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](#_Toc167730933)

[二、工作流程設計 3](#_Toc167730934)

[三、爬蟲、資料清理 4](#_Toc167730935)

[四、Word2Vec、PCA維度縮減、訓練模型、Bertopic主題模型 7](#_Toc167730936)

[五、留言萃取、社會網路圖 12](#_Toc167730937)

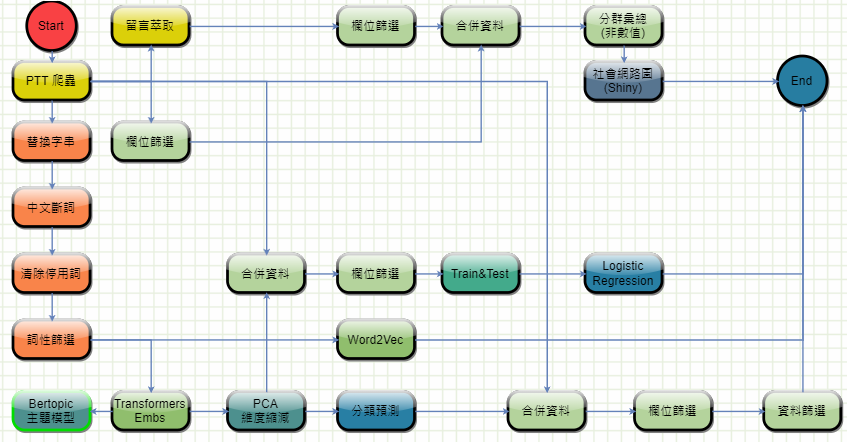
[六、結論 15](#_Toc167730938)

**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4 Final**



* 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「清除停用詞」，將不必要的符號、單字、自定義停用詞去除。
5. 使用「詞性篩選」，將副詞詞性去除。
6. Word2Vec，演算法選擇Skip-grams
7. PCA維度縮減，合併資料。
8. 訓練模型，切割train & test的比例進行模型的訓練。
9. Bertopic主題模型，留言萃取。
10. 合併資料，取得社會網路圖。

**三、爬蟲、資料清理**

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. 資料清理

* 替換字串

1. \n>>，
2. \n\n>>。
3. Sent from>>
4. (BePTT|PttX|JPTT|MeowPtt)>>
5. on my .\*>>
6. Sent from (BePTT|JPTT|PttX|MeowPtt) on my .\*>>
7. ((http|ftp|https)://)[^\s]+(/?[^\s]+)\*>>



* 中文斷詞

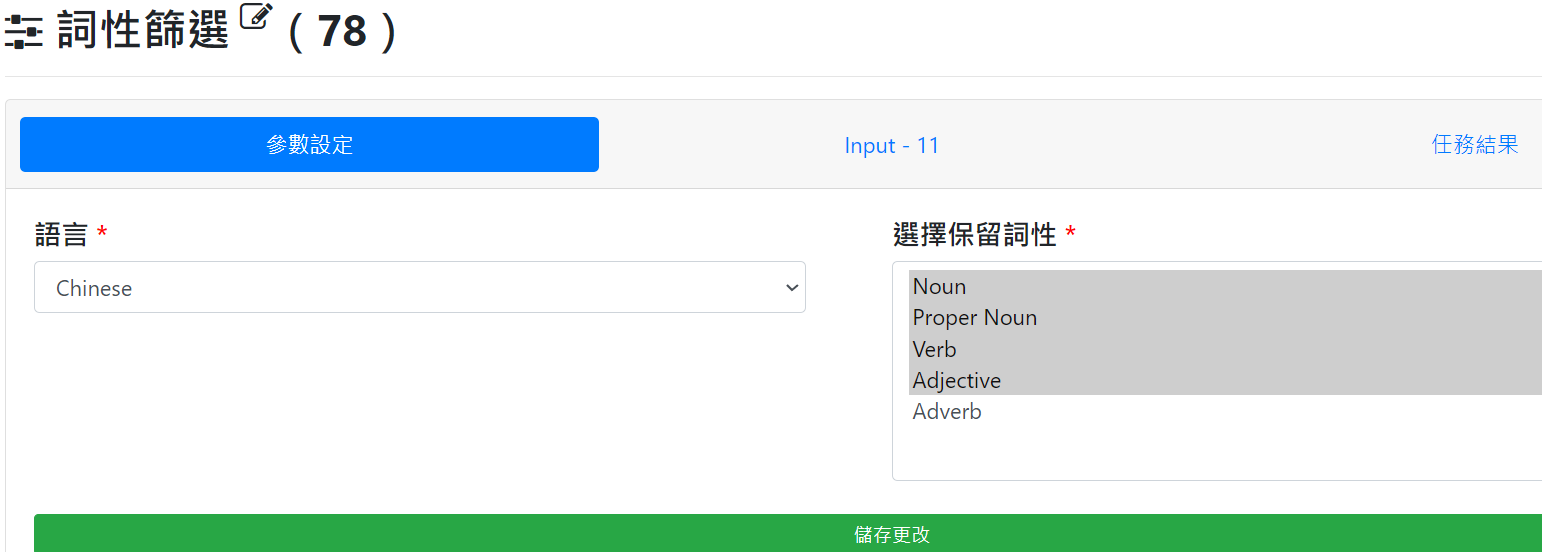


* 清除停用詞



* 詞性篩選

將副詞去除。



**四、Word2Vec****、PCA維度縮減、訓練模型、Bertopic主題模型**

1. Word2Vec

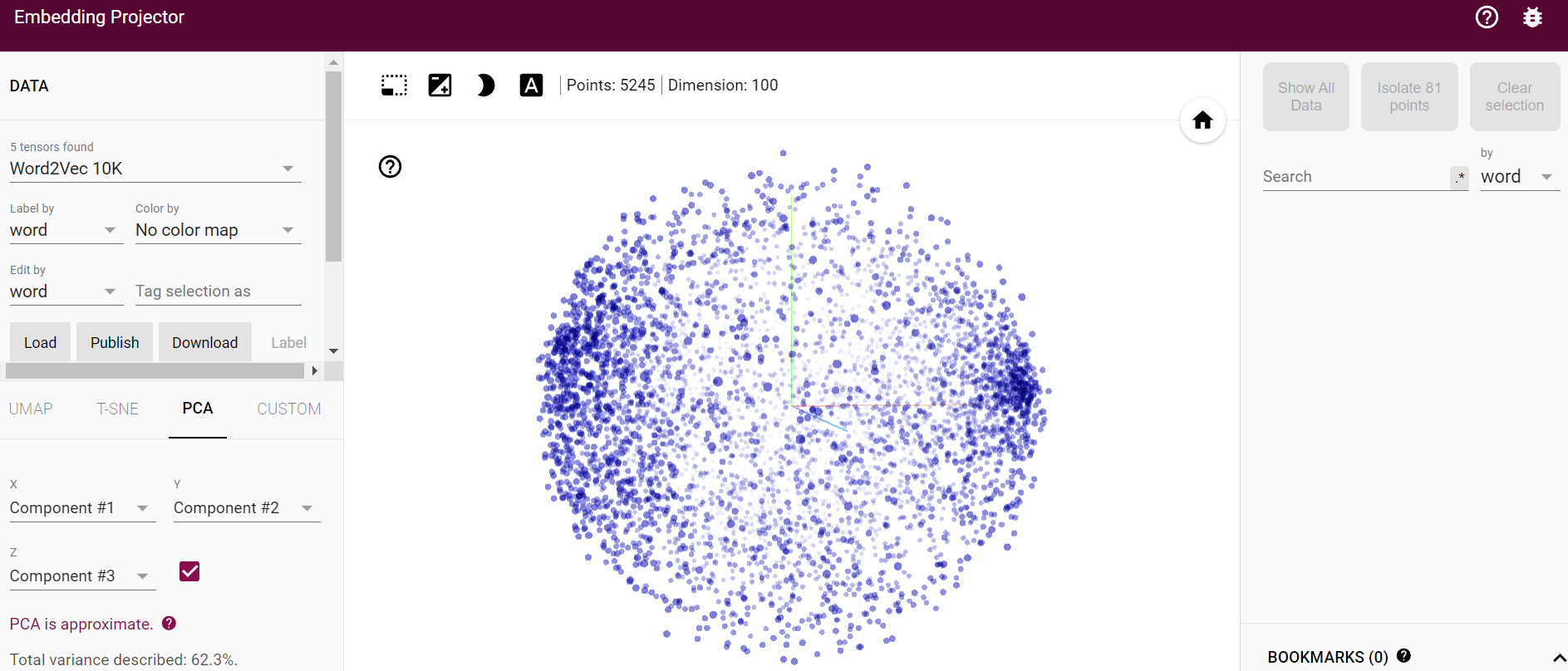
* Word2Vec

利用前述清理完的資料，做Word2Vec，演算法選擇Skip-grams，空間維度設定為100，去除詞頻低於5的詞，參考字詞數為5。



* Embedding Projector

將Word2Vec的結果轉為兩種TSV檔，一個檔案為embedding的向量，另一個檔案為metadata，顯示每個embedding代表什麼意思的。將檔案載入Embedding Projector。

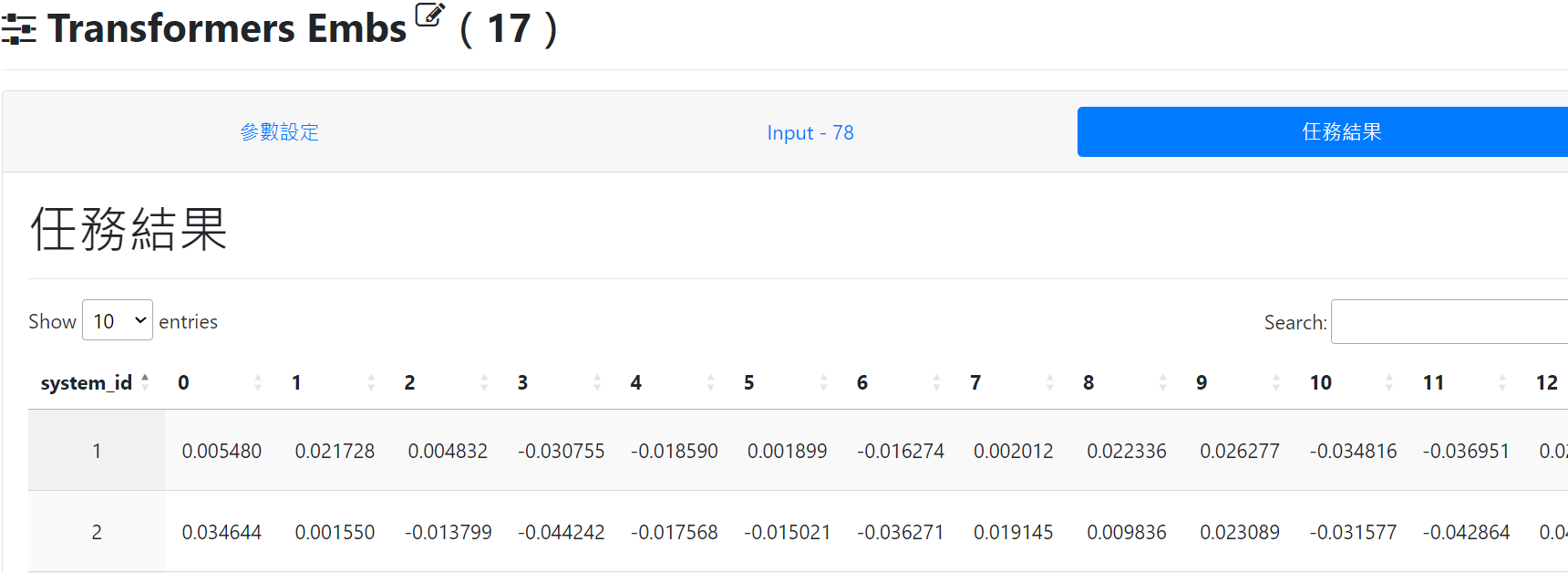


1. PCA維度縮減、訓練模型、Bertopic主題模型

* Transformers Embs

利用前處裡清理完的資料，做Transformers Embs，因為有中、英文，所以模型名稱選擇多語言的模型。輸出結果可看到每個document跟512維度的關係。

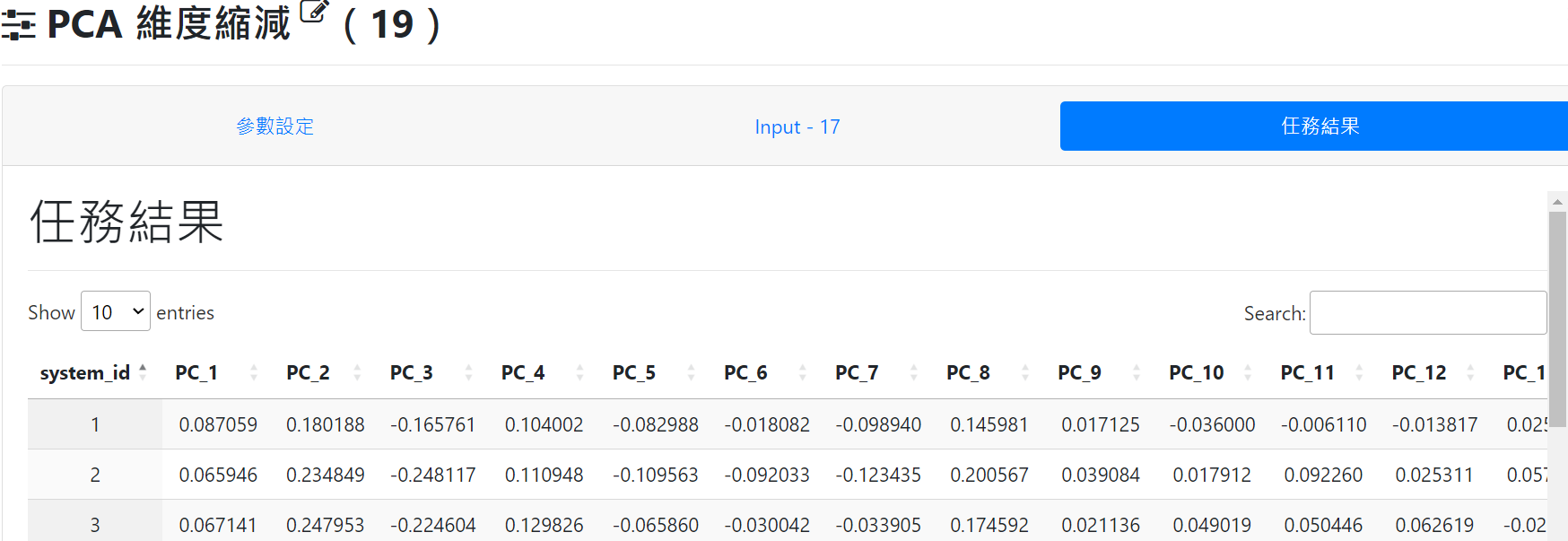




* PCA維度縮減

在做分類模型前，因Transformers Embs有512維度，因此可先透過PCA維度縮減功能，將維度縮減為20。

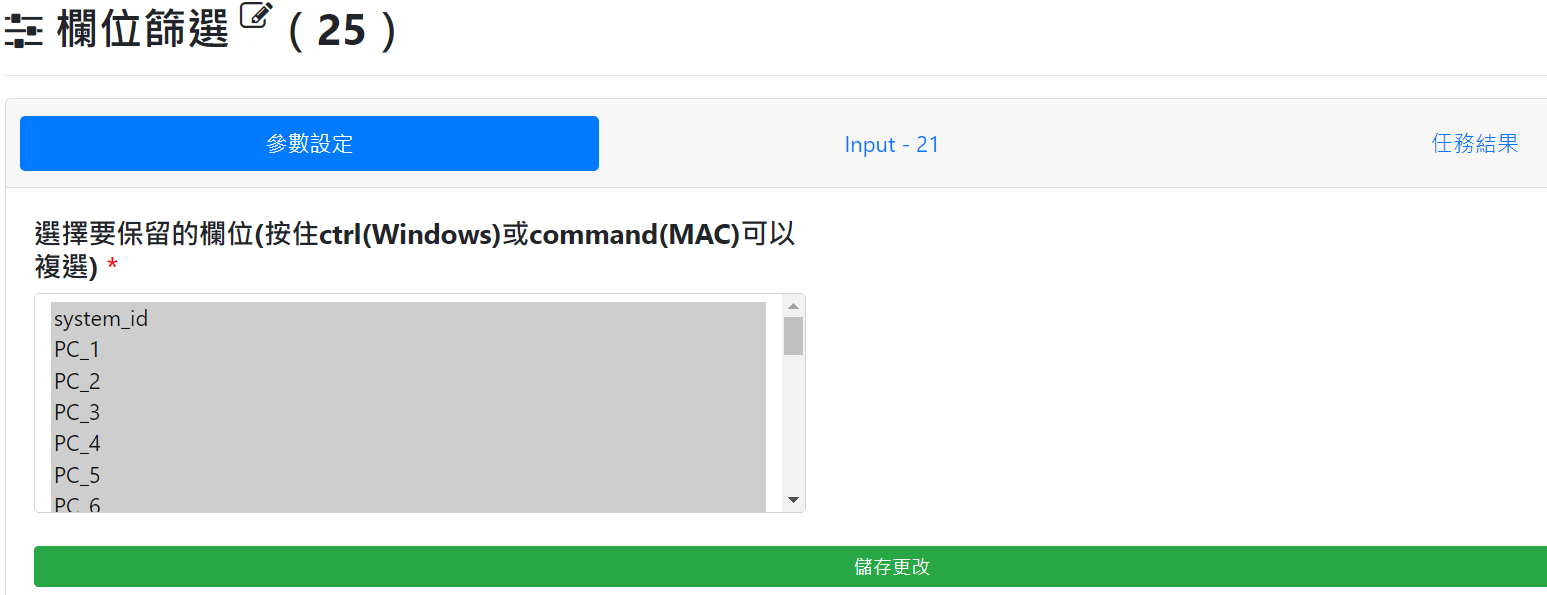




* 合併資料、欄位篩選

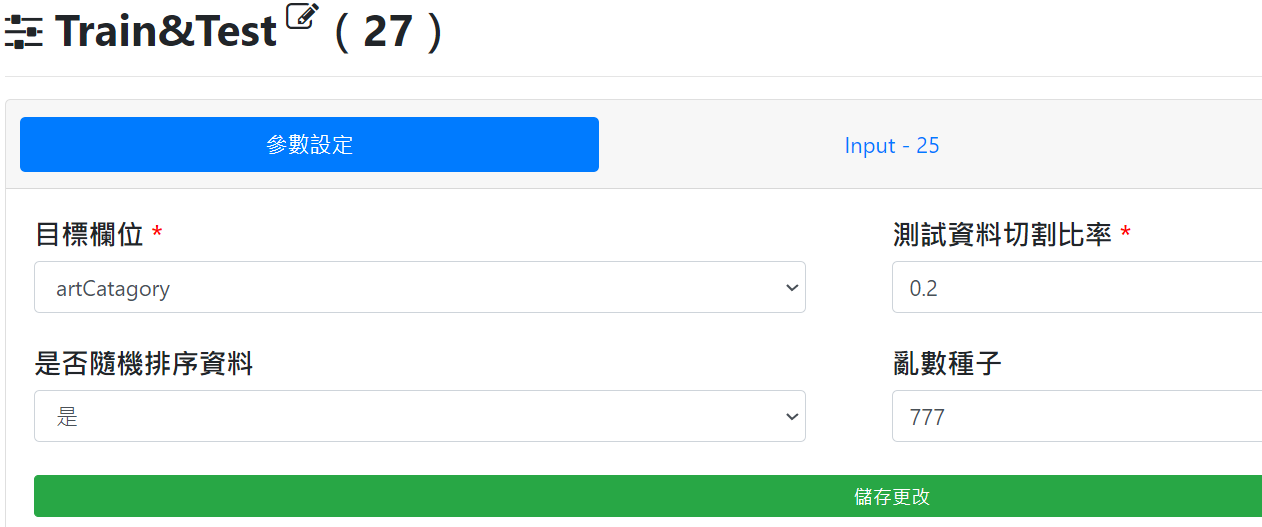
將縮減維度後的資料與PTT原始爬蟲資料利用system\_id合併，並篩選需要的欄位（system\_id、類別、各維度）。





* 訓練模型

將篩選後的欄位，切割train & test的比例進行模型的訓練，使用Logistic Regression，得到F1值為0.884。

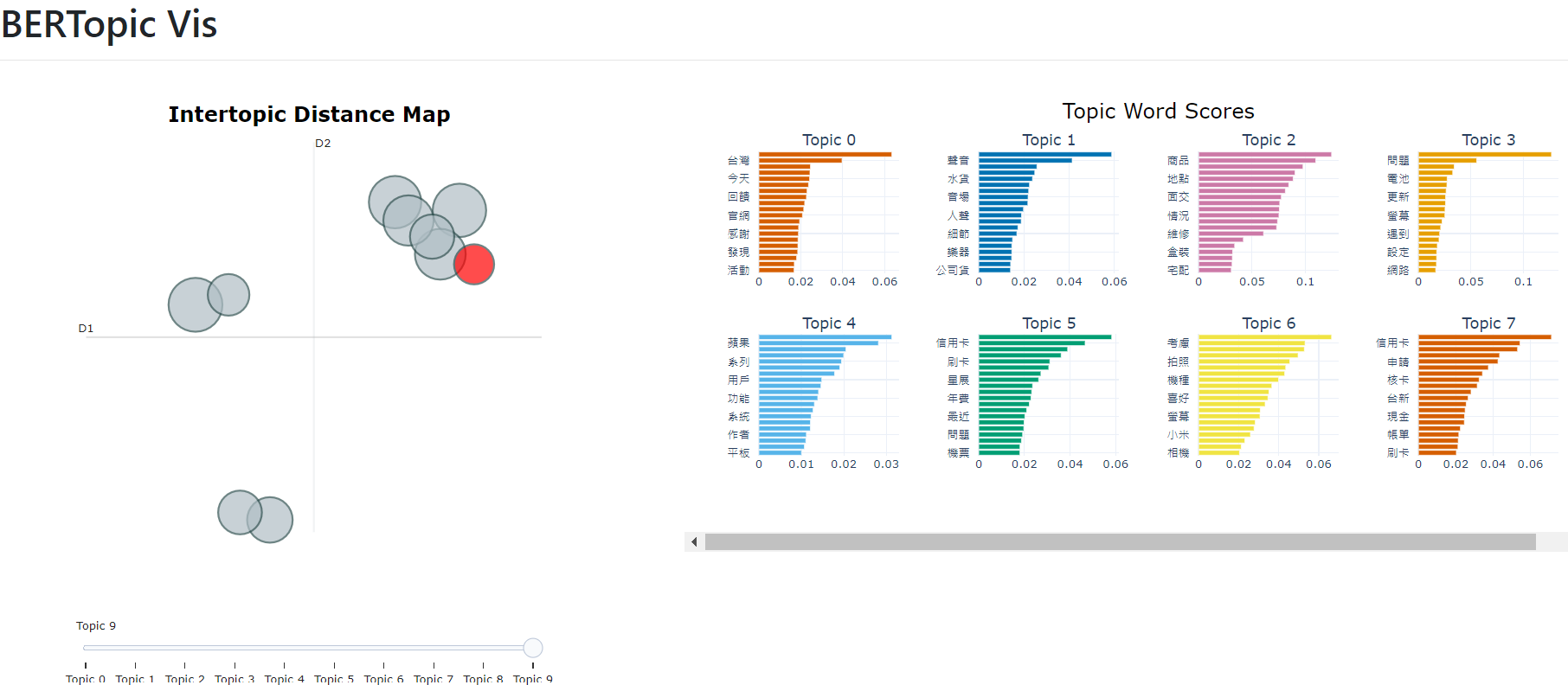




* Bertopic主題模型

利用Bertopic主題模型做分類，參數設定上，語言因有中英文，所以選擇多語言，主題數為10，文章向量縮壓維度為10，主題保留關鍵字數量為10。圖形發現，雖分為三群，但卻沒有分得很清楚，右上角為手機通訊與耳機，左下角為信用卡，左上角則是不確定主題。





# 五、留言萃取、社會網路圖

* 留言萃取

利用留言萃取的功能，將留言與文章拆開，結果可以得到留言的總數量。



* 篩選欄位、合併資料

將原始爬蟲資料、留言萃取後資料篩選所需要的資料，並將兩者藉由system\_id合併，因此就可以得到某一篇文章的發文者是誰，以及某一則留言的留言者是誰。

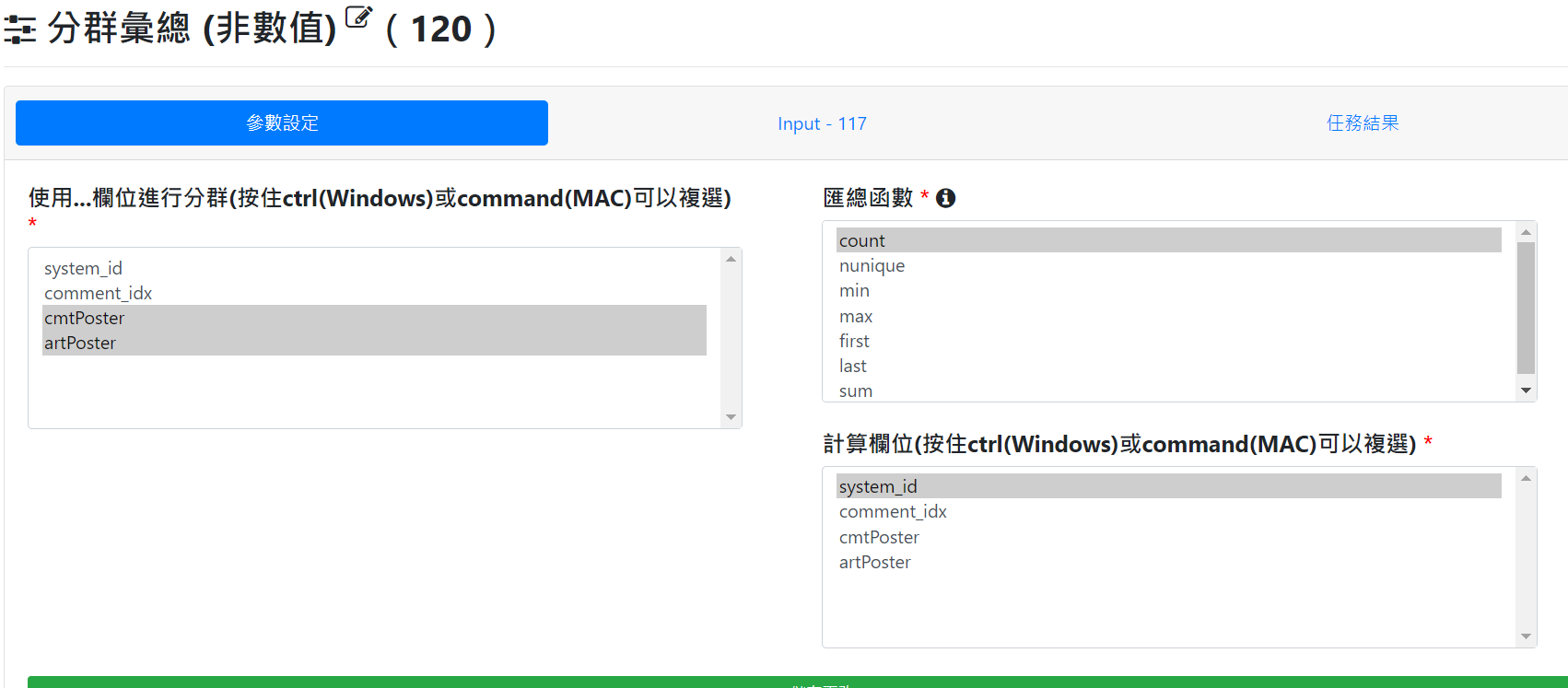


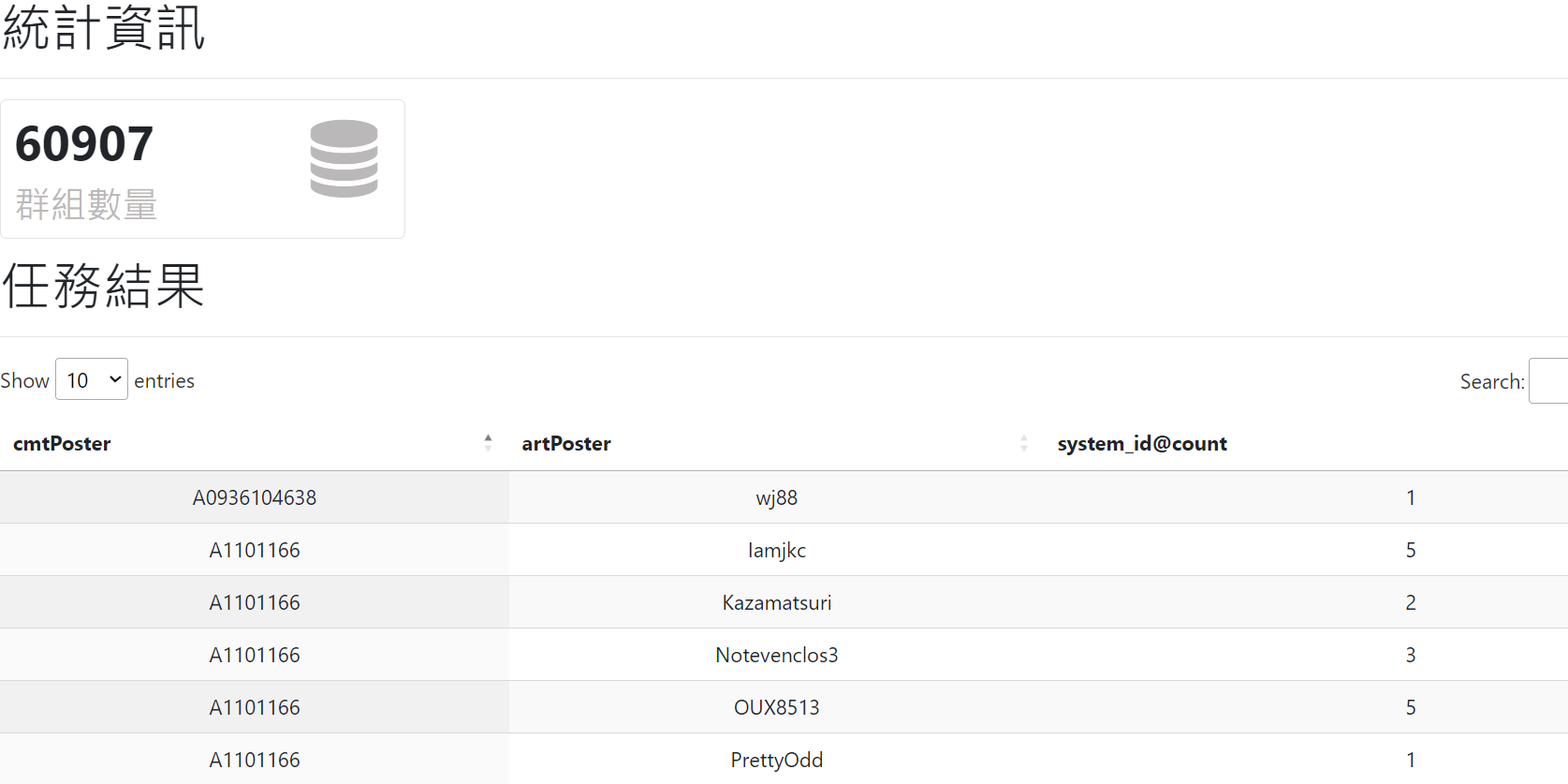




* 分群彙總

要計算發文跟回文之間互動的次數是多少，因此使用分群彙總，另外因為要計算的是system\_id，因此使用非數值的分群匯總。根據發文者與回文者，然後去計算(count)，回文者回覆某個發文者幾次。

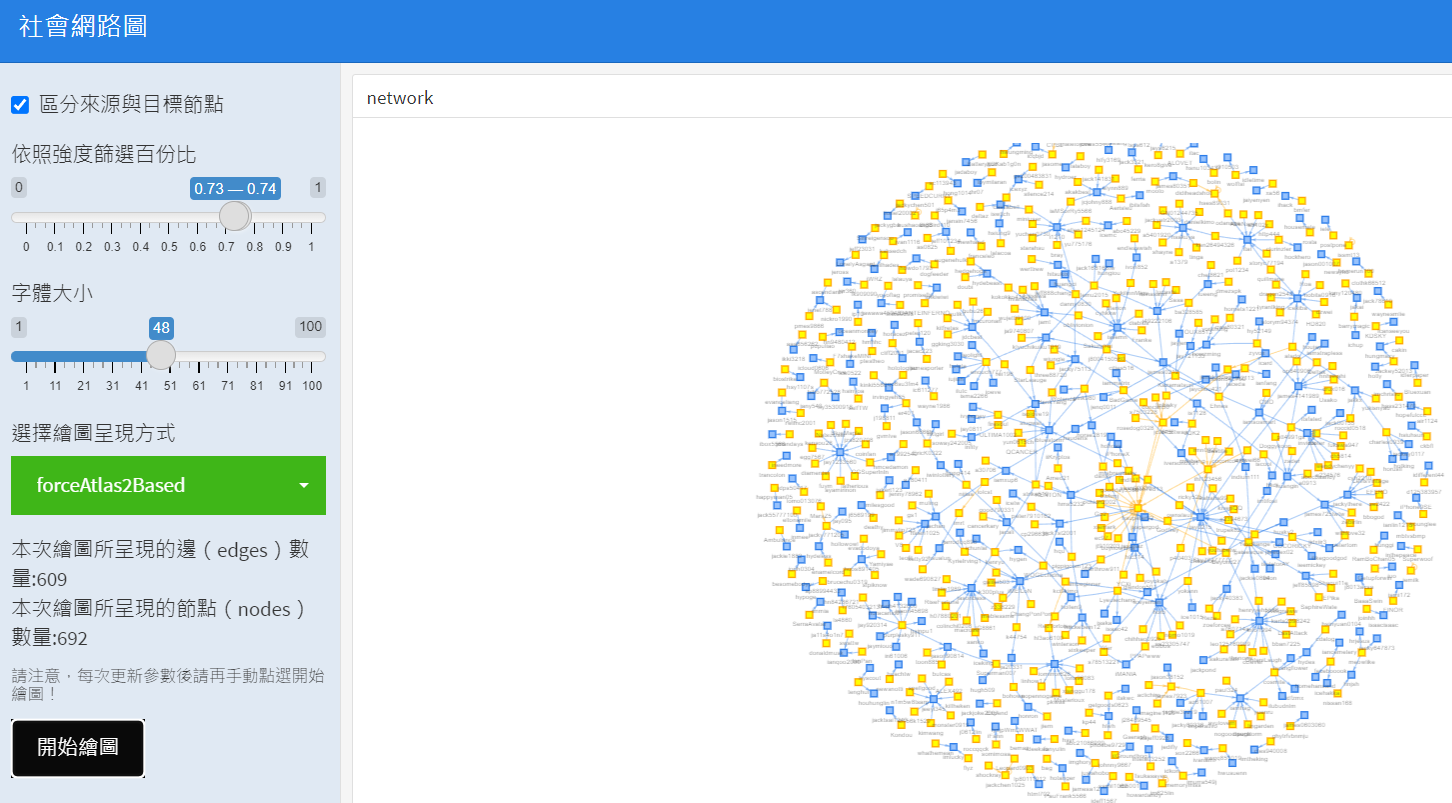




* 社會網路圖

設定參數，節點欄位（來源）：回文者，節點欄位（目標）：發文者，連結欄位：system\_id@count。透過社會網路圖，可觀察到發文者與回文者之間的互動關係，





# 六、結論

資料清理是確保數據質量和分析準確性的基礎。通過對爬取的資料進行替換字串、中文斷詞、清除停用詞及詞性篩選等操作。利用Word2Vec技術，我們對清理後的資料進行詞向量化處理。選用Skip-grams演算法，並將空間維度設定為100，使得模型能夠有效捕捉詞語之間的語義關係。

通過Embedding Projector，我們進一步可視化了詞向量，這對理解詞語在不同文本中的語義差異具有重要意義，在PCA維度縮減與模型訓練方面，我們利用Transformers Embs生成的512維度向量進行PCA縮減至20維度，這不僅降低了數據的維度複雜性，還保留了大部分的重要資訊。隨後，我們通過Logistic Regression對數據進行分類，獲得了F1值為0.884的高效模型，證明了我們的方法在分類任務中的有效性。

此外，我們使用Bertopic主題模型對資料進行主題分類。儘管分群結果顯示出三個主要主題，但分類效果並不十分明顯，特別是在區分手機通訊、耳機和信用卡等類別上。這提示我們，未來的研究中可以考慮引入更多元的特徵和更精細的主題模型，以進一步提升分類準確性。最後，我們進行了留言萃取和社會網絡圖的構建，通過分離留言與文章並計算發文與回文之間的互動次數，清晰可以看見發文者與回文者之間的互動關係。