**社群媒體分析期末報告**

指導老師：黃三益老師

組別：第11組

成員：

N104220015 陳韻元

N104220011 呂伉弘

N104220029 陳聖宗

N104020002 廖英捷

目錄

[一、 動機和分析目的 2](#_Toc137398326)

[（一） 專案說明 2](#_Toc137398327)

[（二） 專案緣起 2](#_Toc137398328)

[（三） 分析平台 2](#_Toc137398329)

[（四） 工作流程 2](#_Toc137398330)

[二、 資料集的描述 2](#_Toc137398331)

[（一） 資料來源 2](#_Toc137398332)

[（二） 關鍵字 2](#_Toc137398333)

[（三） 日期起訖 2](#_Toc137398334)

[三、 資料的分析過程 3](#_Toc137398335)

[（一） 分析模型 3](#_Toc137398336)

[（二） 資料前處理 4](#_Toc137398337)

[四、 視覺化的分析結果與解釋 5](#_Toc137398338)

[（一） 分類模型訓練 5](#_Toc137398339)

[（二） LDA主題模型 13](#_Toc137398340)

[（三） 社會關係網路圖 15](#_Toc137398341)

[五、 結論 29](#_Toc137398343)

[（一） 分類模型訓練 29](#_Toc137398344)

[（二） LDA分類模型 29](#_Toc137398345)

[（三） 社會關係網路圖 29](#_Toc137398346)

# 動機和分析目的

* 1. 專案說明：以文字探勘工作流程設計平台分析PTT上電玩遊戲議題，進行留言萃取建立分類及主題模型，並建立社會網路圖及Gephi圖
  2. 專案緣起：分類模型須人工定義類別，透過PTT看板則可迅速擷取其句子類別，且不同類別的資料筆數不得相差過大，有鑑於身為遊戲玩家常透過社群媒體進行情報交換或經驗分享，因此，本組將運用社群媒體分析所學知識，進行以下討論：
     1. 利用PTT不同類別遊戲看板建模，是否可有效進行分類預測？
     2. 電玩遊戲有哪些相關主題分類？
     3. 電玩遊戲的社會網路圖如何呈現？
     4. 電玩遊戲的Gephi圖如何呈現？
  3. 分析平台：文字探勘工作流程設計平台(<https://tarflow.mis.nsysu.edu.tw/>)
  4. 工作流程：
     1. Final\_Project
     2. Final\_Project\_Inference

# 資料集的描述

* 1. 資料來源：PTT的PC\_Shopping(電蝦)、PlayStation(PS遊戲機)、Steam(Steam)
  2. 關鍵字：手遊、遊戲、電玩
  3. 日期起訖：
     1. 文件分類：

2022/11/01~2022/12/31(訓練模型)

一張含有 文字, 螢幕擷取畫面, 行, 數字 的圖片

自動產生的描述

2023/01/01~2023/01/31、2023/02/01~2023/02/28、2023/03/01~2023/03/31、2023/04/01~2023/04/30 (測試模型)

一張含有 文字, 行, 螢幕擷取畫面, 數字 的圖片

自動產生的描述

* + 1. 主題模型、社會網路圖：2022/01/01~2022/12/31，僅針對PlayStation(PS遊戲機)

一張含有 文字, 螢幕擷取畫面, 行, 數字 的圖片

自動產生的描述

# 資料的分析過程

## 分析模型

一張含有 文字, 螢幕擷取畫面 的圖片

自動產生的描述

## 資料前處理

* + 1. 替換字詞

本組將ptt文章內文中常出現的「\n」及「\n\n」進行替代，以去除雜訊。

一張含有 文字, 字型, 螢幕擷取畫面, 數字 的圖片

自動產生的描述

* + 1. 中文斷詞

將2022年熱門PlayStation及Stem的遊戲名稱詞彙，進行斷詞，一張含有 文字, 螢幕擷取畫面 的圖片

自動產生的描述

一張含有 文字, 字型, 行, 數字 的圖片

自動產生的描述

* + 1. 清除停用詞

將單字元、英文字母、換行符號、html tag、數字、標點符號等停用詞清除，以去除雜訊。

一張含有 文字, 數字, 字型, 行 的圖片

自動產生的描述

# 視覺化的分析結果與解釋

## 分類模型訓練

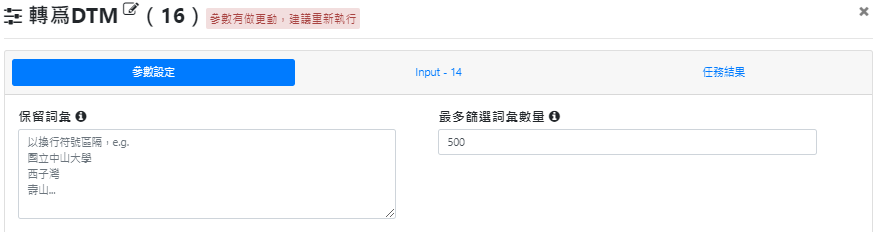
(訓練模型)

一張含有 螢幕擷取畫面, 圖表 的圖片

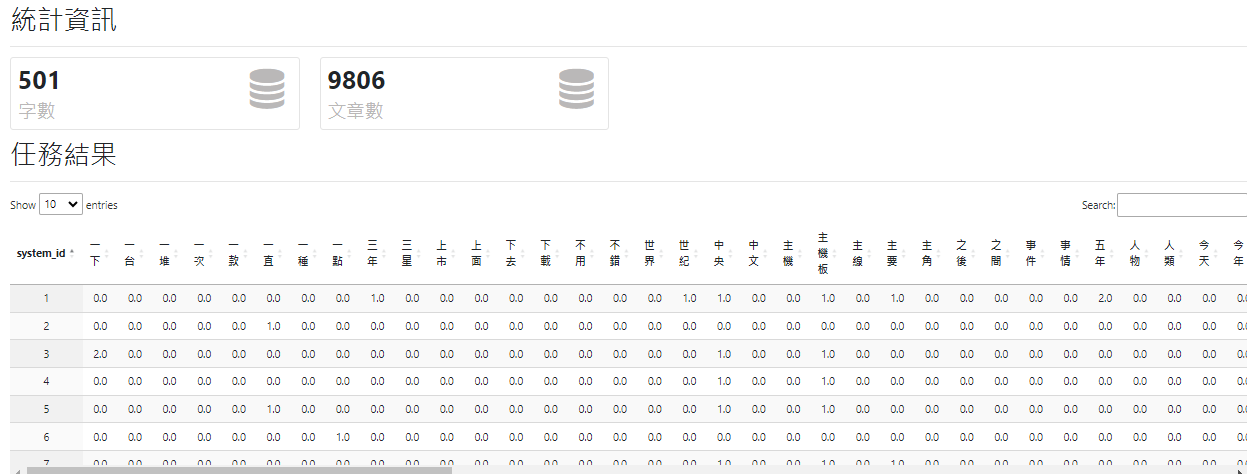
自動產生的描述

* + 1. 轉為Document-Term Matrix(DTM)

輸出文件矩陣(DTM)，以了解每個文件內有哪些字。



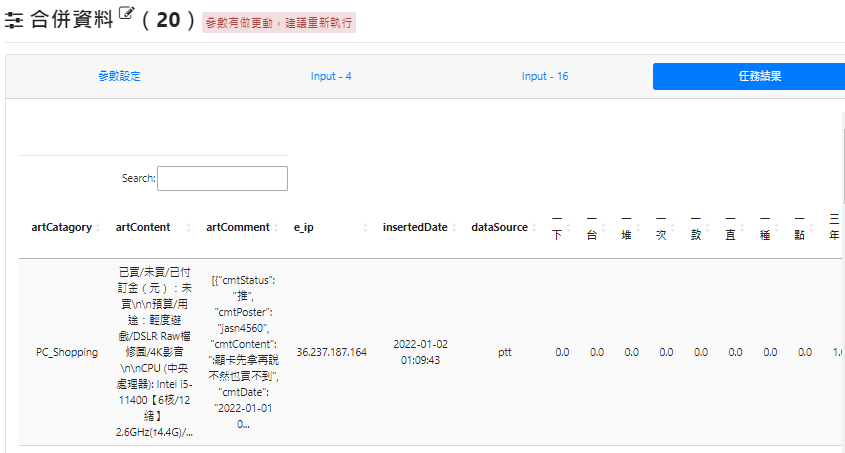
我們知道每一個文件內有哪些字，繪製矩陣，即能知道出現字詞的頻率，比較相關文件之間的相似度，例如使用Cosine Similarity。



* + 1. 合併資料

將原始文本與文件矩陣DTM以system\_id為鍵值，進行資料合併。

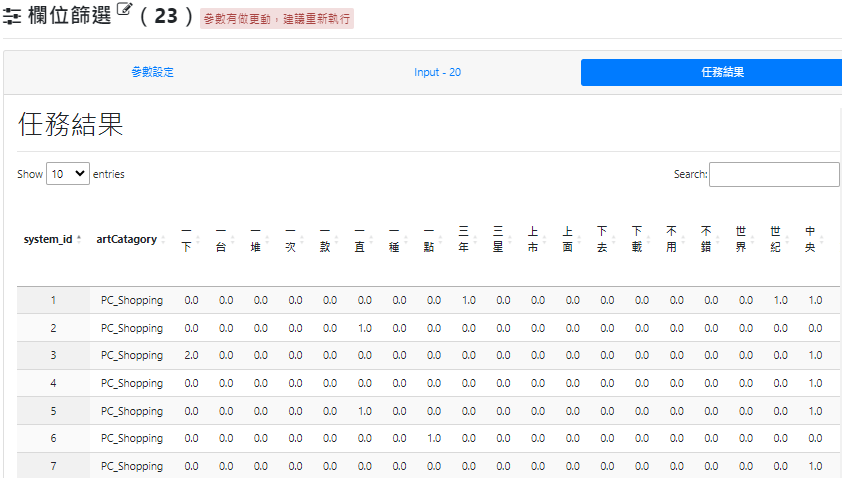




* + 1. 欄位篩選

將文章類別(artCategory)與文件矩陣的詞彙篩選出來，用以訓練模型。





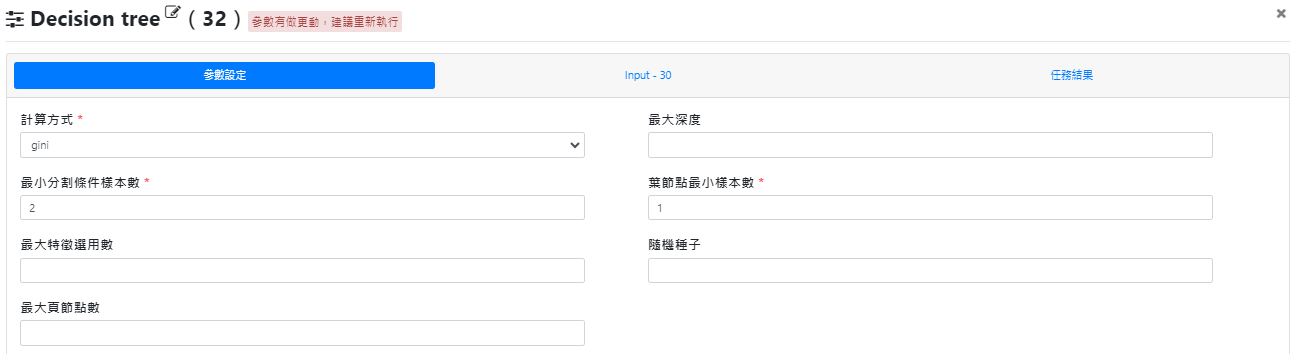
* + 1. 訓練模型-決策樹
       1. Train & Test

將artCatagory欄位隨機排序，以7:3設定訓練資料與測試資料比率，設定亂數種子為168。



* + - 1. Decision tree

建置決策樹



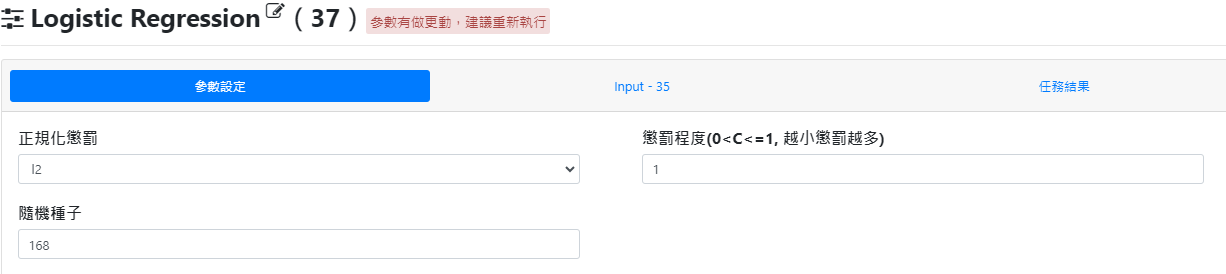
* + 1. 訓練模型-羅吉斯迴歸
       1. K-Fold

本組將資料集進行5-Fold交叉驗證，亂數種子為168。

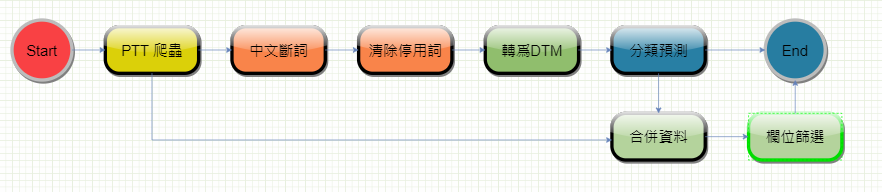


* + - 1. Logistic Regression

建立羅吉斯迴歸模型。

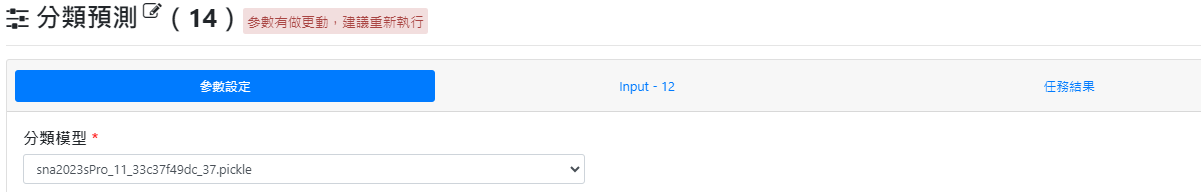


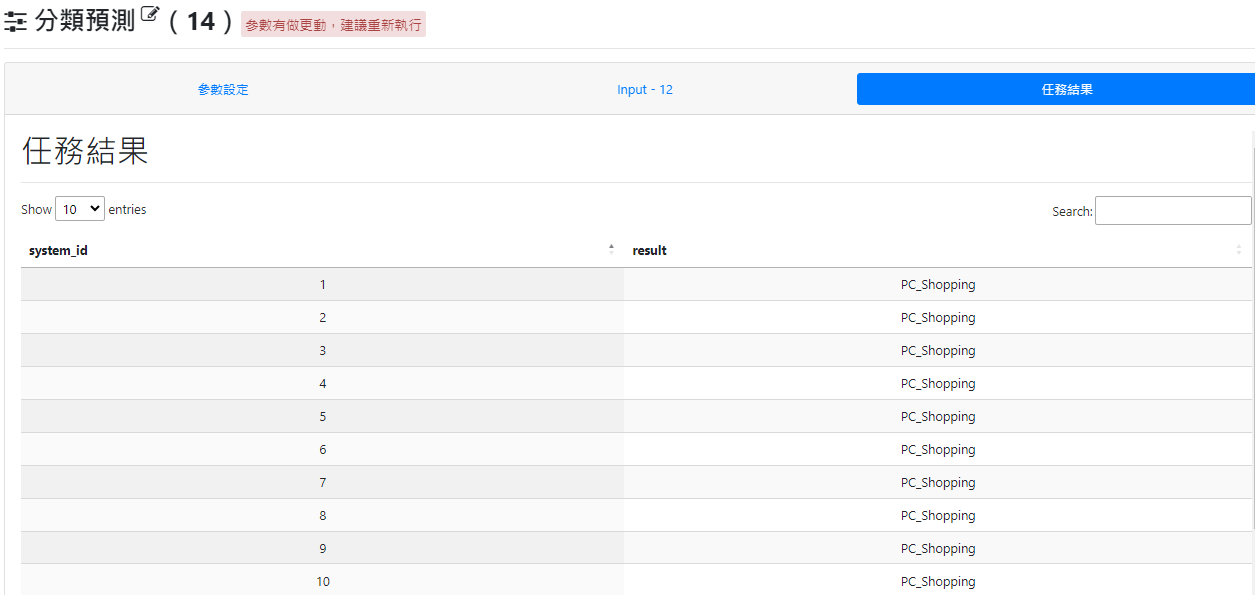
(測試模型)



1. 本組將交叉分析，進行測試以判斷其準確度。
2. 分類預測

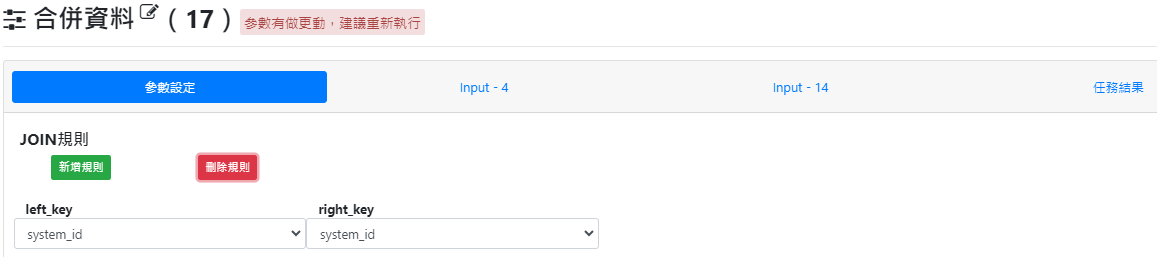
選擇分類模型，產生每個文件所對應的預測類別(result)。





1. 合併資料

將原始文本與文預測類別(result)以system\_id為鍵值，進行資料合併。



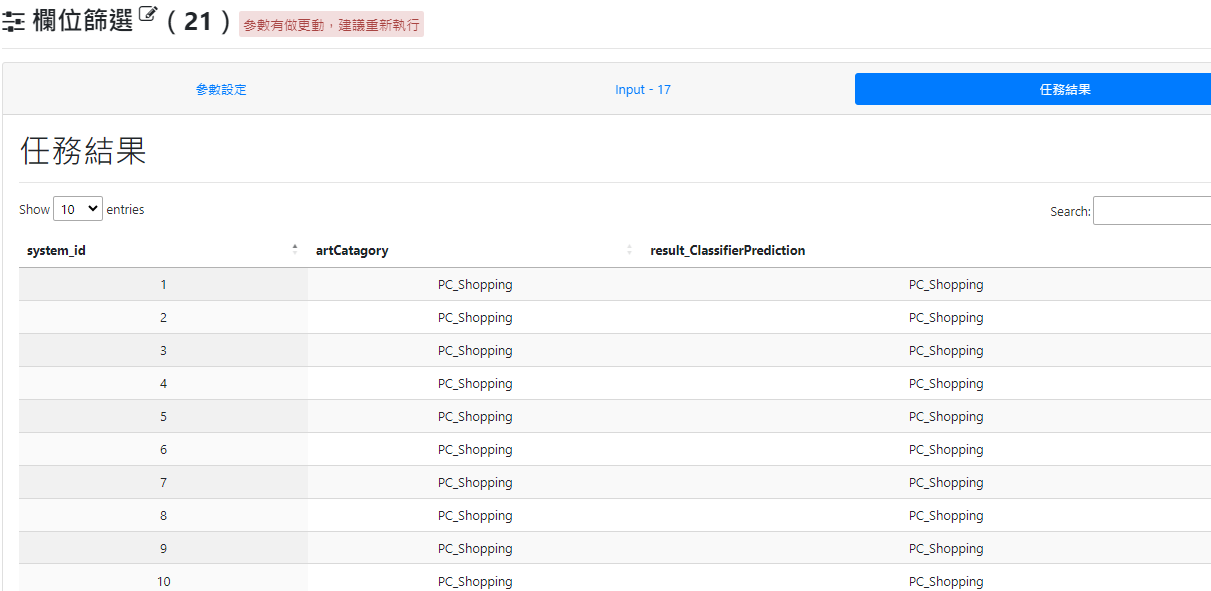
所得資料，可追蹤其文本與所預測類別的差異，做為持續優化分類模型的依據。



1. 欄位篩選

將system\_id、artCatagory、result\_ClassifierPrediction欄位篩選出來，我們將進行預測準確度的計算。





1. 模型調教

本組將Rawdata進行前置處理，我們可以看出有經過前處理，排除雜訊，以及選擇適當DTM大小，會有助於獲得較佳的準確度。

1. 方案一

DTM=500

去除英文=YES

定義詞彙=YES

Train&Test=0.3

Logistic Regression

一張含有 文字, 字型, 白色, 螢幕擷取畫面 的圖片

自動產生的描述

1. 方案二

DTM=300

去除英文=NO

定義詞彙=NO

Train&Test=0.3

Logistic Regression

一張含有 文字, 字型, 白色, 螢幕擷取畫面 的圖片

自動產生的描述

接著，我們嘗試透過Train&Test、K-Fold分別搭配羅吉斯迴歸與決策樹，交叉比對。本組發現，Train&Test搭配羅吉斯迴歸，可獲得相對較佳的準確度。

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

1. 模型預測

建立模型後，我們將模型拿來對後續文章進行預測分析，可以發現，1月份的準確度是0.79、2月份準確度是0.77、3月份準確度是0.76、4月份準確度是0.76。準確度隨時間持續下降，其模型會有過時的問題。

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

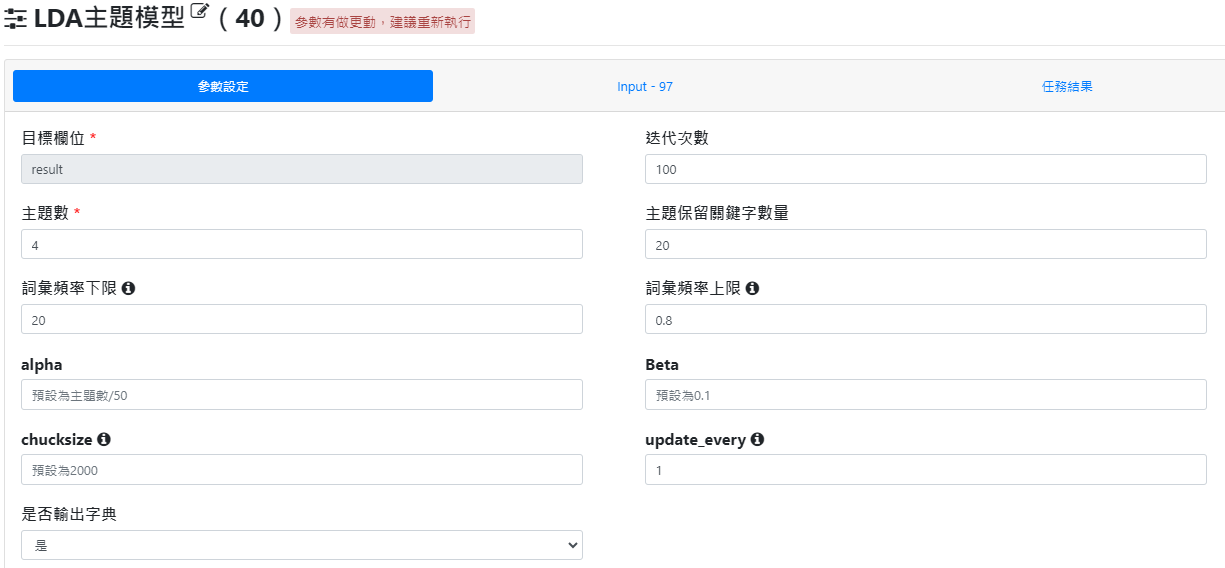
自動產生的描述

但本組發現，在PC\_Shopping(電蝦)、PlayStation(PS遊戲機)、Steam(Steam)版，因為文章內容較無太大幅的變動，因此準確度下降較慢。若是在八卦版這種議題討論多元，變動較大的討論版，準確度可能會有較大幅變動。

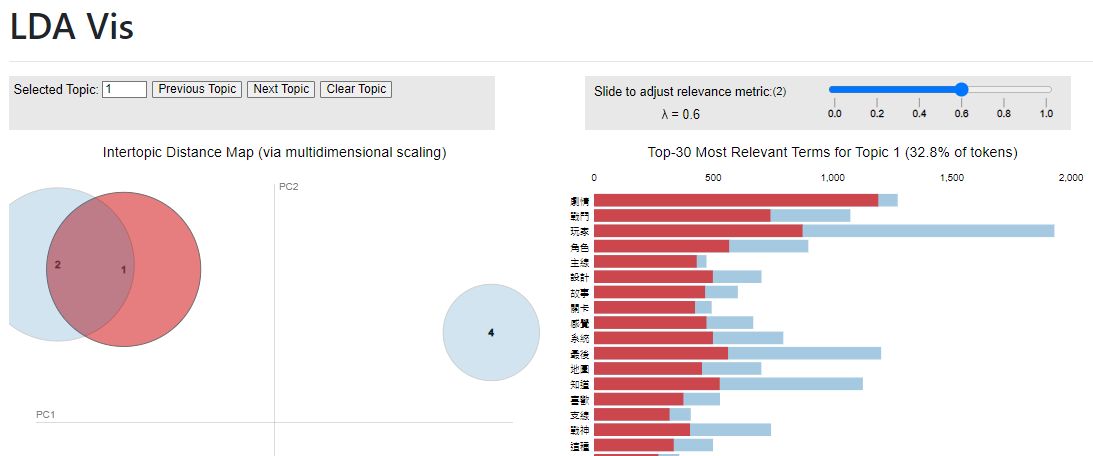
## LDA主題模型



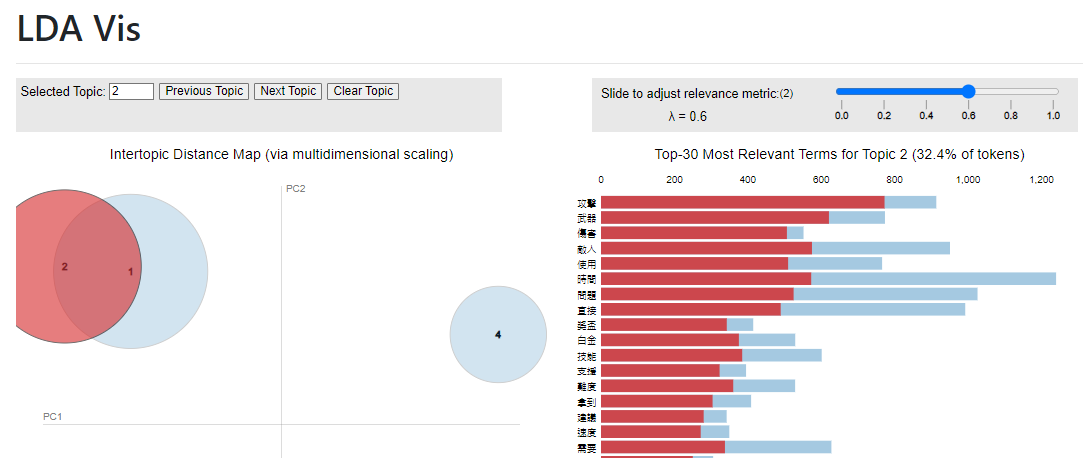
* + 1. 本組擷取主題數為4，並設定詞彙頻率下限須至少出現於20個文本中，且出現頻率不得超過8成以上的文本，再將迭代次數設定為100，抓取主題分類。



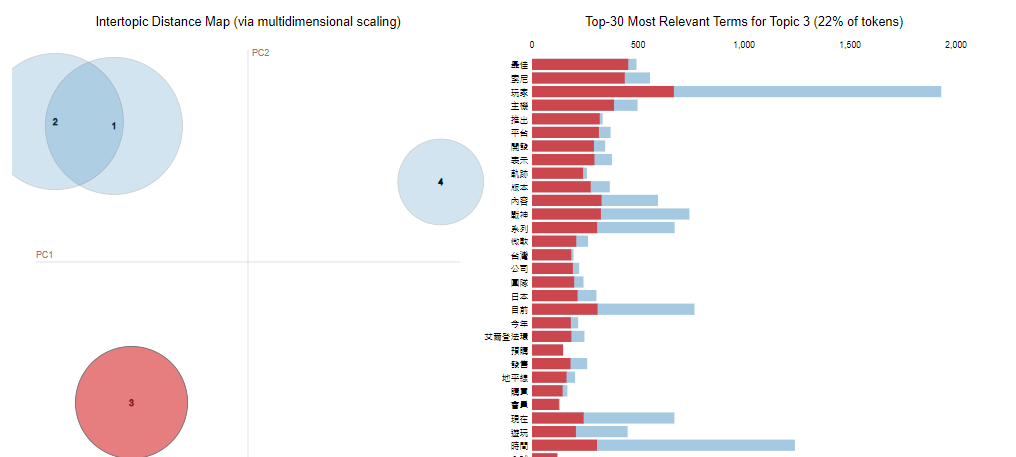
* + 1. 主題一(Topic 1)的主題相符率為32.8%，前10大詞彙為劇情、戰鬥、玩家、角色、主線、設計、故事、關卡、感覺、系統，判斷為「遊戲關卡的設計體驗」。



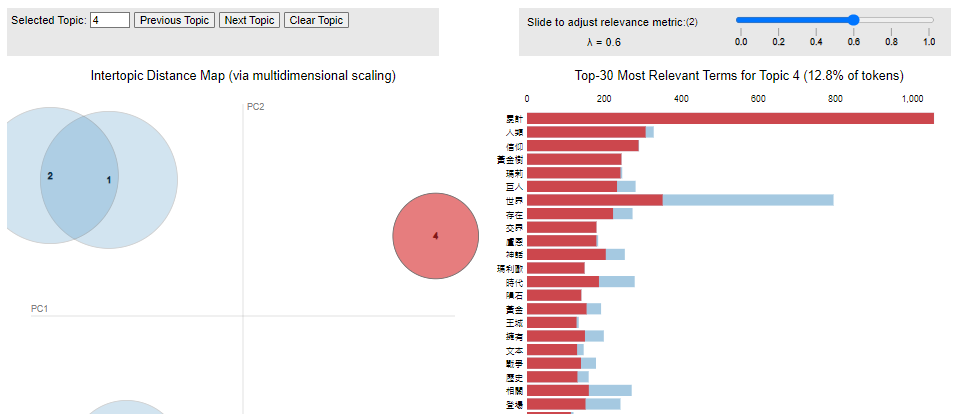
* + 1. 主題二(Topic 2)的主題相符率為32.4%，前10大詞彙為攻擊、武器、傷害、敵人、使用、時間、問題、直接、獎盃、白金，判斷為「遊戲進行中的競技體驗」，其與主題一的意涵相近。



* + 1. 主題三(Topic 3)的主題相符率為22%，前10大詞彙為最佳、索尼、玩家、主機、推出、平台、開發、表示、軌跡、版本，判斷為「玩家對遊戲公司及遊戲機的感受」。



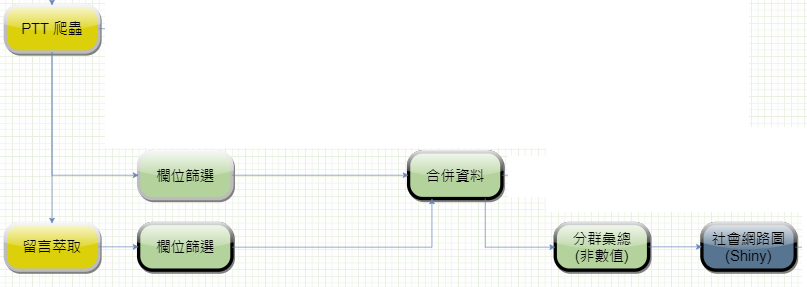
* + 1. 主題四(Topic 4)的主題相符率為12.8%，前10大詞彙為累計、人類、信仰、黃金樹、瑪莉、巨人、世界、存在、交界、盧恩，其中涉及「進擊的巨人」、「瑪莉歐」、「艾爾登法環|ELDEN RING」、「人類：一敗塗地」等多款遊戲，及PS4周邊「PlayStation PS信仰之燈」，判斷為「商城購買遊戲及周邊」。



透過文字探勘工作流程設計平台，我們可以很快知道2022年PlayStation的多款熱門遊戲，以及網路熱賣商品「信仰之燈」。



## 社會關係網路圖



1. PTT爬蟲→欄位篩選
   * + 1. PTT爬蟲

爬取PTT上PlayStation(PS遊戲機)看板上2022/01/01~2022/12/31的文章。

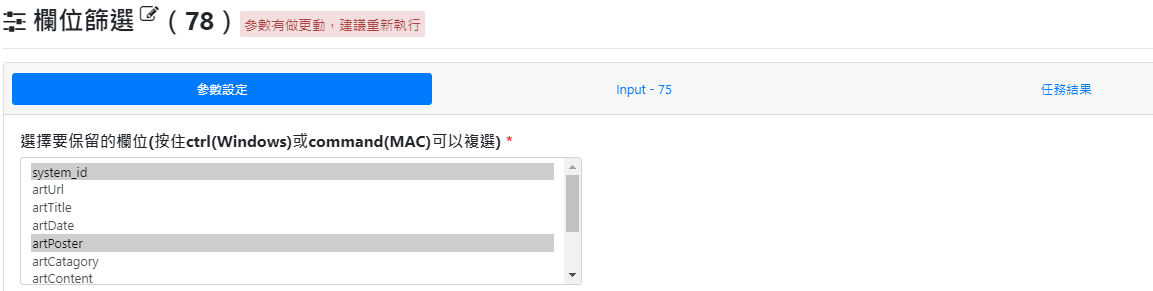


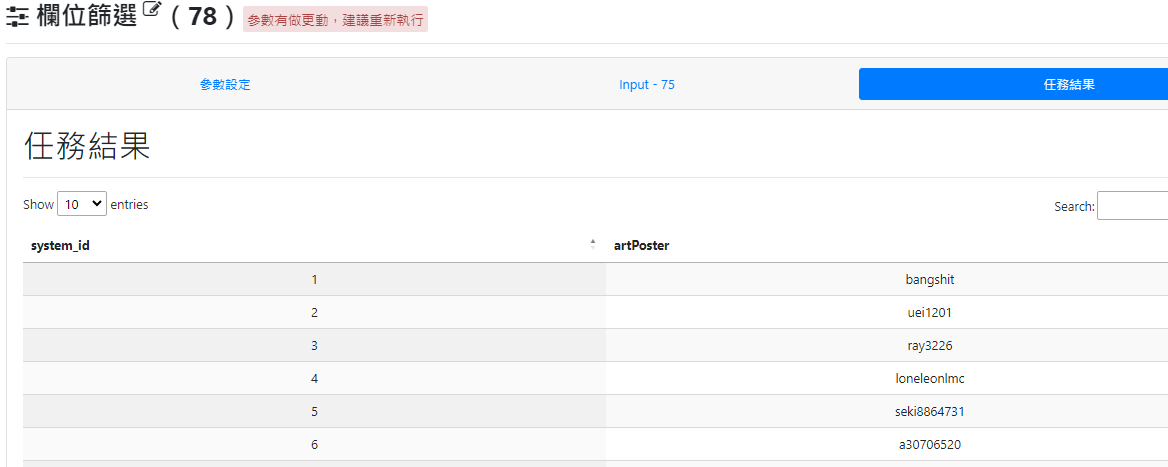
計有2302筆資料。



* + - 1. 欄位篩選

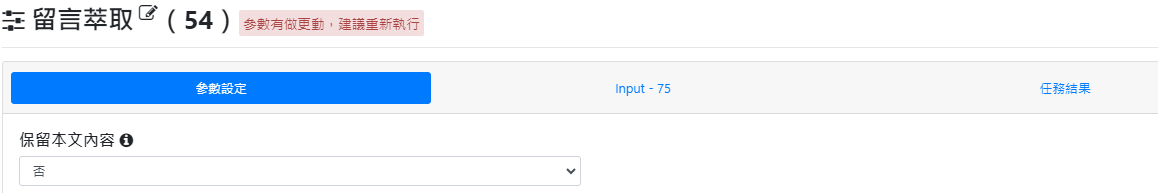
篩選system\_id、artPoster欄位，取得貼文者資訊。

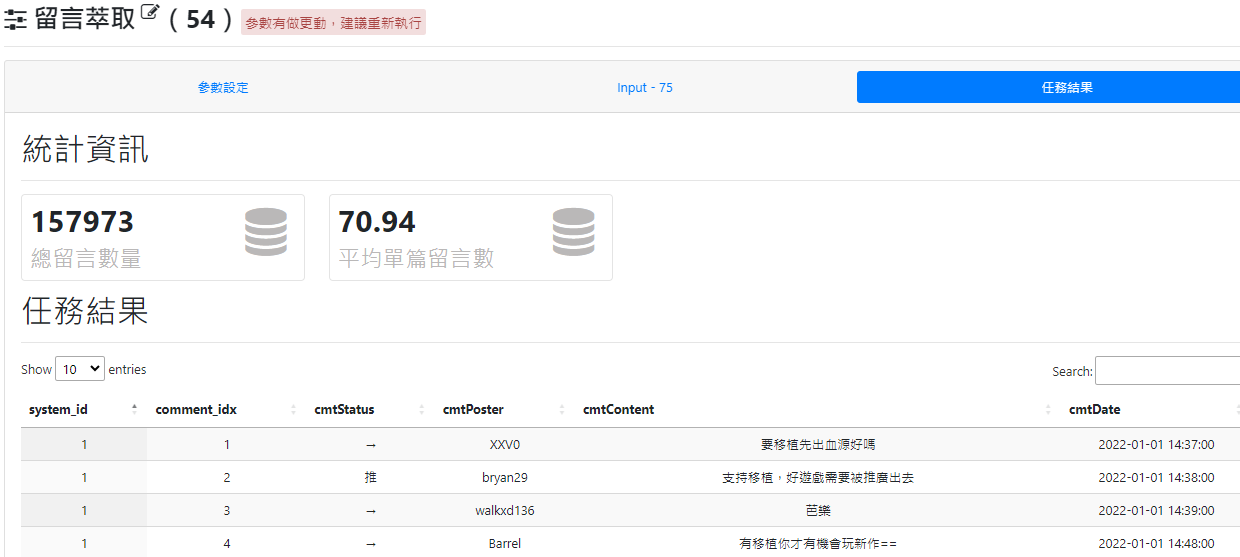




1. 留言萃取→欄位篩選
2. 留言萃取

不保留本文內容，取得相對的留言資料，共計157973筆留言。

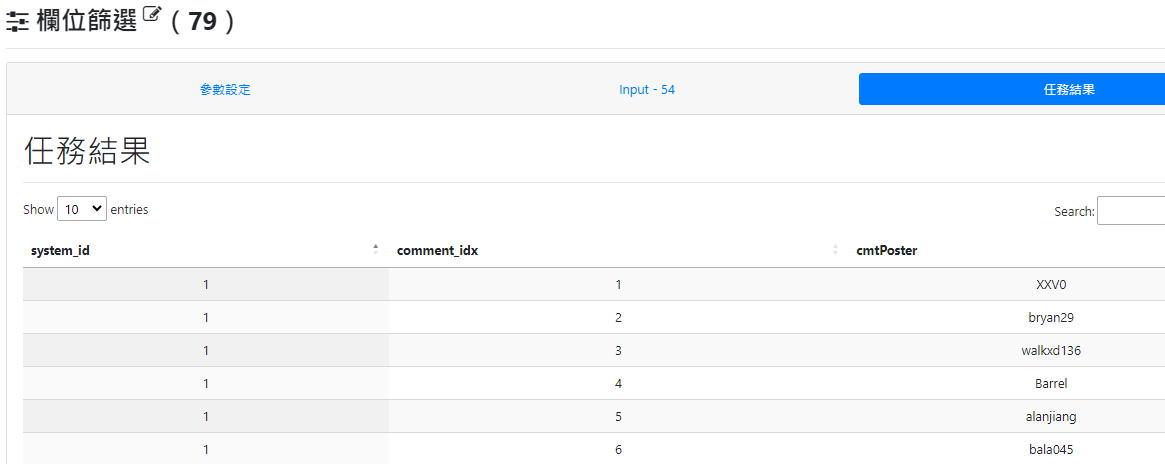




1. 欄位篩選

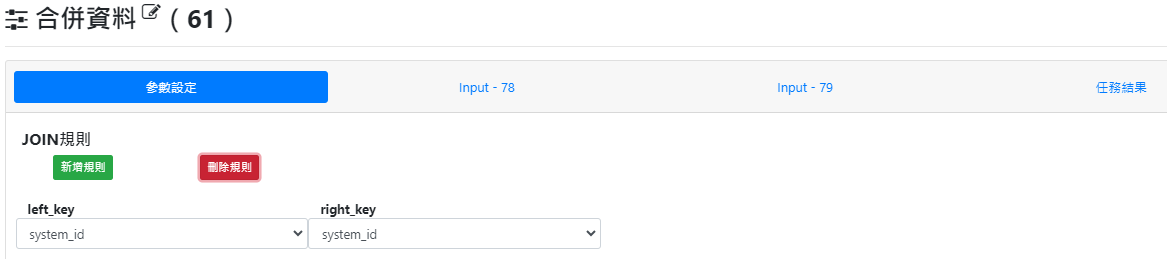
篩選system\_id、comment\_idx、cmtPoster欄位，取得留言者資訊。

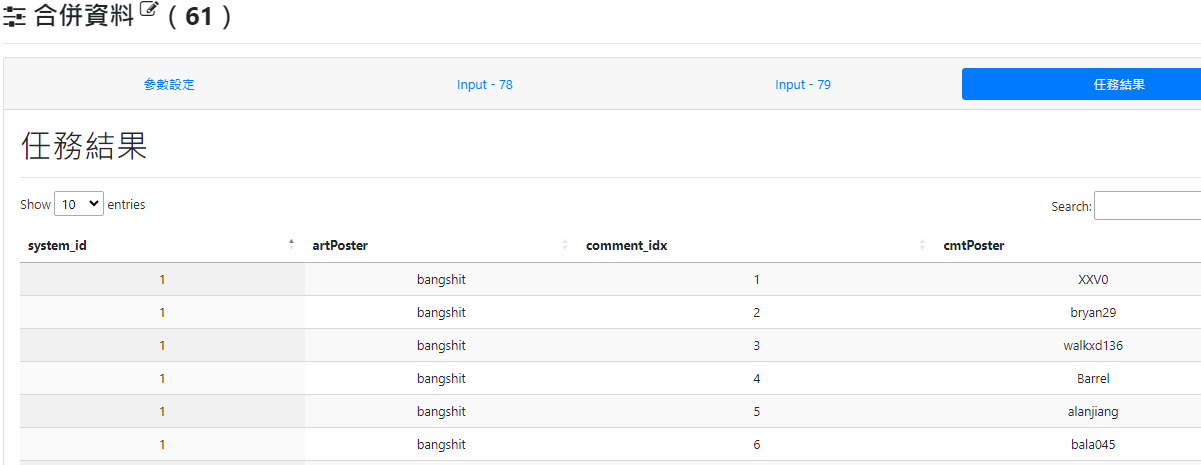




1. 合併資料

將貼文者篩選資料與留言者篩選資料以system\_id為鍵值，進行資料合併。



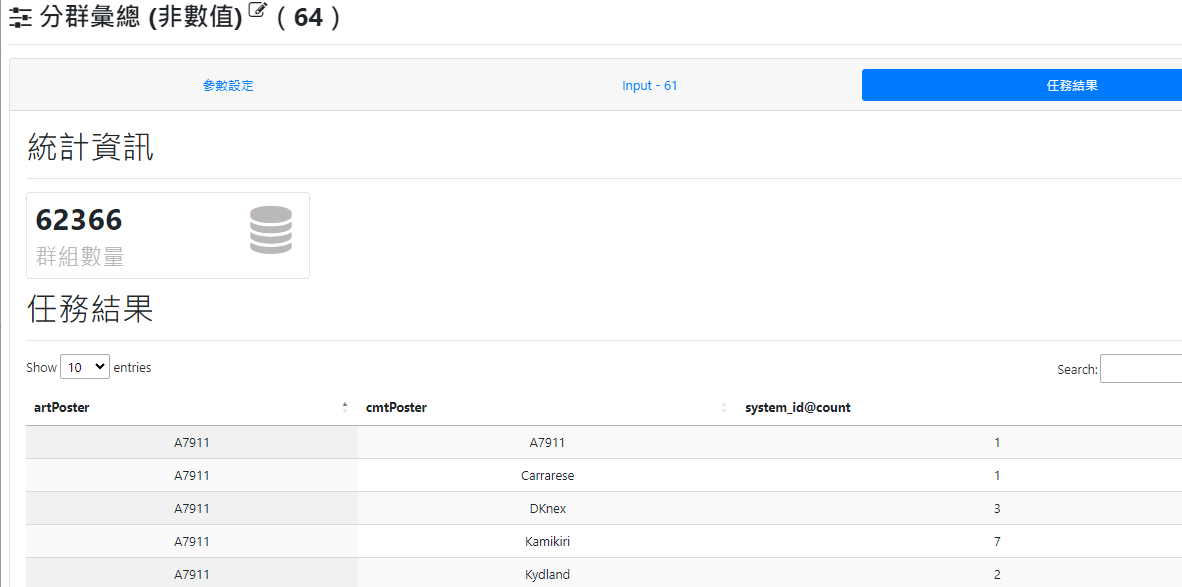


1. 分群彙總(非數值)

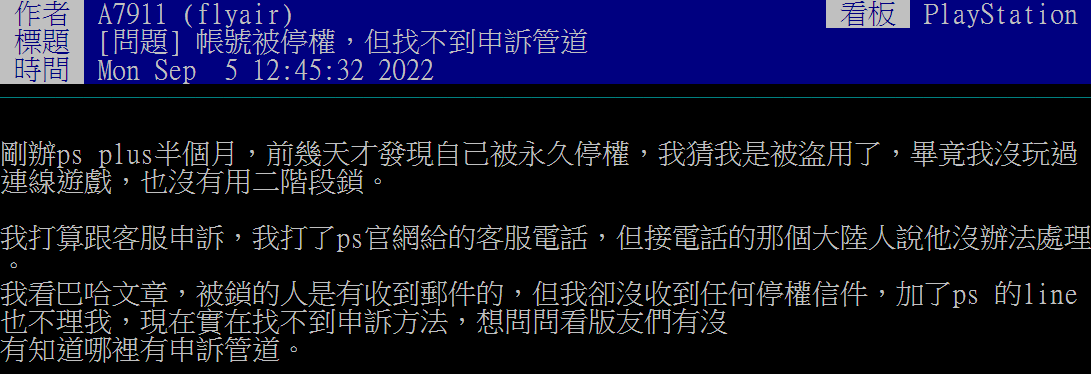
本組以(artPoster, cmtPoster)作為鍵值，進行分群彙總，計算count(system\_id)，找出貼文者與留言者間有多少留言數。

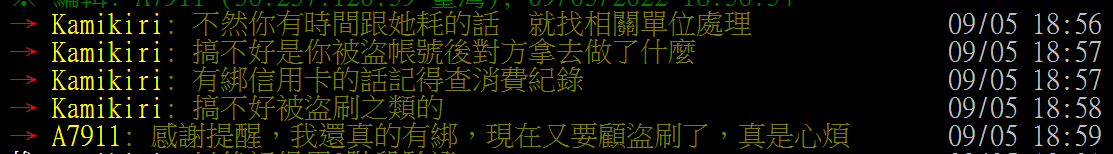


取得網路圖間的關聯，特別一提的是，因貼文者也會透過留言回復其他的留言者(如A7911)，因此，可能形成節點(node)自己的邊(edge)。



PTT內，A7911於2022/09/05新增一則貼文，並於留言處，回復Kamikiri進行了留言。



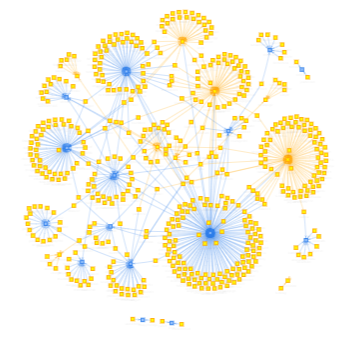


1. 社會網路圖(Shiny)

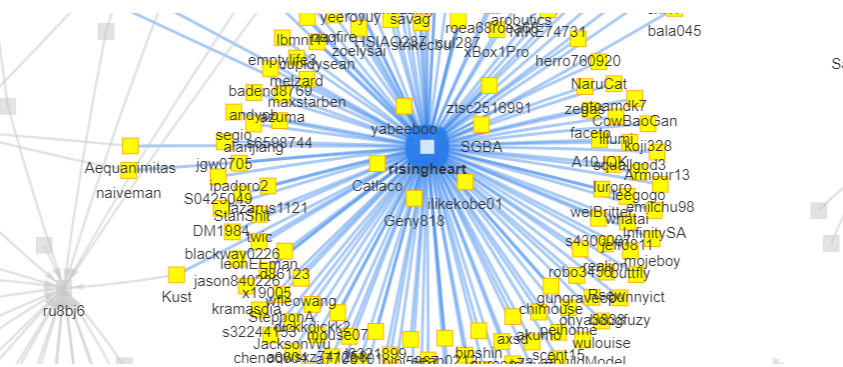
由貼文者(artPoster)連結至留言者(cmtPoster)，邊加上權重(留言總數, system\_id@count)



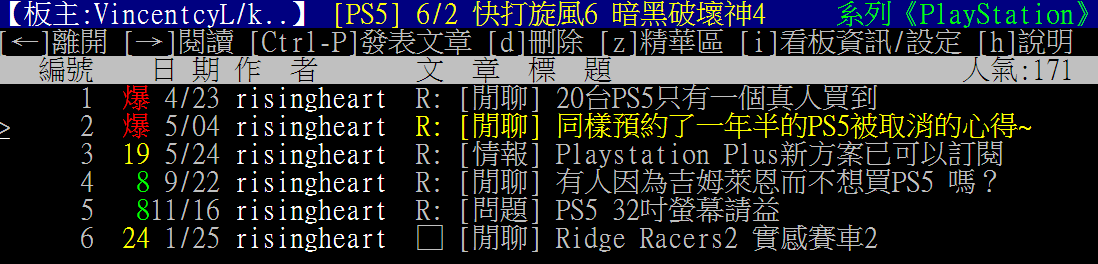
勾選【區分來源與目標節點】，區分來源(藍色)節點與目標(黃色)節點，可繪製社會網路圖。

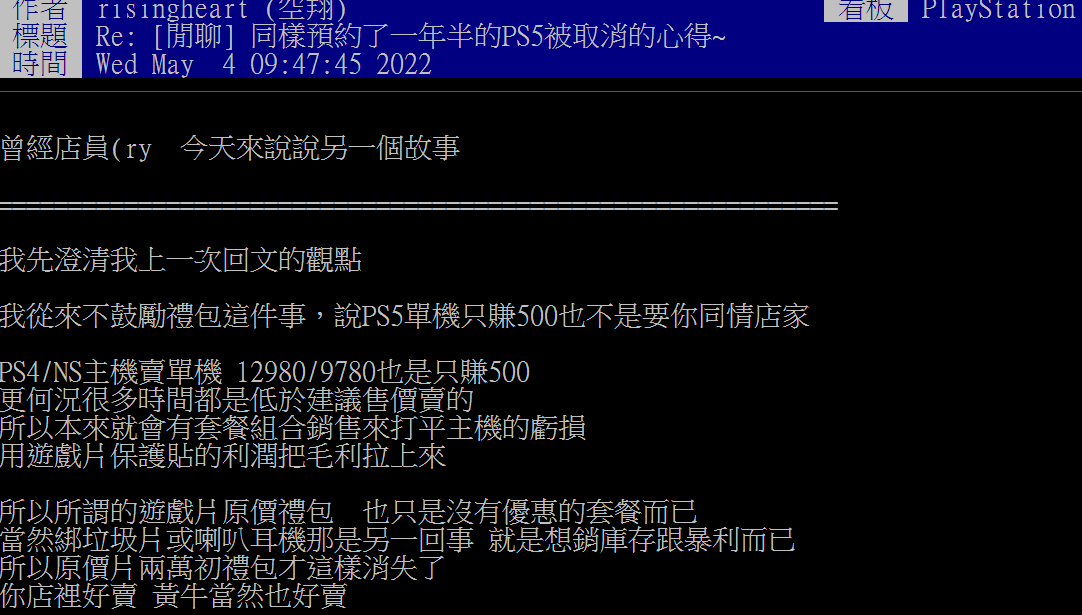


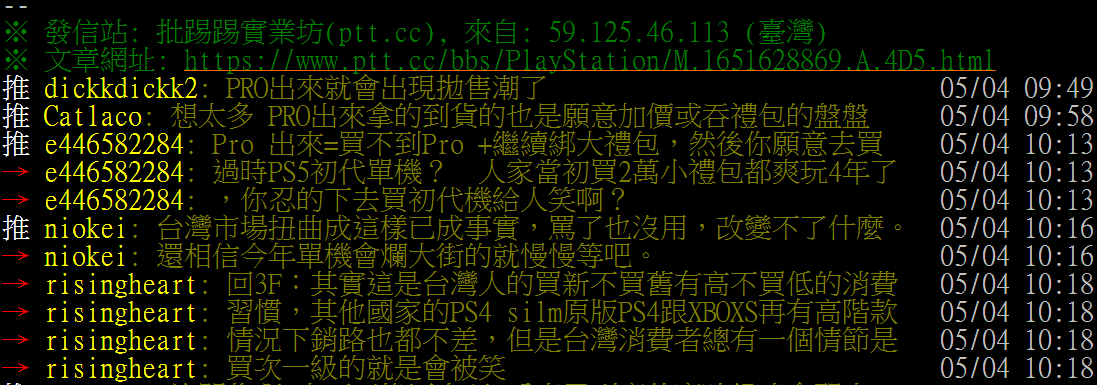
我們可看到有形成數個聚落，其中有一個最大的聚落，其中心點為發文者risingheart。

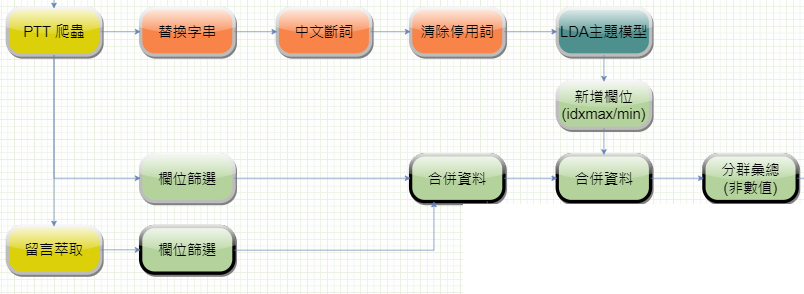


PTT中，可以看到risingheart貼文，以2022/05/04為例，更是與多位留言者互動討論，非常活躍。





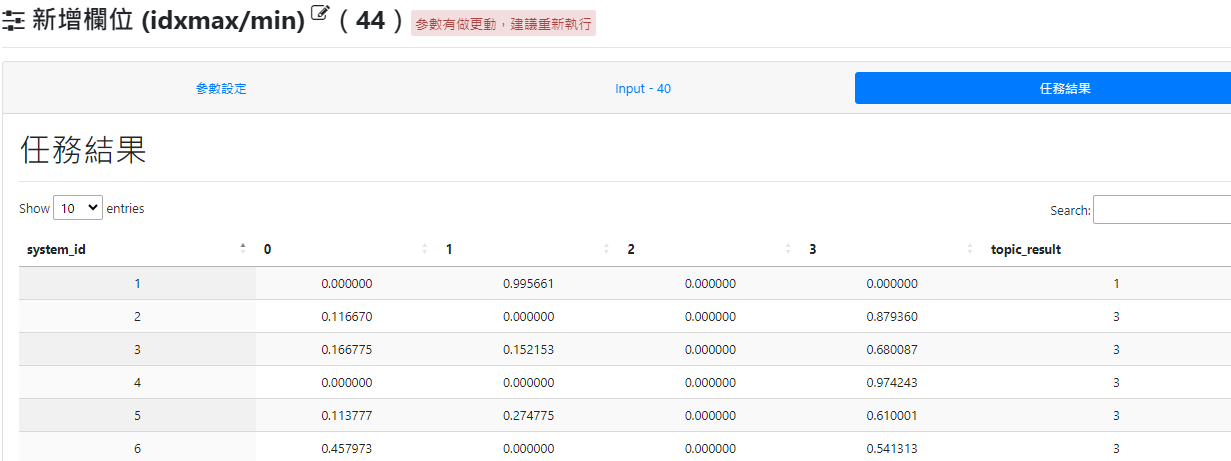




1. 新增欄位(idxmax/min)

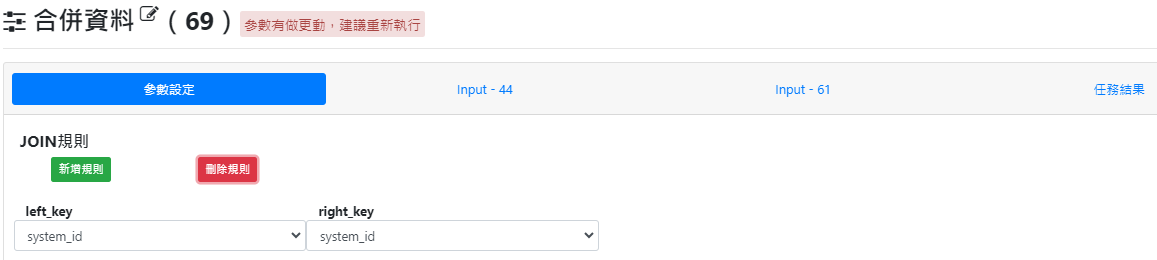
我們將LDA主題分類所得資料，計算其分類欄位的數值，並取最大值作為該文件的主題類別(topic\_result)。

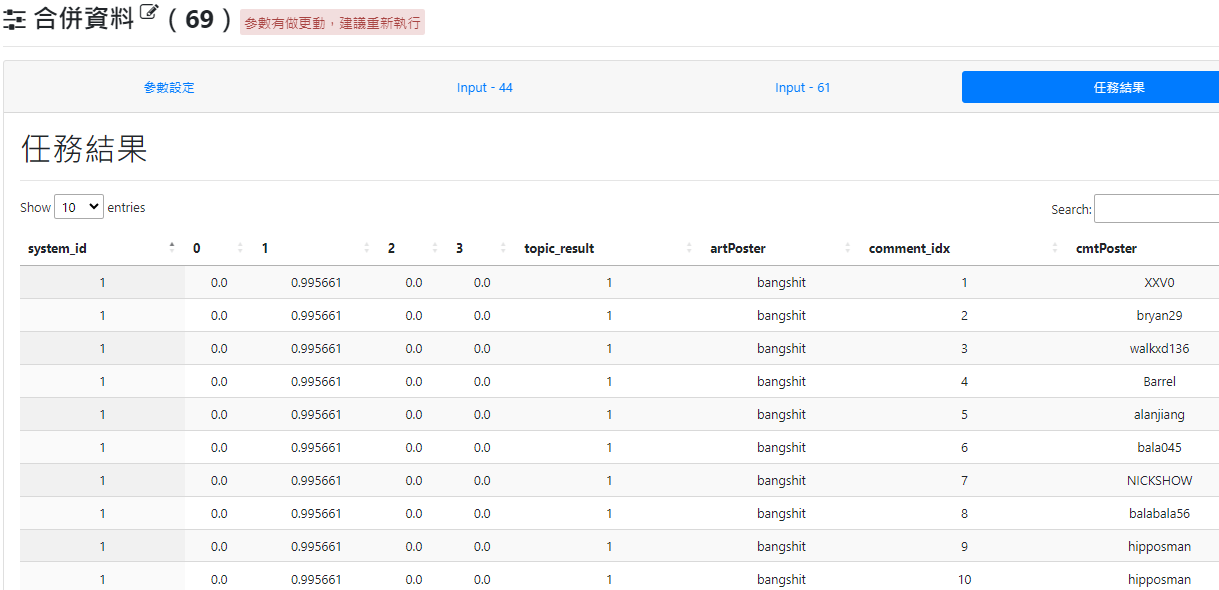




1. 合併資料

再將處理社會網路圖的合併來的資料，再次與新增欄位(idxmax/min)(44)的資料合併，以取得所有文本的推定主題類別。

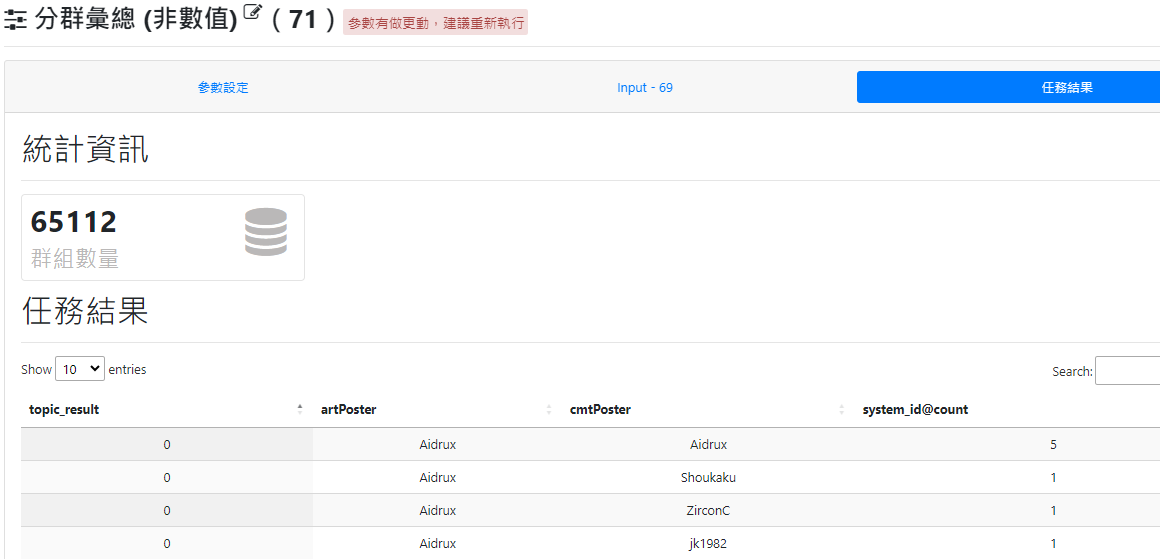




1. 分群彙總(非數值)

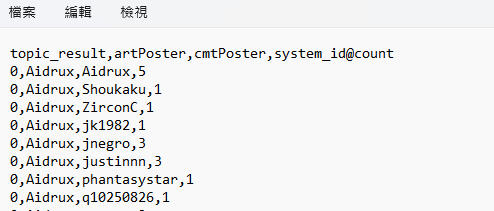
最後，本組以(topic\_result, artPoster, cmtPoster)作為鍵值，進行分群彙總，計算count(system\_id)，找出貼文者與留言者間有多少留言數。

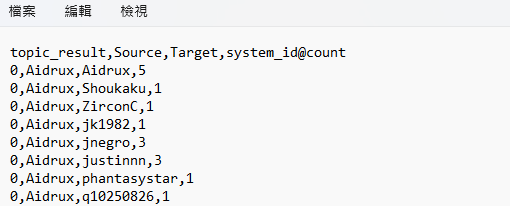




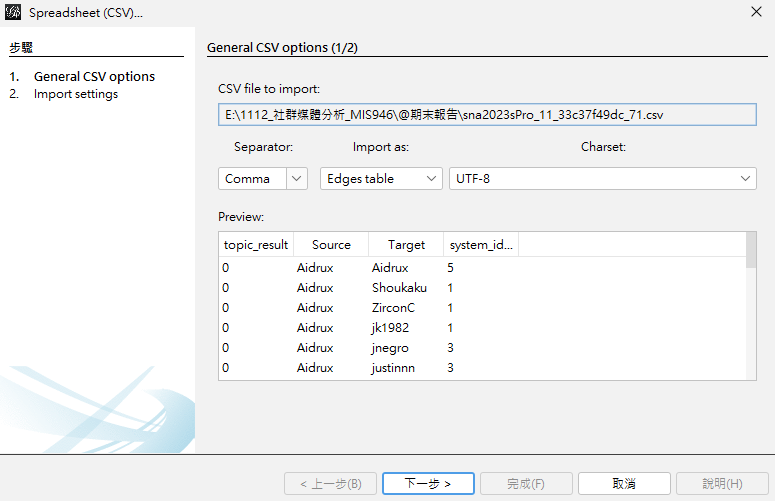
1. Gephi繪圖
2. 前處理

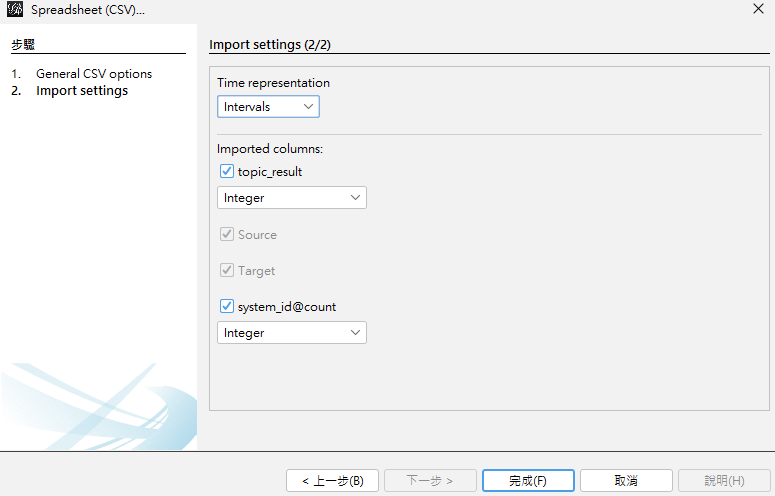
將上步驟所得csv匯出，並修改表頭名稱，artPoster→Source；cmtPoster→Target。

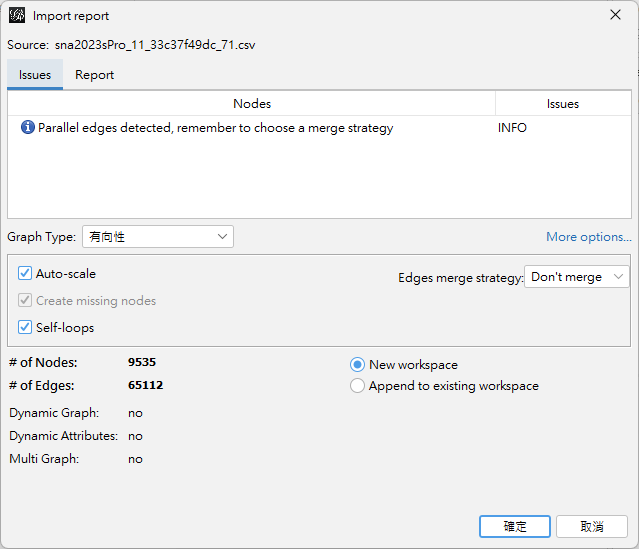




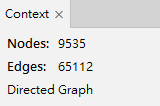
1. 匯入檔案



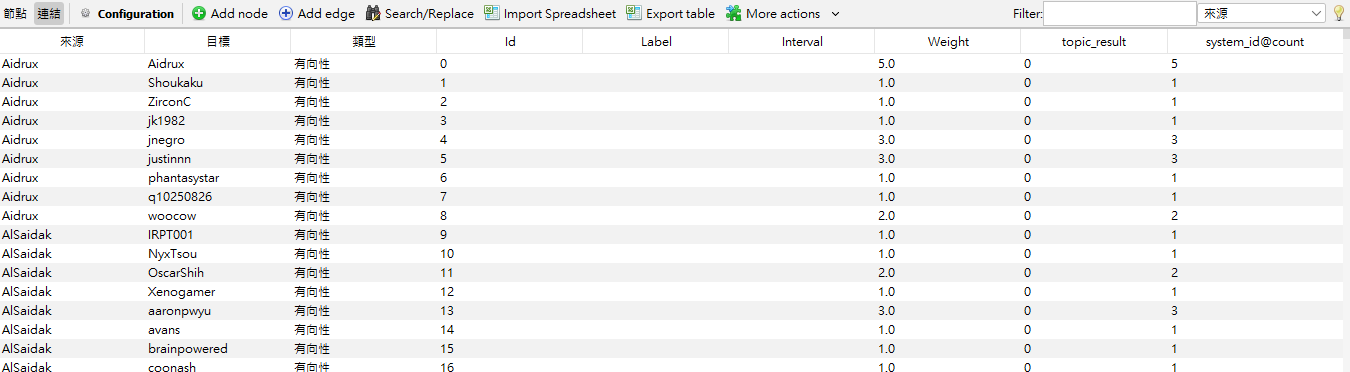




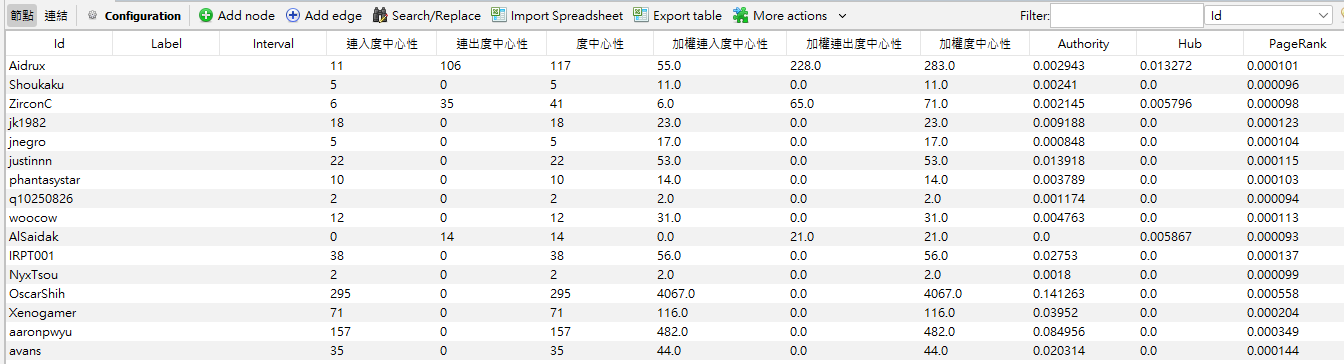
1. 匯入資料共有9535個節點(Nodes)與65112個邊(Edges)。



1. 將system\_id@count複製到Weight欄位。



1. 利用統計(Statistics)功能計算各項數值，包含平均度中心性degree centrality、平均權重程度(Avg. Weighted Degree)、PageRank、HITS等。



1. 找出群體內主要貼文者
   1. 本組使用篩選(Filters)功能，將連出度中心性(Out-degree centrality)設定大於851，繪製圖如下：

一張含有 文字, 字型, 行, 數字 的圖片

自動產生的描述 一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

一張含有 行 的圖片

自動產生的描述

* 1. PlayStation(PS遊戲機)看板主要貼文者，有chirex、kokokko416。

1. 找出群體內主要留言者
   1. 本組使用篩選(Filters)功能，將連入度中心性(In-degree centrality)設定大於217，繪製圖如下：

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述 一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

一張含有 行, 圖表, 繪圖 的圖片

自動產生的描述

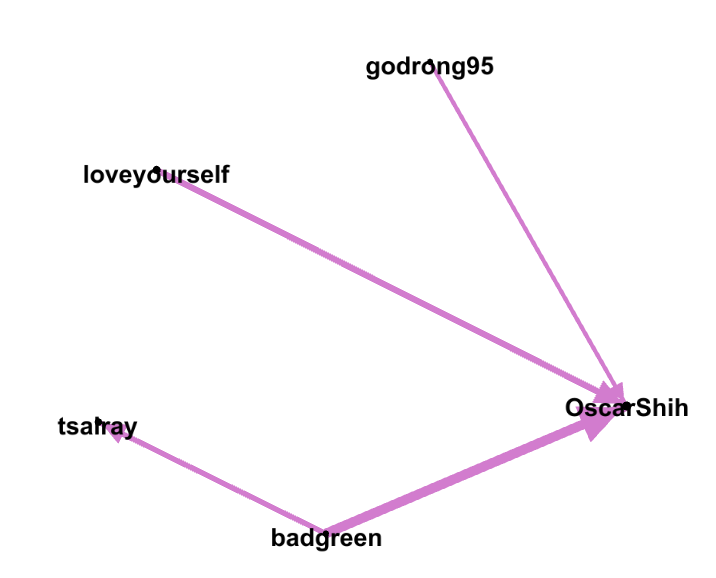
* 1. PlayStation(PS遊戲機)看板主要留言者，有OscarShih、phix、hipposman、CloudVII。

1. 找出群體內主要意見領袖
   1. 本組使用篩選(Filters)功能，將PageRank範圍設定大於3.355692121582743E-4，以及system\_id@count大於30，繪製圖如下：

一張含有 文字, 螢幕擷取畫面, 字型, 陳列 的圖片

自動產生的描述 一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述



* 1. 本組追朔原始檔案，可發現ID為OscarShih的留言者與其他貼文者(如loveyourself、badgreen、godrong95)互動頻繁，連結程度高。另外，這幾位的PageRange值都相當高，可視為PlayStation(PS遊戲機)看板的意見領袖。

一張含有 文字, 字型, 螢幕擷取畫面 的圖片

自動產生的描述

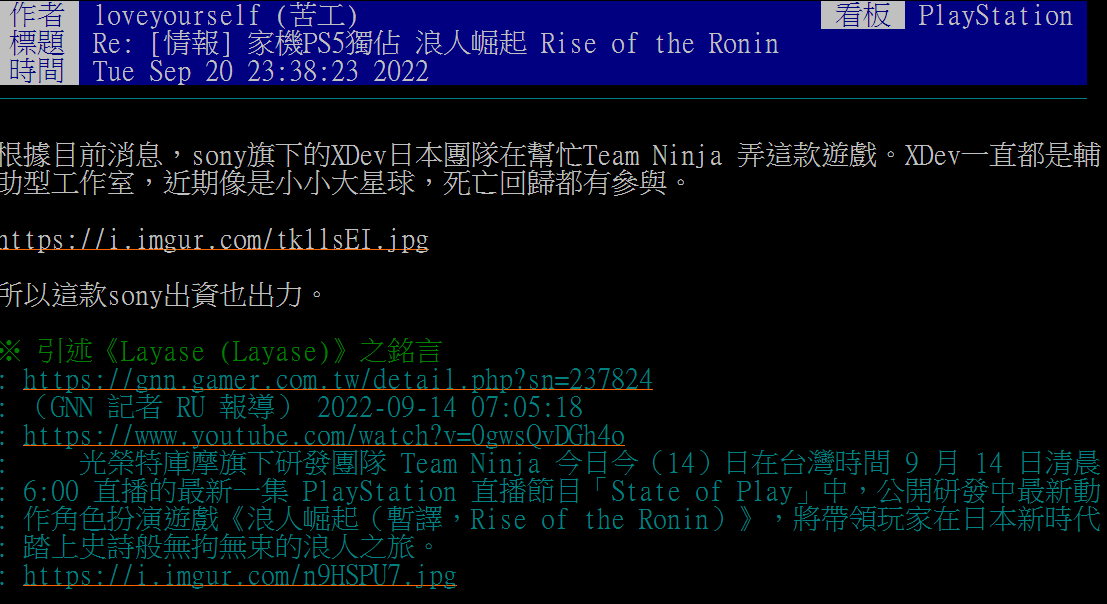
一張含有 文字, 字型, 螢幕擷取畫面, 代數 的圖片

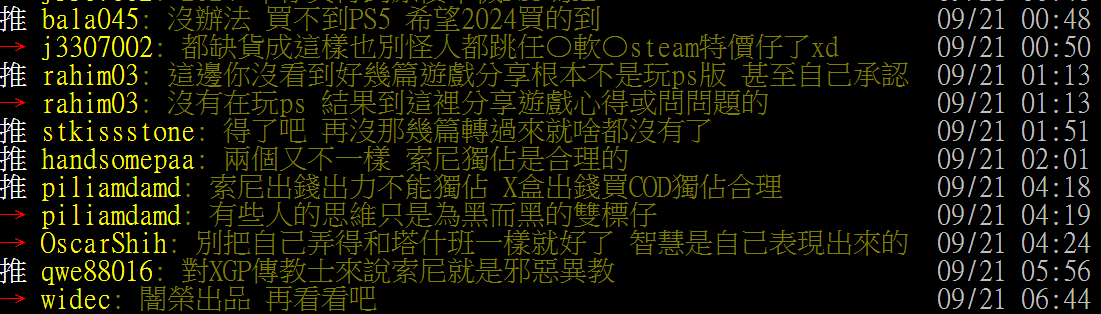
自動產生的描述

一張含有 文字, 字型 的圖片

自動產生的描述

* 1. 如loveyourself於2022/09/20的貼文，OscarShih就曾留言討論。





# 結論

透過本次課程，我們學習到......

## 分類模型訓練

1. Train & Test + Logistic Regression獲得較佳準確度。
2. 前置作業(斷詞+清除停用字)使rawdata越乾淨，模型訓練出來的精確率會越高。
3. 模型隨時間推移，準確度越來越低。

## LDA分類模型

使用LDA能夠很快速的分類主題，也讓我們體驗到文字平台的強大。

## 社會關係網路圖

使用社會網路圖，讓我們透過視覺化圖表，能快速找到具有中心性的KOL。