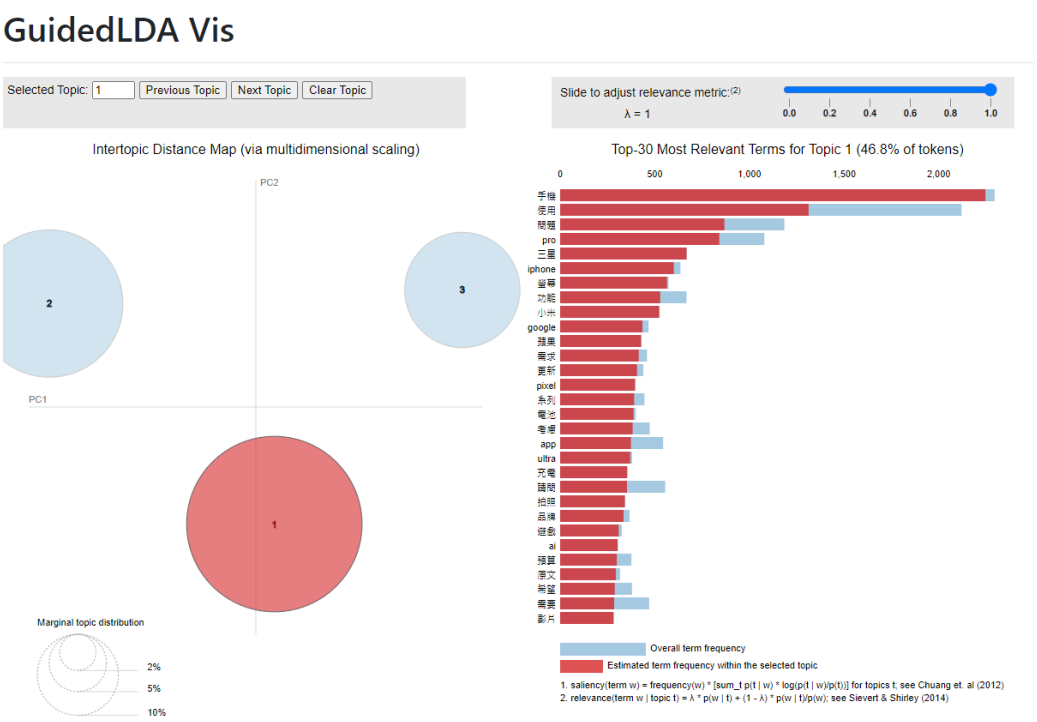
Showing 1 to 3 of 3 entries Previous 1 Next

五、GuideLDA 主題模型

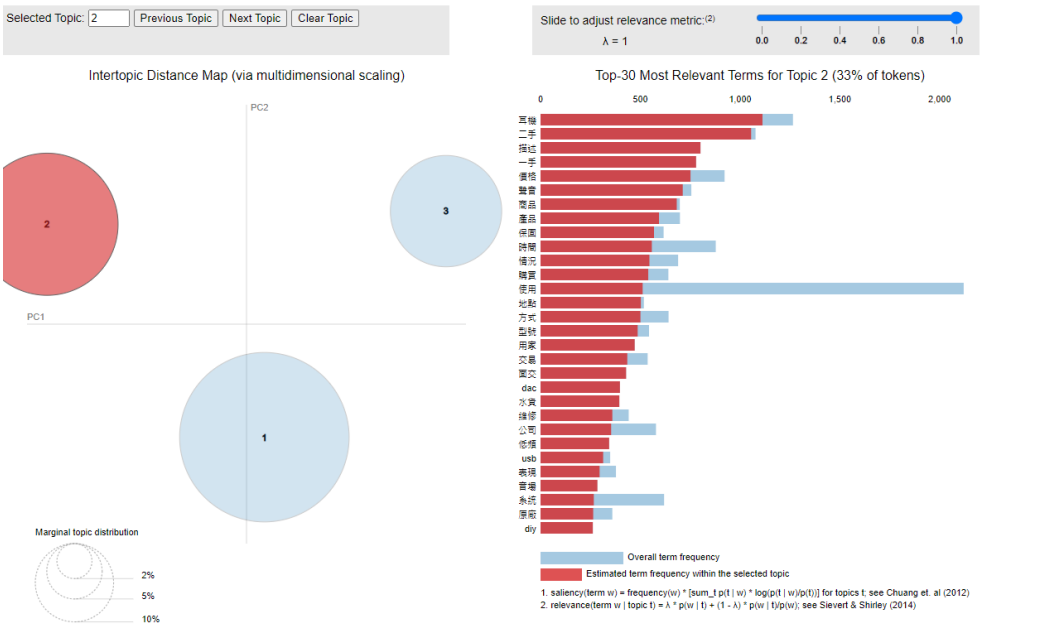
- GuideLDA 主題模型

  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

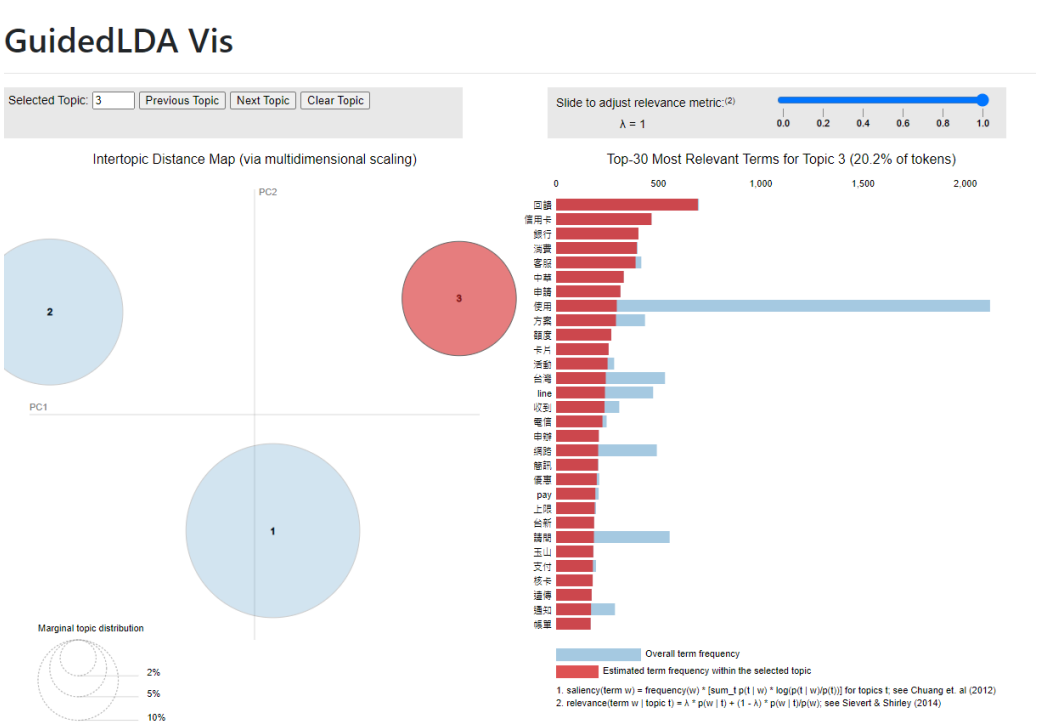
(1) 首先看第 1 群，可以看到主要出現「手機」、「三星」、「使用」、「iphone」等關鍵字，與手機通訊較為相關之關鍵字。



(2) 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。



(3) 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：

(4) 第 1 群(手機通訊)：923 篇

(5) 第 2 群(耳機)：765 篇

(6) 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。