



誰識KOL?

2020台灣總統大選在 Facebook 上的社群網絡分析

游騰林

202009

開始前來點小調查

專案 PM



- 你的產品有哪些客戶?
- 怎麼找出重要的客戶?
- 他們是怎麼看待產品?

分析師



- 社群網絡分析是什麼?
- 要怎麼分析文本資料?
- 如何結合SNA跟NLP?



游騰林 數據分析師



東華大學 社會所



國泰世華 數據部



TLYu0419@gmail.com



<https://github.com/TLYu0419>



By collaborating with Cathay's subsidiaries and consolidating the group resources, DDT provides innovative digital products and services driven by data with brilliant customer experiences, aiming to become the top in Asia Pacific Fintech ecology.



國泰金控
Cathay Financial Holdings

國泰產險
國泰證券
國泰創投
國泰投信
國泰人壽
國泰世華

時間安排

介紹

- 專案目標
- 使用套件
- 資料結構

SNA

- 如何評估重要性?
- 找出 KOL
- 視覺化

NLP

- NLP 基礎
 - LDA 分析
 - 視覺化
- SNA + NLP
 - 更多應用

總結



「專案介紹 Introduction」

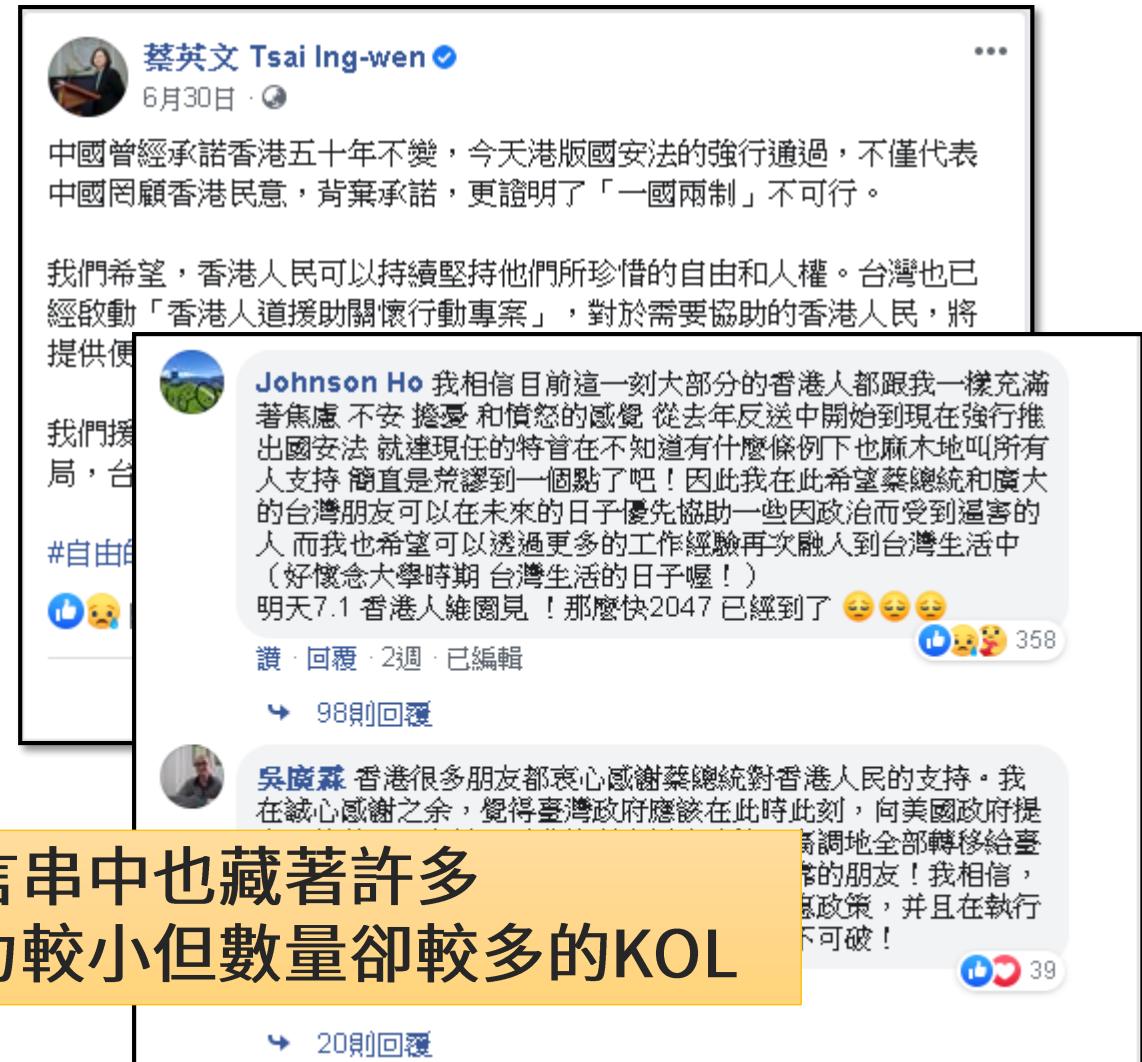
說到 KOL 你會想到誰？

介紹

SNA

NLP

總結



在留言串中也藏著許多
影響力較小但數量卻較多的KOL

專案目標

找出 KOL

透過 SNA 找出位在網絡核心的 KOL

認識 KOL

透過 NLP 認識 KOL 討論的熱門話題

為什麼需要 SNA？

介紹

SNA

NLP

總結

INSIDE 5G AI 新創 評論 焦點 活動 Jobs 好工作 繁 / 簡 訂閱

趨勢
抓到網軍！Twitter 永久刪除來自中、俄、土耳其的 3.2 萬個帳號
2020/06/12 · Heemie · Twitter、網軍、言論、仇恨言論
除了刪掉中國核心網軍的大量帳號，Twitter 也發現有額
軍帳號所發表內容的正當性。

中時電子報 真道理 真愛台灣
新冠肺炎 即時 政治 言論 生活 娛樂 財經
首頁 / 中國時報

網軍帶風向價碼 網站全
04:10 2020/06/26 | 中國時報 | 吳家豪
拍賣網站驚見網軍出售帶風向的服務，且價格全都露。圖為國民黨去年率眾到行政院陳情，要求消滅網軍東廠。(本報資料照)



為什麼需要 SNA ?

介紹

SNA

NLP

總結



Python 套件

介紹

SNA

NLP

總結



Ref:

<https://tlyu0419.github.io/2020/03/17/Crawl-Facebook-Pages>

誰識KOL? 2020台灣總統大選在 Facebook 上的社群網絡分析

Mail: TLYu0419@gmail.com

Python 套件

介紹

SNA

NLP

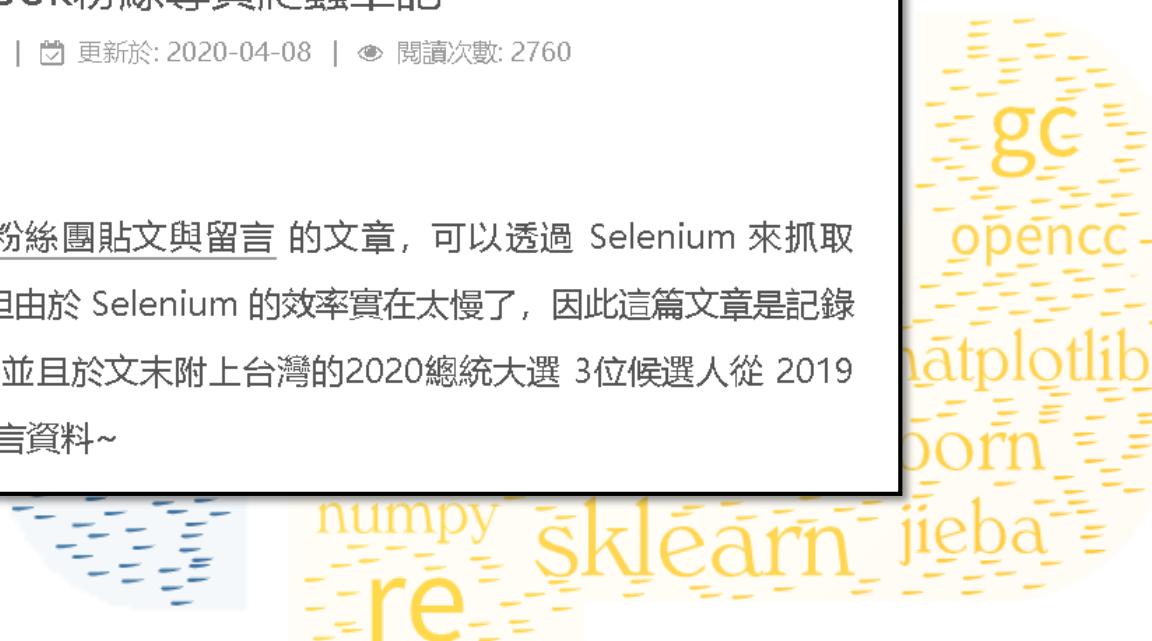
總結



Facebook粉絲專頁爬蟲筆記

發表於 2020-03-17 | 更新於: 2020-04-08 | 閱讀次數: 2760

先前曾發表過 [網路爬蟲_Facebook粉絲團貼文與留言](#) 的文章，可以透過 Selenium 來抓取 Facebook 粉絲專頁的貼文與留言，但由於 Selenium 的效率實在太慢了，因此這篇文章是記錄我嘗試用 request 抓取資料的筆記，並且於文末附上台灣的2020總統大選 3位候選人從 2019 年 1 月至 2020 年 2 月間的貼文與留言資料~



完整的 Python 程式碼如下，有興趣的人請參考
<https://github.com/TLYu0419/FindAndMeetKOLs>

Ref:

<https://tlyu0419.github.io/2020/03/17/Crawl-Facebook-Pages>

誰識KOL? 2020台灣總統大選在 Facebook 上的社群網絡分析

Mail: TLYu0419@gmail.com

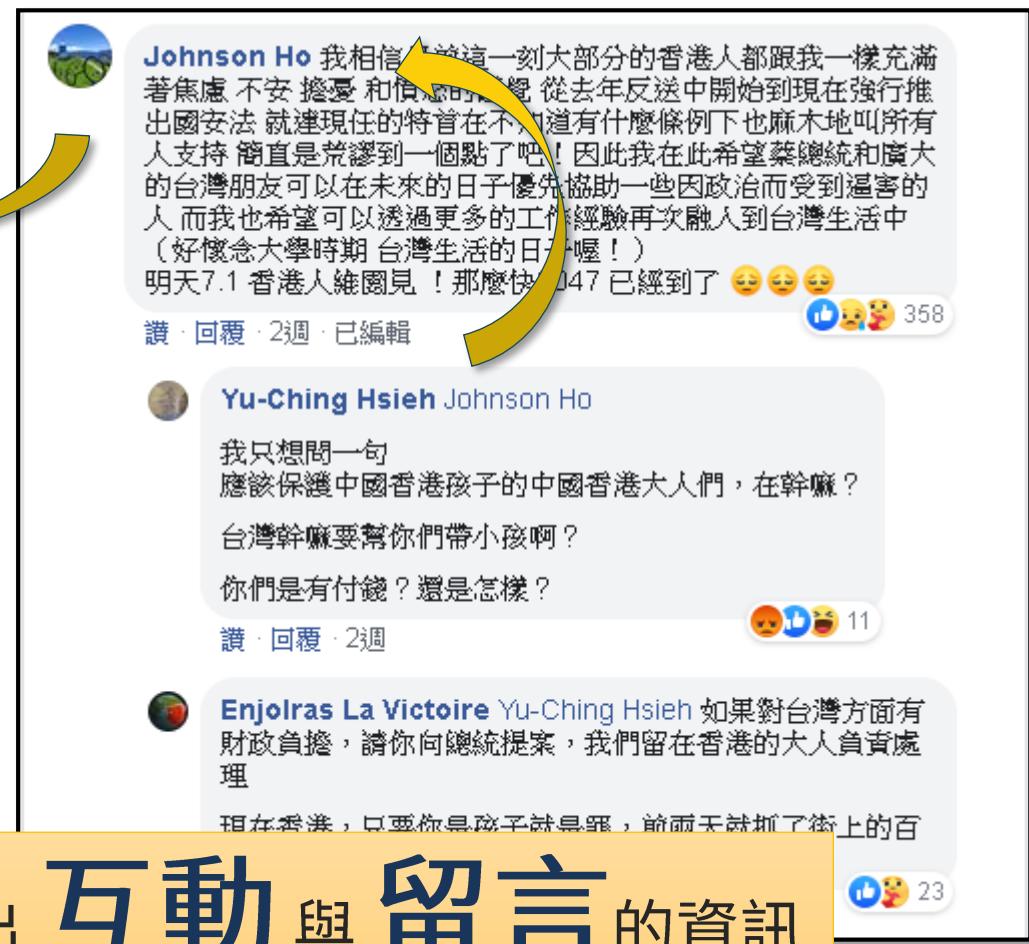
Facebook 上有哪些資料？

介紹

SNA

NLP

總結



Facebook 可以萃取出 **互動** 與 **留言** 的資訊

資料結構

介紹

SNA

NLP

總結

互動資料表：559,763 rows * 5 columns

SOURCE_NAME	SOURCE	TARGET_NAME	TARGET	WEIGHT
Bill***	1471*****	蔡英文 Tsai Ing-wen	4625*****	1
Ash***	5673*****	蔡英文 Tsai Ing-wen	4625*****	1
Billy***	1000*****	蔡英文 Tsai Ing-wen	4625*****	1
Vicky***	1660*****	蔡英文 Tsai Ing-wen	4625*****	1

留言資料表：1,865,409 rows * 5 columns

NAME	AUTHOR	TIME	TEXT
謝**	1000*****	2020-01-01 20:15	松山到了，宋上時刻到了！
Lai***	1000*****	2020-01-01 16:13	棄草棄菜挺宋，選賢與能愛台灣，台灣要蛻變...
Bruce***	1000*****	2020-01-01 22:45	唉，現在的政治卡小 藍綠白黃，都是個鳥樣. ..
楊***	1000*****	2020-01-04 20:04	韓市長加油 告到底 老虎不發威當成病貓 可惡...

註：時間範圍為大選前2個月(2019/11/1 ~ 2020/1/10)

誰識KOL? 2020台灣總統大選在Facebook上的社群網絡分析

Mail: TLYu0419@gmail.com



「社群網絡分析」

Social Network Analysis

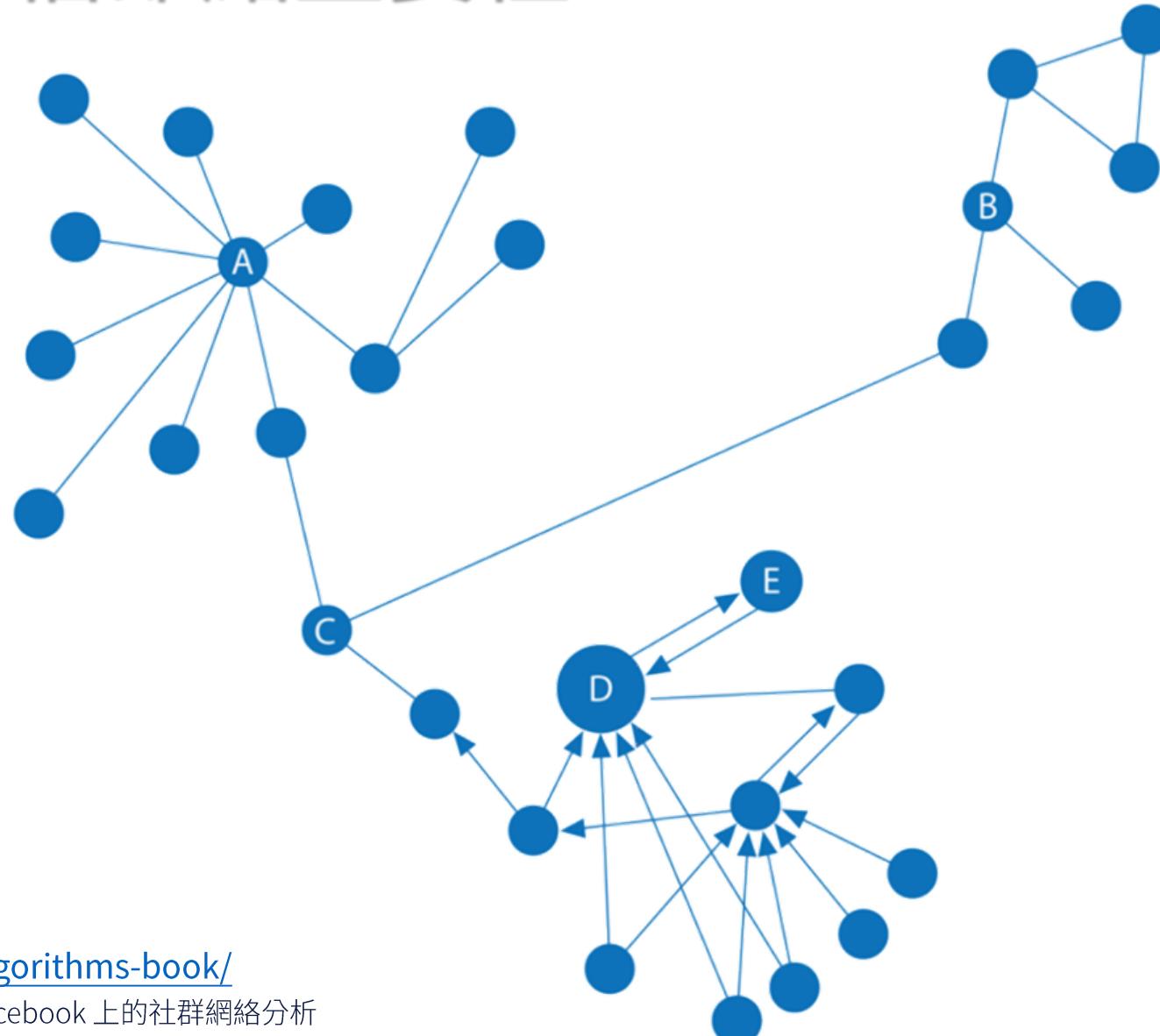
如何評估節點重要性?

介紹

SNA

NLP

總結



Ref:

<https://neo4j.com/graph-algorithms-book/>

誰識KOL? 2020台灣總統大選在Facebook上的社群網絡分析

Mail: TLYu0419@gmail.com

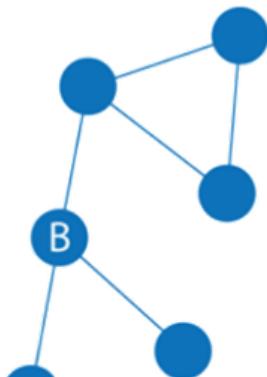
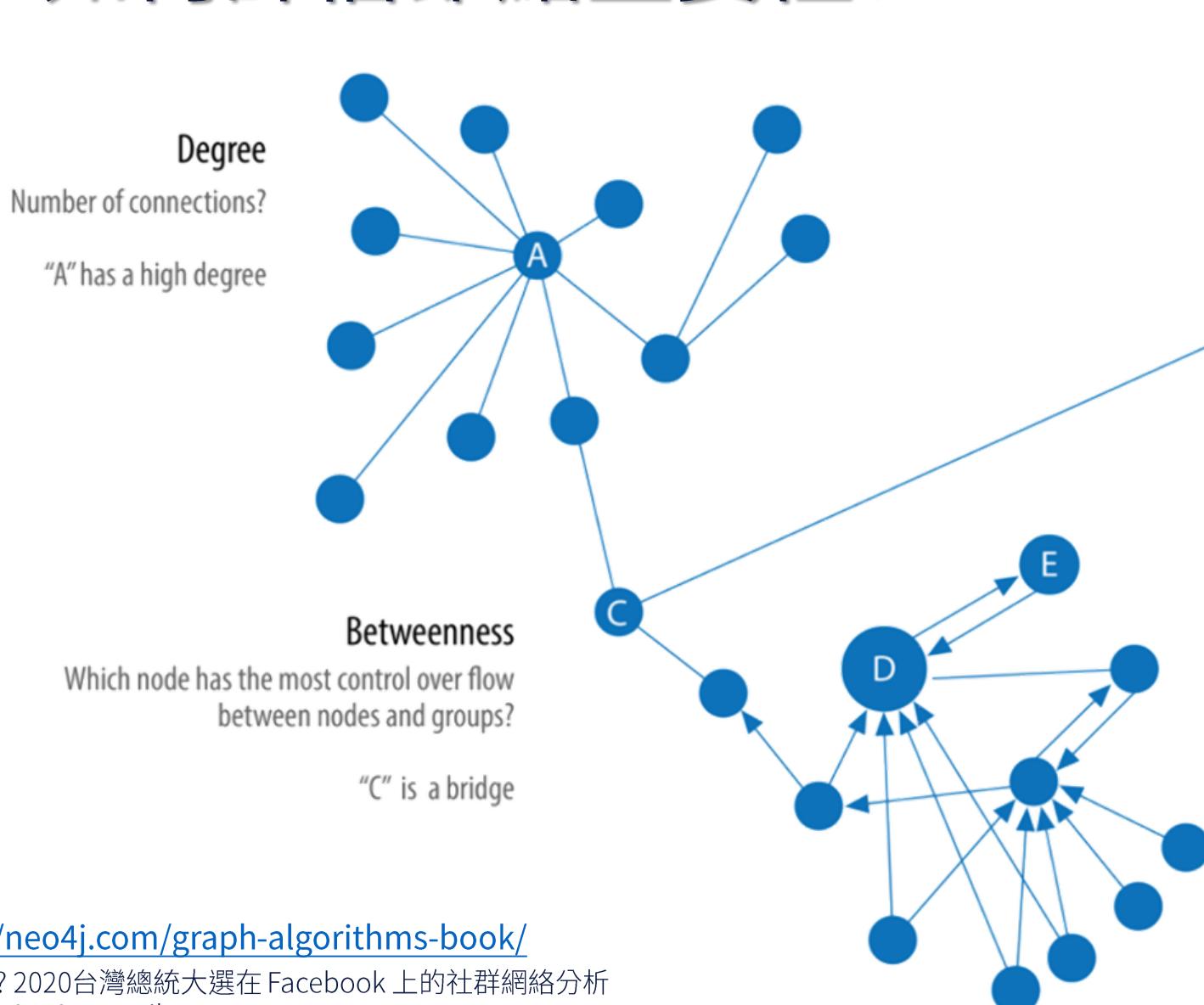
如何評估節點重要性?

介紹

SNA

NLP

總結



Closeness
Which node can most easily reach all other nodes in a graph or subgraph?
"B" is closest with the fewest hops in its subgraph

PageRank
Which node is the most important?
"D" is foremost based on number & weighting of in-links
"E" is next, due to the influence of D's link

Ref:

<https://neo4j.com/graph-algorithms-book/>

誰識KOL? 2020台灣總統大選在Facebook上的社群網絡分析

Mail: TLYu0419@gmail.com

Pagerank 簡介

介紹

SNA

NLP

總結



- 計算網頁的重要性
- 想法
 - In-link越多越重要
 - 來自重要網頁的引用較重要

如果是不重要的節點(假帳號)
藉此創造出的互動也不會有價值

Ref : <http://jpndbs.lib.ntu.edu.tw/DB/PageRank.pdf>

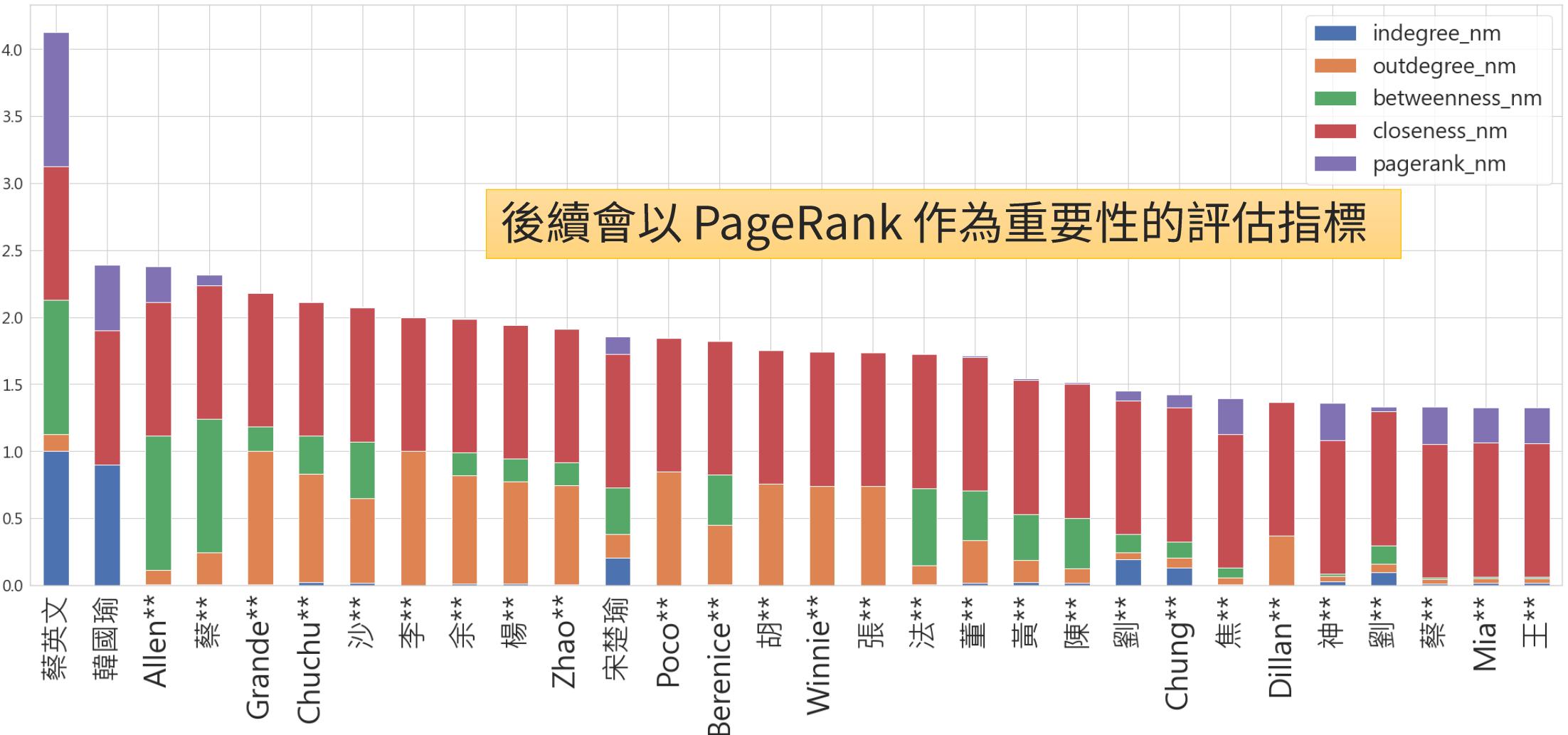
誰是網絡中最重要的人？

介紹

SNA

NLP

總結



SNA 視覺化

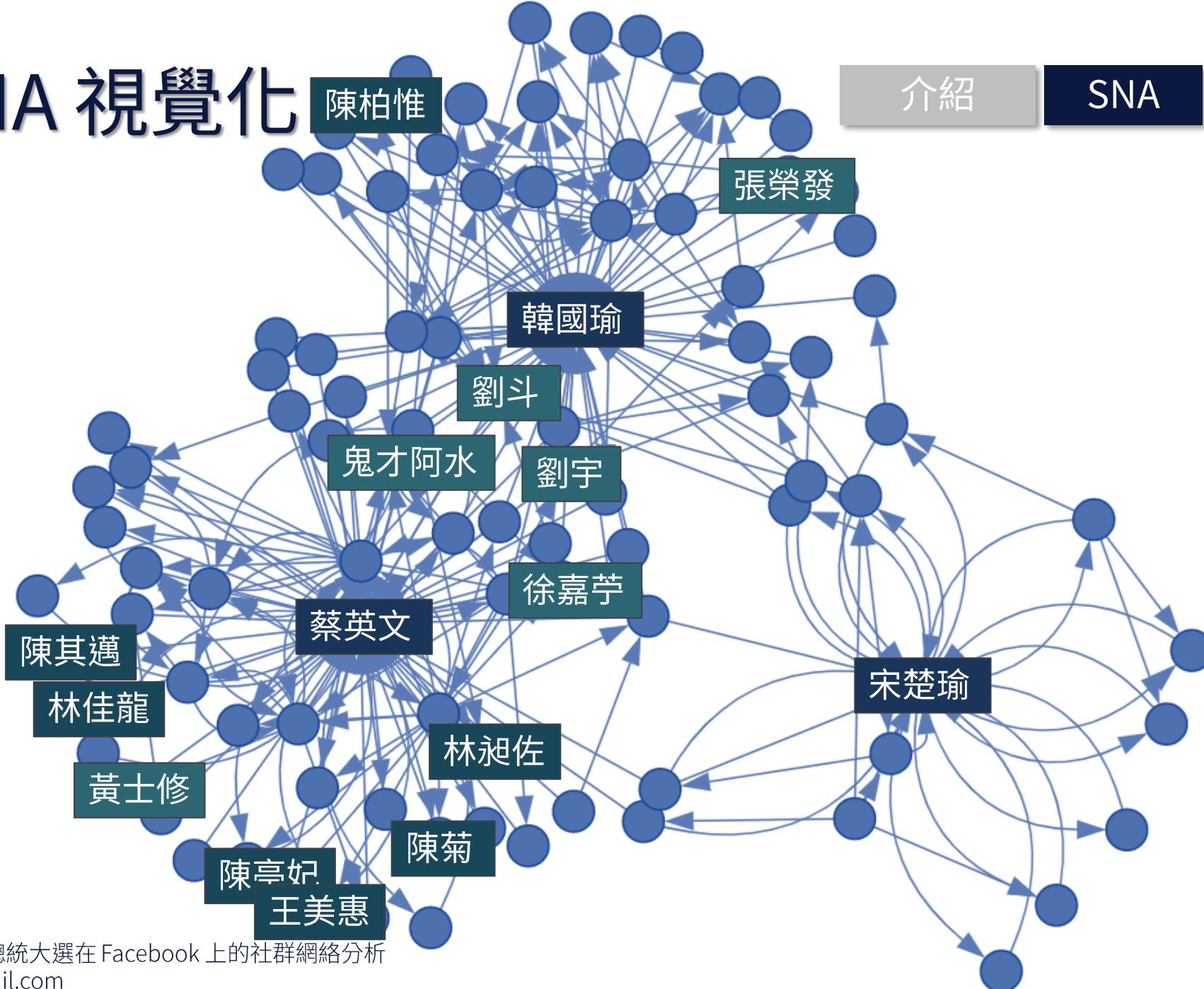
陳柏惟

介紹

SNA

NLP

總結



SNA 視覺化

陳柏惟

介紹

SNA

NLP

總結



陳柏惟 高雄失去了市長 1,074



陳柏惟 「難民法」！所以台灣跟中國一邊一國惹！謝韓韓支持台獨！
對了，知道難民法是國對國就好



SNA 視覺化

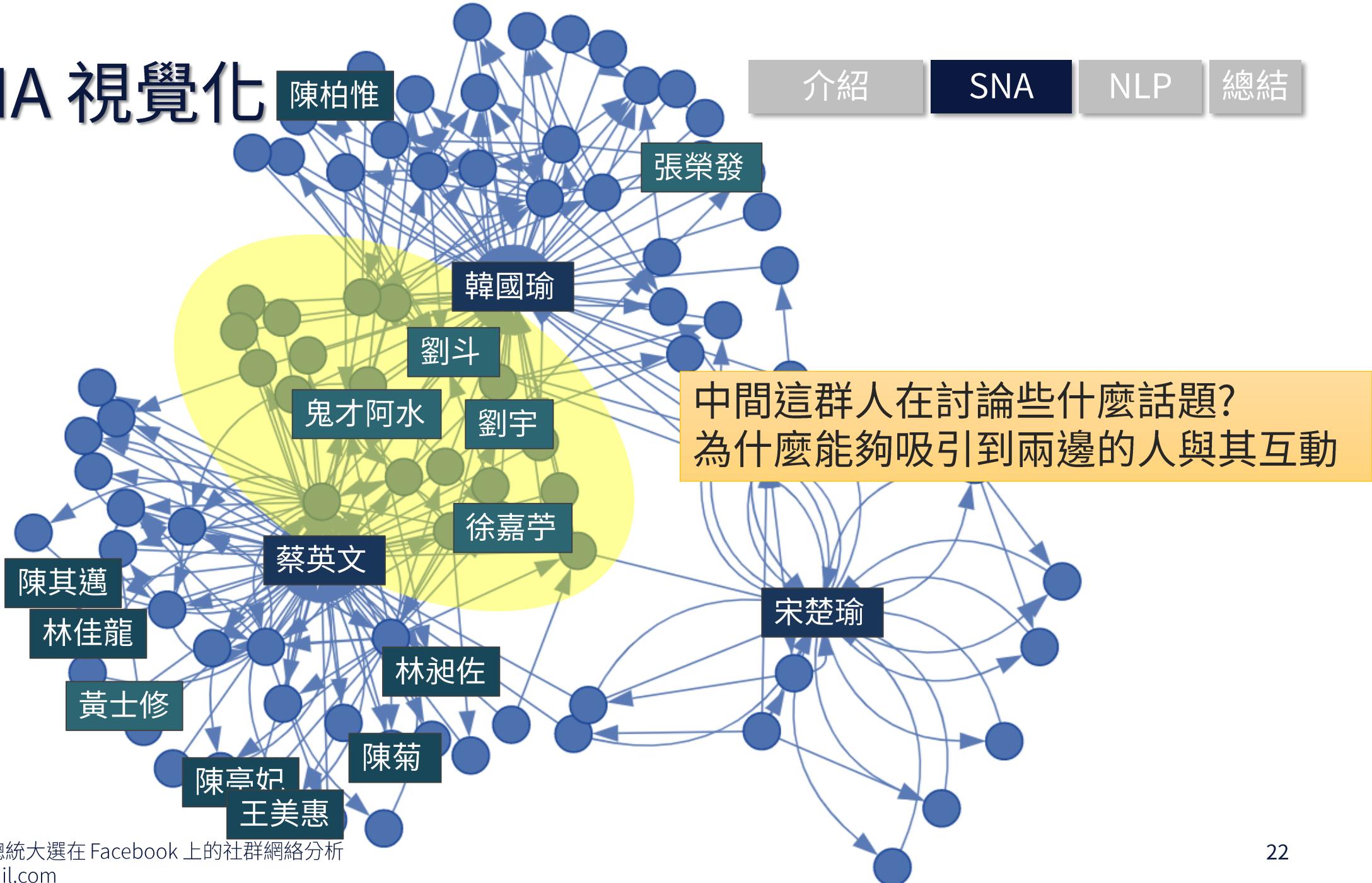
陳柏惟

介紹

SNA

NLP

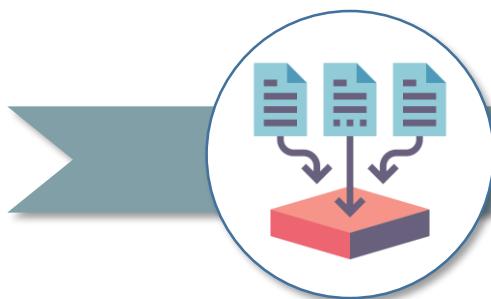
總結





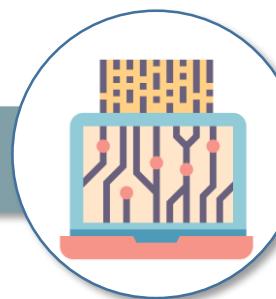
「自然語言處理」

Natural Language Processing



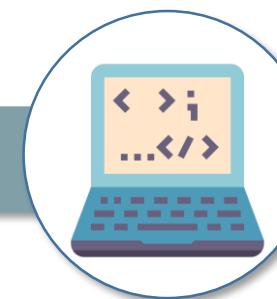
收集語料

- 社群網站
- 新聞文本
- 商品評論
- ...



預處理

- 同義詞
- 停用詞
- 特殊符號
- 斷詞
- ...



特徵工程

- 詞頻矩陣
- TF-IDF
- 詞向量
- ...



訓練模型

- 文本分類
- 情感分析
- 主題模型
- ...

LDA-主題分析

介紹

SNA

NLP

總結

AUTHOR	TEXT3
100000319616052	民主 當飯 民主 連飯 小英 點到 苗栗 ...
100002013267977	清白 檢驗 民眾 遊行 遮羞布 發言 候選...
100013992199086	治國 下台 韓投 定了 韓會 乾淨 茶滿 ...
706072435	習大大 兒子 過去 往前 高雄 不可能 豪...
100002862989012	下台 論文 矮油 加油 下台 論文 頭殼 ...
100000239611720	樹菊 阿嬤 台灣 良心 認證 庶民 加油 高...
100001876335501	支持 菜陰魂 出兵 解放 香港 嘴砲 護香...
100001601896453	翻轉 下架 小英 全民運動 下架 小英 通...
100001520370048	宋伯伯 加油 支持 經驗 政治 頭腦 裡面...
100000404266312	準備 出征 特權 盜採 砂石 質疑 出一 ...
744679351	韓總 香港人 支持 支持 香港 暴徒 所作...
100000990153827	高雄 天上 星星 不一定 高雄 四處 奔波...

帶職參選

烙跑 溜之大吉 草包 高雄

博士論文

英國 倫敦 博士 論文

國家認同

親中 共匪 統治 子子孫孫

.....

....

LDA 分析技巧

1. 去除重複
2. 轉成長文本
3. 限制詞頻

Ref :

<https://www.machinelearningplus.com/nlp/topic-modeling-python-sklearn-examples/>

文本主題結果

政治做出立委
輔助投資
過半選擇團結
蔡總統小英連任
支持發展
台灣人選舉民進黨
選舉國民黨
剩下政黨
聯合政府
距離
經濟
谷草包
站出來
國家台灣人民

市民政策 當選潘恆旭 笑死政府 只有
高雄 中國 韓粉 支持 國家
票投 高雄人 台灣人
國民黨 年輕人國瑜 新莊 發大財
落跑 刑法 罷免 民調
香港 廢除 市長 選舉 民進黨
吳斯懷

介紹

SNA

NLP

總結

文本主題結果

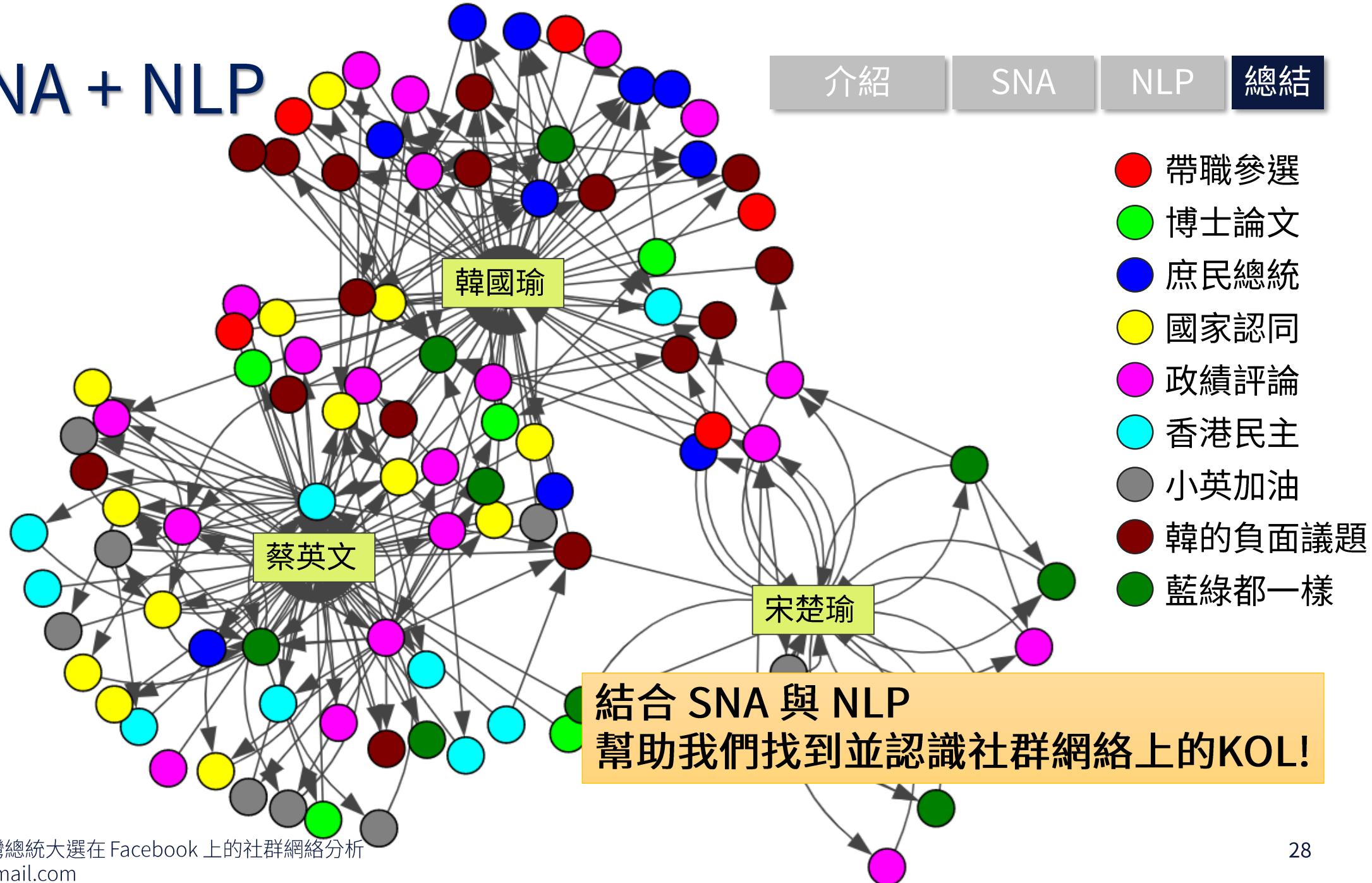
介紹

SNA

NLP

總結





SNA 的更多應用

介紹

SNA

NLP

總結



社群

找到並認識社群 KOL



客服

分析完整的服務旅程



電信

重新評估客戶的價值



電商

抽獎活動的中獎機率



金融

偵測理專的異常金流



其他

其他更多應用

總結

介紹

SNA

NLP

總結

專案 PM



- 你的產品有哪些客戶?
- 怎麼找出重要的客戶?
- 他們是怎麼看待產品?

分析師



- 社群網絡分析是什麼?
- 要怎麼分析文本資料?
- 如何結合SNA跟NLP?