交大宅宅森77

靠北交大貼文分析

Group 7
0316213 蒲郁文 0316238 蔡孟軒 0316248 何鳳雯

Table of Contents

Introduction

靠北交大，交大人不可或缺的佈告欄。

學校發的公告沒人理，靠交上的資訊卻能迅速地傳達給交大人。

Introduction

的佈告欄。

上的資訊卻能迅速地傳

Introduction

Problem Definition

1. 靠北文有什麼常見的主題？

2. 靠北文主題和時間的關係？

Problem Definition

Input：

- 靠北交大貼文資料集

Output：

- 各貼文的主題
- 各主題的時間分佈

Related Works

A Comparison of Document Clustering Techniques (KDD 2000) Steinbach, M., Karypis, G., & Kumar, V.

- K-means
  - better
- agglomerative hierarchical clustering
  - performs poorly
    - since, in many cases, the nearest neighbors of a document are of different classes

Challenges

- Posts on social networking sites are dirty.
- Have a lot of grammar errors, spelling errors, etc.

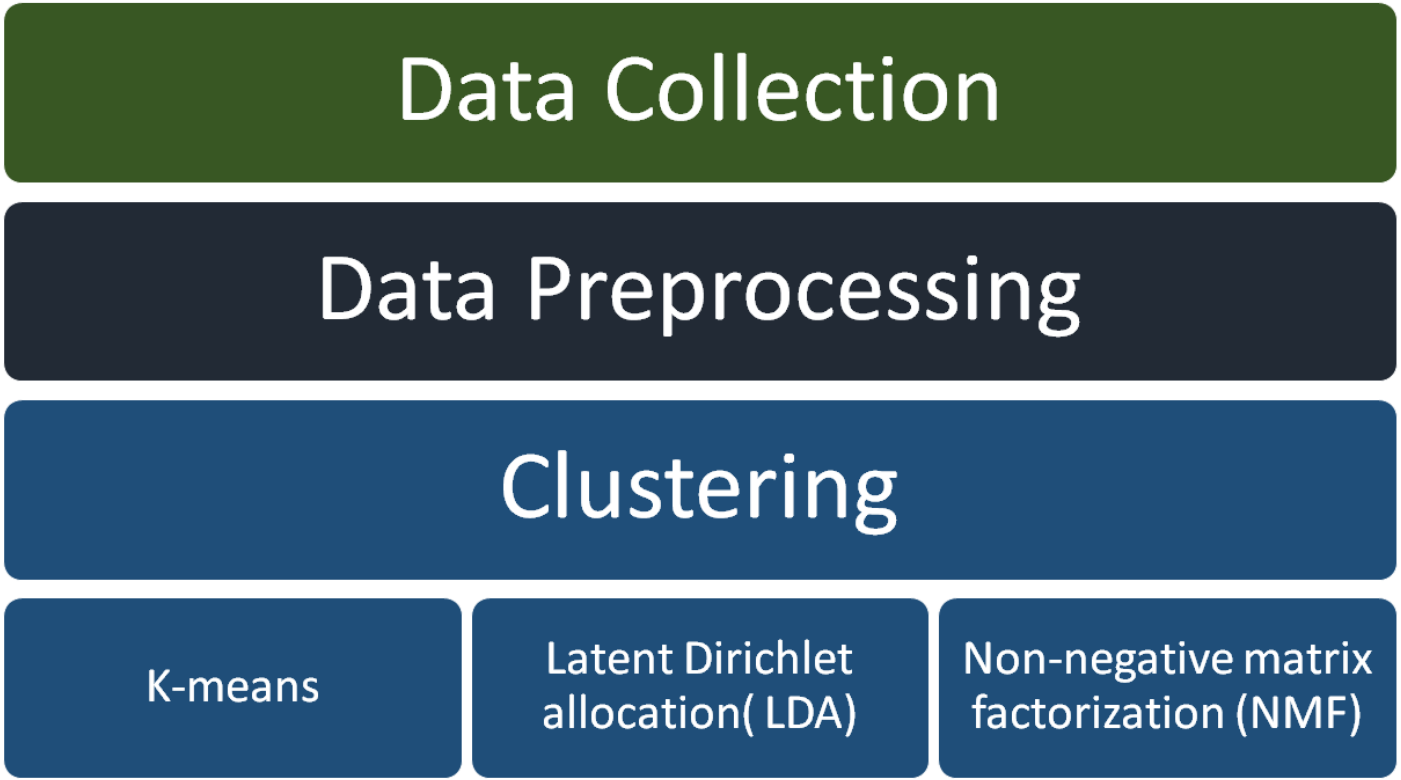
Dataset Descrption

	ID	Link	Message	Created_ time	Story
Data Type	粉專編號_文章編號	連結：分享文、來源文章連結	貼文內容	貼文發布時間	動態描述：分享文、來源粉專資訊
Example	55_94	https://www.fb.com/NCTULIB/photos/a.686796024509631.10737	#靠交49561 圖書館用我的照片都不用講一下的?https://www.fb.com/NCTULIB/posts/1659934950729062	2017-12-05T04:34:12+0000	靠北交大 shared 國立交通大學圖書館's post.

Dataset D

	ID	Link	reated_ me	Story
Data Type	粉專編號_文章編號	連結來源	貼文發布時間	動態描述：分享文、來源粉專資訊
Example	55_94	http com pho 602 737	017-12-05T 4:34:12+000	靠北交大 shared 國立交通大學圖書館's post.

Methods

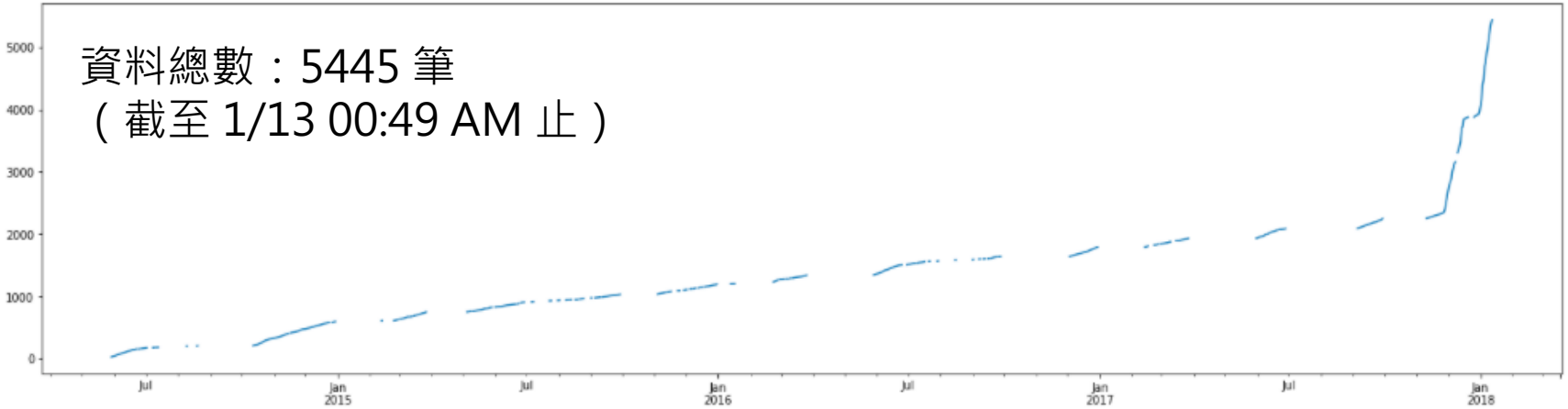

Data Collection
Data Preprocessing
Clustering
K-means
Latent Dirichlet allocation(LDA)
Non-negative matrix factorization (NMF)

Data Collection

不斷透過 Facebook Graph API 蒐集資料

（最後更新時間：2018/01/03 00:49）

Data Observations



累計貼文數量對時間的關係圖

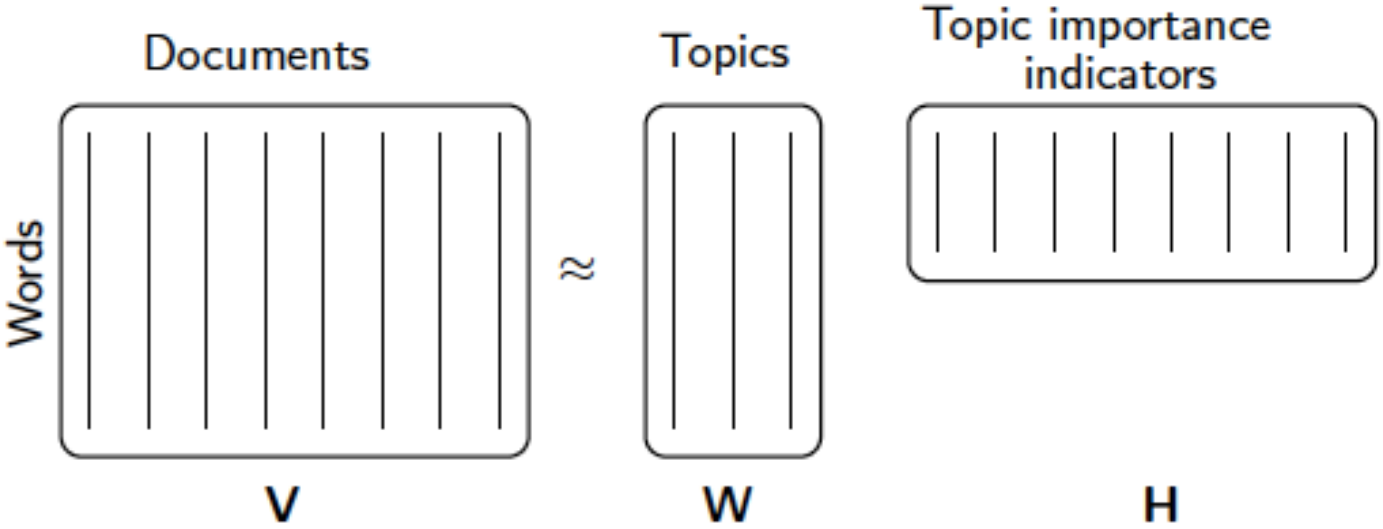
Data Preprocessing

1. Remove posts that have no content.
2. Word Segmentation: Jieba + human intelligence
   (e.g. 梅竹, 13舍)
3. Removing Stopwords: SnowNLP stopword list +
   human intelligence (e.g. 😂, 哈哈哈)
4. Build TF–IDF / TF document vectors.
   (using both unigrams and bigrams; 14275 features in total)

Clustering Methods

1. Hard-Clustering
    a. K-means
2. Soft-Clustering
    a. Latent Dirichlet Allocation (LDA)
    b. Non-negative Matrix Factorization (NMF)

Clustering (NMF)



source: http://www.cnblogs.com/pinard/p/6812011.html

Clustering (LDA)

Topics

gene	0.04
dna	0.02
genetic	0.01
. . .	

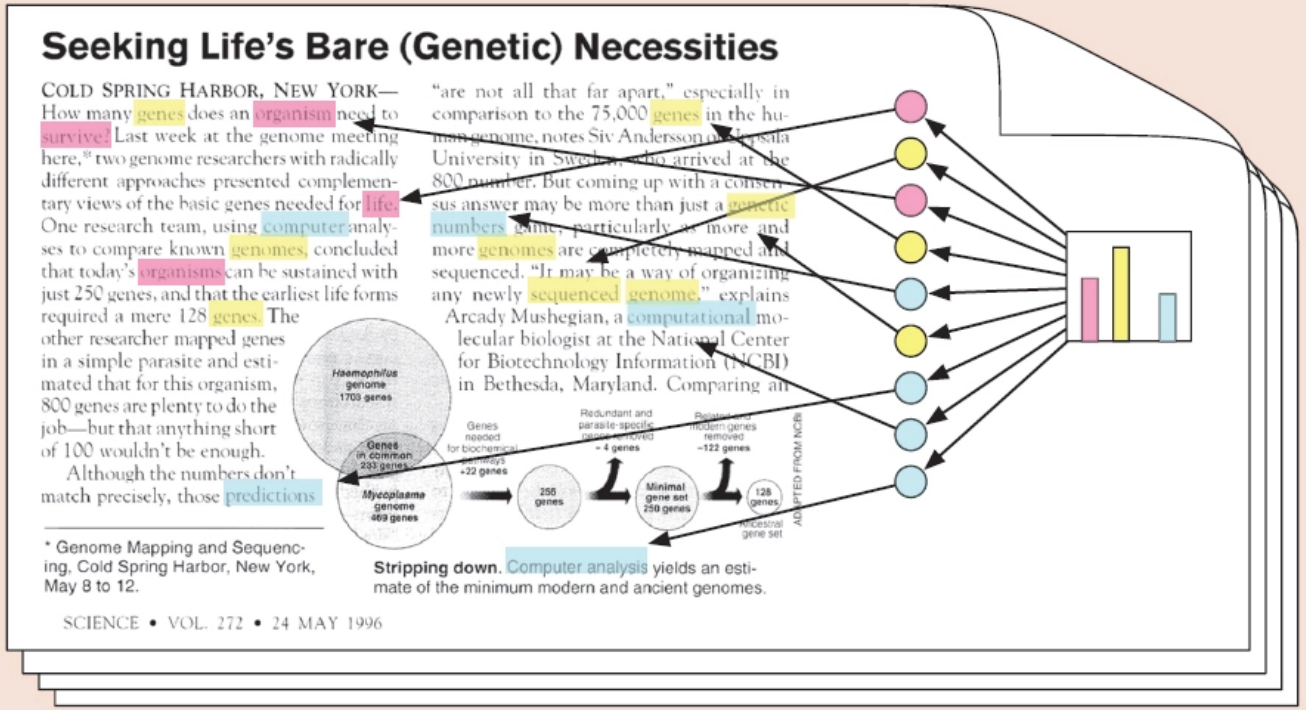
life	0.02
evolve	0.01
organism	0.01
. . .	

brain	0.04
neuron	0.02
nerve	0.01
. . .	

data	0.02
number	0.02
computer	0.01
. . .	

Documents

Topic proportions and assignments



source: http://deliveryimages.acm.org/10.1145/2140000/2133826/figs/f1.jpg

Experimental Results

Experimental Results (NMF)

Topic 0: 2592

別人 / 講 / 尊重 / 事 / 只會 / 討厭 / 懂得 / 大聲 / 讀書 / 別人 讀書 / 感到 / 整天 / 衣服

Topic 6: 102

助教 / 學店 / 分數 / 課 / 學生 / 教授 / 作弊 / 考卷 / 可愛 助教 / 題目 / 成績 / 安安 / 跪求

Topic 8: 156

吃 / 吃 吃 / 滷味 / 蔬菜 / 路 / 二餐 / 好吃 / 滷味 吃 / 我加 / 面 / 一餐 / 夜市 / 姐妹 / 女二

Topic 11: 120

梅竹 / 清 / 梅竹 清 / 必勝 / 倒數 / 比賽 / 梅竹賽 / 女籃 / 宣傳 / 清大 / 梅竹 倒數 / 清 女籃

Topic 12: 178

考試 / 作業 / 寫 / 寫 作業 / 交 作業 / 期末考 / 交 / 考試 作業 / 期中考 / 上課 / code / e3

Experimental Results (LDA)

Topic 3: 135

臺灣 / 電影 / 路上 / 幸福 / 中 / 幸福 路上 / 社團 / 故事 / 希望 / 手機 / 透過 / 學生 / 領袖

Topic 4: 210

辦 / 活動 / 營隊 / 系學會 / 學妹 / 梅竹 / 參加 / 社團 / 妳 / 女生 / 找 / 系 / 明明 / 學校

Topic 7: 180

吃 / 狗 / 我會 / 留言 / 宿舍 / 感謝 / 同學 / 撿 / 咬 / 私訊 / 怕 / 留言 我會 / 希望 / 樓

Topic 10: 147

倒數 / 助教 / 題目 / 學生 / 同學 / 時間 / 2de / ICLAB / 期末 / pattern / project / 梅竹 / 學校

Topic 16: 184

系統 / 妳 / Windows / 內容 / 目標 / 環境 / 型別 / 配置 / 攻擊 / 程式設計 / 方法 / 原理

Experimental Results (K-means)

Topic 7: 137

文 / 聽說 / 告白 / 搬運 / 搬 / 帥哥 / 搬運工 / 哥 / 貼文 / 系 / index / 搬 文 / 屁孩 / 🙌

Topic 14: 148

找 / 系學會 / 明明 / 學妹 / 朋友 / 辦 / 時間 / 社團 / 幫忙 / 活動 / 還要 / 學長 / 營隊 / 逆

Topic 15: 72

梅竹 / 清 / 比賽 / 梅竹賽 / 倒數 / 必勝 / 梅後 / 梅竹 倒數 / 宣傳 / 梅竹 清 / 明年 / 清大 / 號

Topic 19: 106

助教 / 微積分 / 寫 / 掛 / 考試 / 作業 / 物理 / 微積分 大會考 / 物理 微積分 / 題目 / 分數

Topic 24: 90

八卦 / 畢業 / 實驗室 / 門檻 / 畢業 門檻 / 畢業 找 / swap 八卦 / swap / 女友 / 找 / 聽說

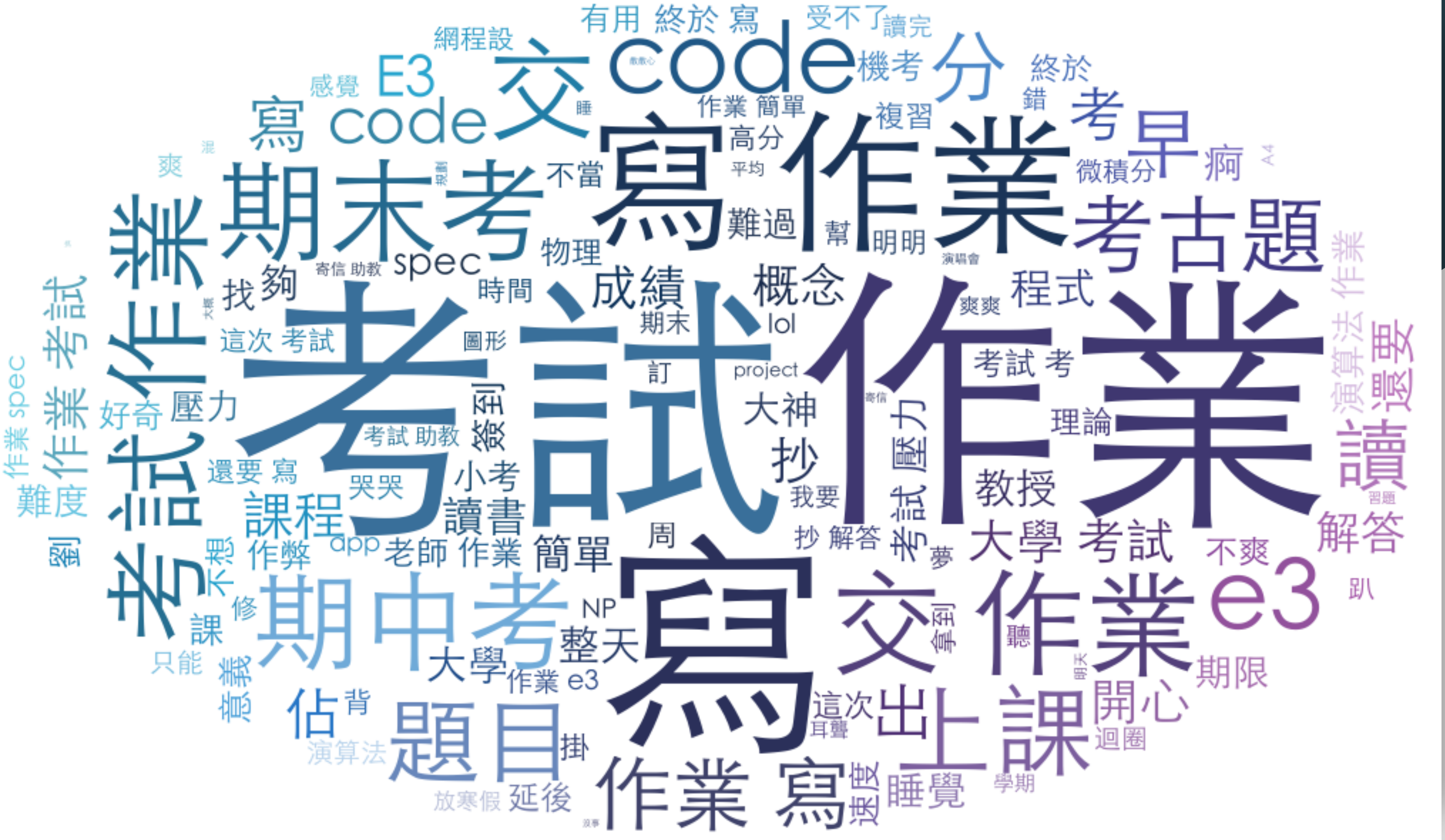
Results Visualization (NMF)

室友
昨天
異性戀
室友 吵
不好
聽 分享
自言自語
消夜
我會
半夜
想到
室友 室友
大便
考
發現 室友
鬧鐘
宿舍
整天
住
陰莖
買
幫 室友
怎辦
機
超差
約會
興奮
吹
回 房間
轉頭
室友 朋友
讀書
靠杯
三人
洗澡
偷
女朋友
救救
發現
回
房
唱歌
凝視
室友 睡覺
沒
快世界
期末
夫妻
唱歌 難聽
住宿
尷尬
事情 想到
難道
室友 鎖
吵到
交了
算
屁
室友 睡
肥宅
睡覺
衣服
小心 室友
小孩
帶
女生 胸部
走
床上
屁 屁
外面
btw
聽
聲音
一整天
冰箱
室友 吃
親愛 室友
好開心
兩個
game
帶 回來
新 室友
房間
女友
晚上
味道
睡著
就算
沒用
愛上
難聽
手槍
室友 男朋友
兩個 室友
講
率
自閉症
肥宅 室友
Tag
長
睡 不好
半小時
吃飯
學期
回來
室友 走
臭
聖誕
其他人
煩惱
不介意
房間 室友
乾脆
廢物
點
期末考
室友 大便
糾正
垃圾
回家
大聲
種種
室友 可愛
男朋友
室友 帶
堅信
不在
衛生習慣
陌生人
晚上
半夜
事情
念
怕
玩
幫
鎖
lol 夠
睡
躺
求解
耳機
好好
我該
親愛 放
排擠
行行好
小心
好醜
跨年
寢室
明明
講話
胸部
期中考
唸書
電動
廁所
一波上課


可愛
男朋友
男朋友 可愛
櫃檯
運管
啦啦隊
蛋包飯
聽說
機會
妹子
你好
女孩
#怪
耶舞
接力
阿西
鑽石
可愛 妳
電資
奈米
名字
水
戴眼鏡
機械
店員 可愛
舔
騷擾
屁股
戀愛
學期結束
可愛 男友
阿西 師
可愛 tag
二餐
笑容
主任
迷人
你好 可愛
學妹
土木系
可愛 女生
室友 可愛
樣子 可愛
皮卡丘
越
倒數
應付
告白
學姊 可愛
認識一下
進場
主持人
姊姊
常
疊字
土木系 名字 李
機
可愛 可愛
師
大叔
心
妳 死會
追
跳
高
妹
死會
婊
打球
梅竹 倒數
號
跳舞
可愛 助教
我家
女神
教
小天使
樣子
電物
馬尾
好奇
為啥
妳 可愛
發
復活
想要
男朋友戴
可惜
馬
男友
女生 可愛
店員
恩
奶廚


吵
讀書
睡覺
大聲
13舍
13舍 樓
寢室
半夜
四樓
整天
吵屁
室友 吵
走路
gogoro
吵到
吵 吵
12舍
十二舍
十舍
走廊
這麼晚
講話
鼻子
滑鼠
房間
醫生
衝三小
電機 資工宿舍
寢室 吵 學校 辦
沒課 還要 吵
閉嘴
早上
晚上
吃飯
期中考
期末考
博愛
老母
平常
音樂
安靜
小聲點
客人
坐在
試試看
檢舉
好不好
竹軒
吉他
星期三
椅子
椅子 吵
樓靠
娘
樓下
混在
進來
放
影響
10舍
門口
回家
二樓
十舍 四樓
嘈音
一群
游泳池
吵 長
醜
就算
吵屁
考試 壓力
登登
笑
膩 點
廢物
辦
三點
狗
浩然
彈 吉他
笨蛋
殺
嘴
聽
快
整晚
還要
林先生
靜音
道
廢
耶舞
小聲
博愛 校區
吵 圖書館
平常 吵
自私
凌晨
敲
黑暗
要死
活動
手指
熄燈
子
房
知
六樓
鬧鐘
坐在 椅子
那群
室友 睡覺
一大早
聲音
校區
折
住
長 醜
辦 活動
睡
講
音量
彈
電機
49427
大叫
室友 朋友
壓力
別系
聚上
小屁孩
可悲
教室
閉上
吵屁
睡覺 讀書
地板
娘 整天
滑鼠 聲
欠 檢舉
講話
大聲
女
加油
怎辦
天花板
馬
懂
一大早 吵
12舍 二樓
大廳
聲
外面
耳機
走廊 大聲
破壞
祝你們
長
整晚
架
某間
手機
e04
麵
孤獨
打字
消失
為啥
閉上 嘴
ni


梅竹
清
必勝
比賽
梅竹 倒數
女籃
梅竹賽
梅竹 清
竹狐是
倒數
宣傳
梅竹 成功
梅竹 竹狐是
直播
瞧不起
成功
號
交清
收拾
清大
清
女籃
加油
男排
擲
梅後
橋藝
閉幕
工具
男排 號
銅板
火力
跑跑
掰掰
貼
啦啦隊
翻
紀念品
支援
昨天
太帥
轉播
熱情
電競
天
硬幣
明年
隔壁
棒球隊
掛
不少
清 不見
打爆
換
連假倒數 女生
清 掛
忘記
梅竹 男排
吃掉
綠色
全體
嗚嗚
後援會
相機
梅竹 19187
火力 班
大學
兩校
商學院
清交
攻陷
排球
難過
擲 銅板 19187
玩
帥
女生 tag
肥宅
不見
記者
嗆 聲
研究生
熱舞
顯眼
梅竹 後援會
班
吹 梅竹
找
吹
下臺
公平
有大
model
夜市
女朋友
體育館
布條
告白
亂



考試
作業
寫
寫
作業
期末考
考試作業
期中考
題目
code
交
code
寫
分
E3
交
作業
e3
上課
作業 寫
考古題
早
考
成績
概念
程式
讀
解答
開心
期限
出
大學 考試
教授
讀書
課程
作業 考試
難度
考試 作業
spec
作業 spec
有用 終於 寫
受不了
讀完
網程設
機考
終於
感覺
作業 簡單
錯
複習
高分
平均
微積分
痾
不當
難過
幫
明明
物理
演唱會
寄信 助教
找
夠
時間
期末
爽爽
lol
這次 考試
圖形
訂
project
考試 考
好奇
壓力
大神
演算法 作業
還要
抄
考試 壓力
簽到
理論
考試 助教
還要 寫
哭哭
小考
我要
作弊
老師 作業
簡單
周
抄 解答
不爽
app
不想
劉
修
課
NP
整天
拿到
大學
作業 e3
只能
意義
估
背
聽
這次
耳聾
掛
演算法
延後
放寒假
速度
睡覺
學期
迴圈
趴
夢
爽
睡
A4
習題


圖書館
網路
圖書館 網路
講話
情侶
吵死
求解
放閃
網路 爛
清大
圖書館 五樓
演唱會
不可思議
力點
爛
垃圾
幾次
煙味
照片
吵 圖書館
用氣音
旁邊
五樓
隨機
活動
聊天
大聲 聊天
滾出去
滑鼠
講
六樓
耳聾
很慢
書
疊
拜託
唸書
安靜
手機
擦屁股
整間
五樓 滑鼠
浩然
為啥
讀書
週末
過期
麻煩
期末考
降
LOL
衛生紙
幾個
周到
討論
草地
洗澡
大聲
一群
回
一首
傘
擦
走
邀
那區
生
男
怕
震
學妹
祝
唱
打字
就算
外面
公共 空間
施工
幾點
圖書館 公共
靜音
拉基
我要
很爛

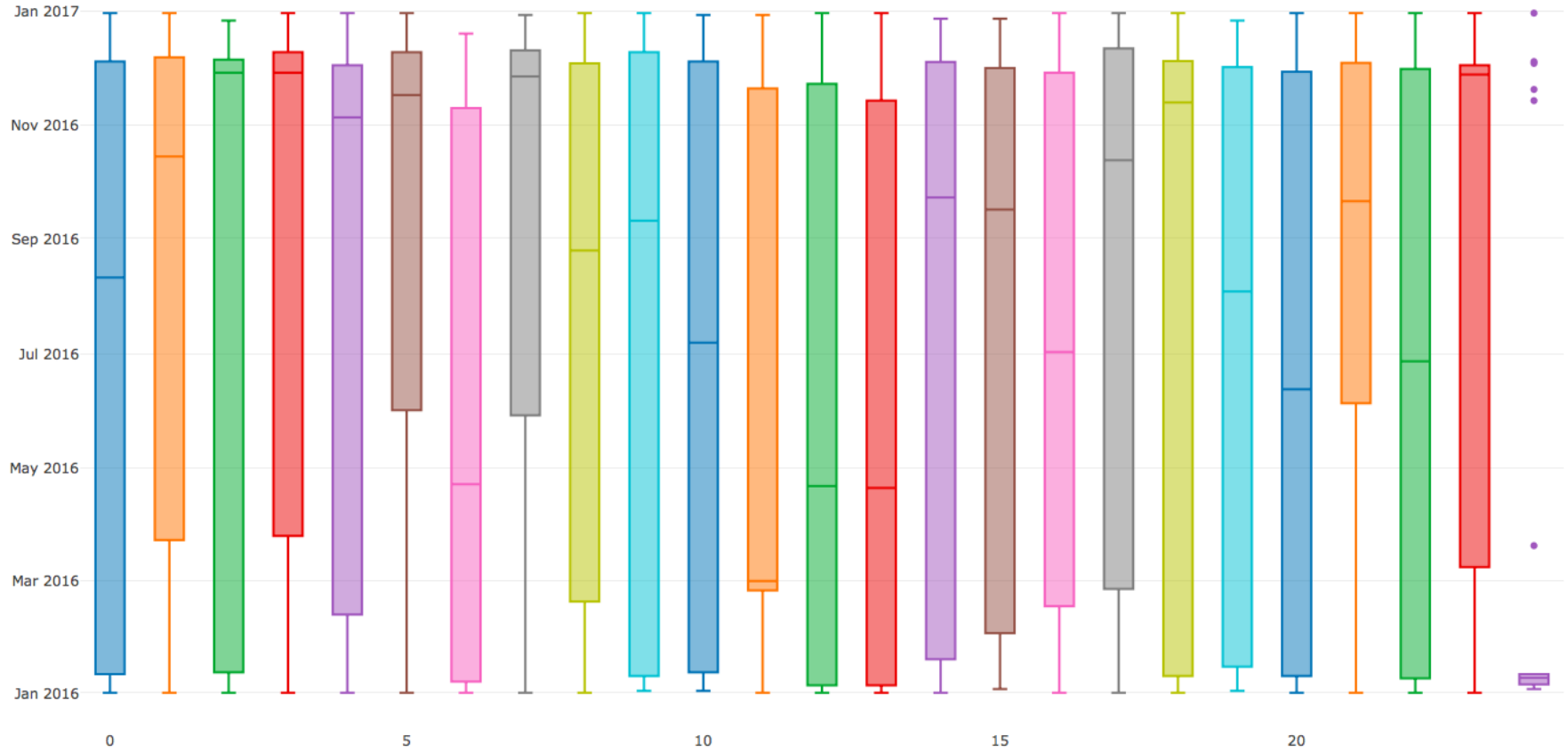

win10
編好 13688
變種
發文
戰
不吃
罵
快
我文
好久
小心
成立
笑
沒梗
龍王
快樂
更新 速度
性侵
小編
男生
封鎖
觀眾
長大
香港
bubble
速度
比不上
辛苦
活著
小編 辛苦
禮拜五
吃 我文
手
吃 文
pub
包庇
電資
昨天
揪 咪
跑
過年
很煩
應徵
私訊
OO
踹
勾結
守口如瓶
戰文
聖誕夜
龍騎士
個資
抓馬
女朋友
謝謝
讚
置頂
夢
不適
忘記
貢獻
謝
編號
共
跟上
踹 共
新 編好
小編 吃
匿名
小編 至少
聖誕節
吃文
靠交小編
新
雞湯
討論
小編 更新
點 睡
聚
粉
一篇
bubble pub
功德
16433
更新
哭哭
小編 文
文
六
不見
至少
平臺
難得
置頂 文

朋友
Tag
Tag 身邊
身邊
身邊 朋友
carry
缺
宅宅
夠 難
男生 朋友
睡不著
嘴
邊緣人
課 朋友
DT
好友
朋友 carry
朋友 他會
男生
找
室友 朋友
笑臉
小聲
吃飯
49879
搭車
揪
句
有時候
女朋友
多麼
整天
Ans
背地裡
感覺
名字
熟
課
生活
姓黃 朋友
很熟
大學
土木系
照鏡子
課堂
姓黃
光粉
半夜 睡不著
順便
臭
有個
選
Tag 各系
朋友 關係
講
XXX
陪
男
載
名字
豪想
睡
選 課
有個 朋友
聊
黑人
傻
開門
夠
偷偷
房間
tag 姓黃
這科
半夜
留言
我該
關係
出
大學 生活
RRR
走
輕鬆
火車站
作業
49286
希望
欠
難
又帥厭世
興趣
看不出
社會


ptt
bbs
ptt bbs
Gossiping
bbs Gossiping
bbs SENIORHIGH
SENIORHIGH
bbs NCTU
NCTU
NCTU TALK
TALK
麥當勞
優質
拒吃
很潮
滾
陶
一首
長
土木
寫個
團購
成效
發了
戰
TO
信徒
19204
一條
禮拜
塊
置頂
國家
3853
52795
中午
置頂文
自由
神父
愛滋
詩
友
文
學弟
113
支援
潑
交上
引戰
E6
獎學金
塊 陶
校園
酸
よ
問卦
新聞
買
便宜
教主
趕快
發現
明天 中午
sex
明天
繼

狗
女二
單身
單身 狗
女二 狗
TAG
tag
身邊
怕 狗
流浪狗
身邊
狗狗
作弊 狗
半夜
半夜 女二
人士
作弊
吠
乱
怕
愛 狗
趕走
走
面前
項圈
清大
清夜
爽
樓梯
謝謝你們
隔壁
外人
阿福
醬油
動物
學生
幹累
老鼠
囉
宿舍區
汪汪
TAG 身邊 49501
49315
攻擊
一群
狗王
抓
大隊
狂吠
社
咬人
電工
閉嘴
清華
幾次
蟑螂
幫
穿衣服
本校
狗 幹累
狗 趕走
九舍 外面
咬
毛
糧
聽
亂 吠
狗 學校
呼籲
夜市
季節
可悲
踩到
不停
學校 狗
念
外面
白色
Ok
進來
六年
單位
蟬
狗本
差點
反擊
九舍
認親
乖
汪
定期
地下室
今晚
吵醒
韓國
跑
交通
路上
雞
計較
女二 狗狗
狗 老
工科
臭
老
快
室內
常駐
長是
狗 圖書館
吵個
汪汪 社了
牠
餵狗
愈 宿舍
社了


中 提到
宿舍
中毒
手槍
救
支援
死忠
希望 光頭
光頭
哥哥
幾篇
回
希望
那位 同學
加油
滿分
微積分 大會考
笑話
直播
留言
求學
聽說
學
直播
憂鬱症
考 滿分
正義
母湯
歐
底下
期末考
哥哥
好帥
中
粉
影片
多人
提到
試試看
版本
分
樣子
看個
哥哥
嘻嘻
娛樂
竟然
大會考
七舍
光頭
那位
大喊
唱
哥哥 母湯
半夜
考
答對
TAG
明明出門
看著
底下 留言
微積分


Jan 2017
Nov 2016
Sep 2016
Jul 2016
May 2016
Mar 2016
Jan 2016
0
5
10
15
20

Issues Faced

- Data Collection
  - Facebook 對資料蒐集的限制
- 自然語言處理
  - 文本分群演算法的選擇
  - 最佳化各方法的參數
- 成果呈現
  - 如何視覺化

Other Efforts

- Agglomerative Clustering
  - 效果不佳
  - 與文獻結論一致

Conclusions

- 自動化找出貼文的主題是可行的
- 分群效果：NMF > k-means > LDA
- 分群時間：LDA (10.1 s) > k-means (8.08 s) > NMF (997 ms)

Future Works

- 結合貼文讚數、留言
  - 議題熱門程度分析
- 結合天氣資料
  - 下雨、寒冷是否會讓人更容易 7pupu
- 將本專案整理成開源套件
  - 目前處理繁體中文社群網站文本的套件仍十分稀少

Job Description

- 蒲郁文(34%) - data preprocessing; clustering
- 蔡孟軒(33%) - clustering; visualization
- 何鳳雯(33%) - data collection; clustering

References

1. Xu, W., Liu, X., & Gong, Y. (2003, July). Document clustering based on non-negative matrix factorization. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 267-273). ACM.
2. Steinbach, M., Karypis, G., & Kumar, V. (2000, August). A comparison of document clustering techniques. In KDD workshop on text mining (Vol. 400, No. 1, pp. 525-526).
3. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3(Jan), 993-1022.

Q & A

Thank You!