高維度資料分析 07小組期末報告

探討影響網路新聞分享量之因素

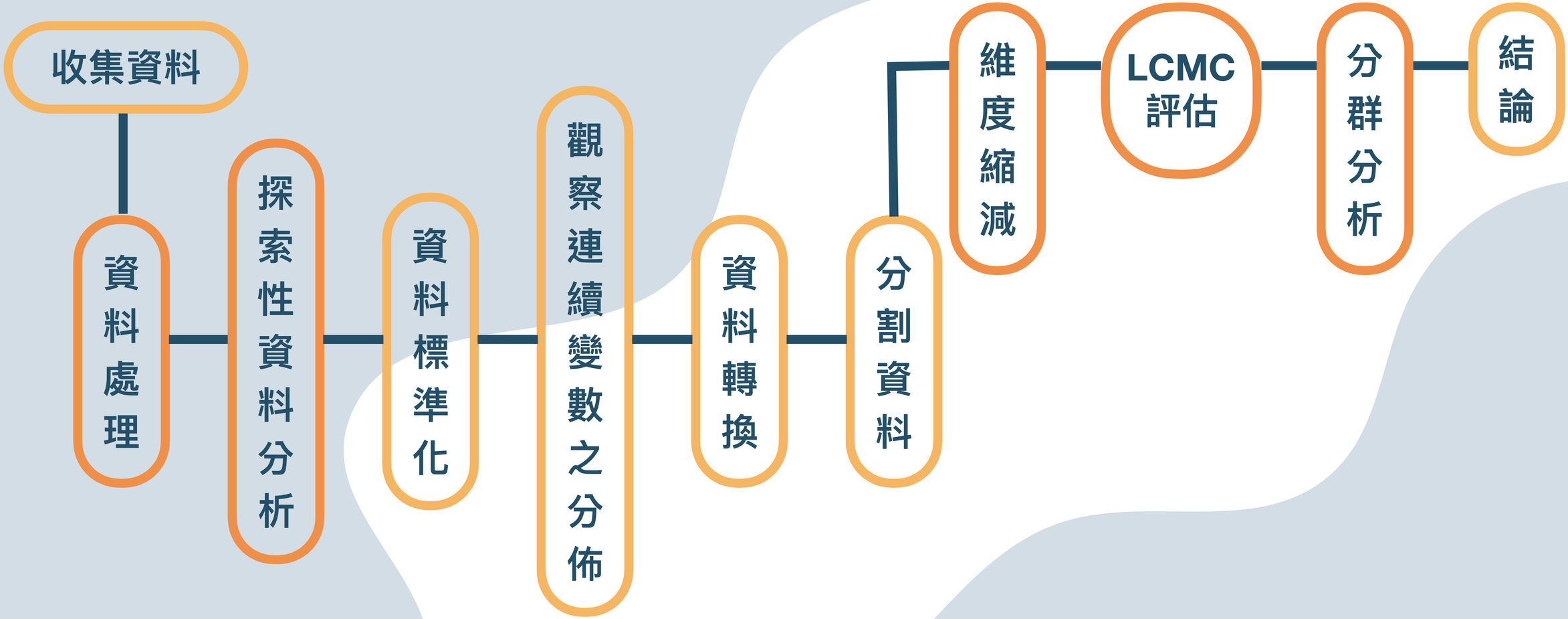
統碩一 710933117 陳逸瑄
統碩一 710933109 林政寬
統碩一 710933120 簡亦萱
單位：國立臺北大學統計學系
報告日期：2020.12.23

目錄

1 導論 INTRODUCTION

- 分析流程
- 研究背景
- 研究動機與目的

導論 分析流程


收集資料
資料處理
探索性資料分析
資料標準化
觀察連續變數之分佈
資料轉換
分割資料
維度縮減
LCMC評估
分群分析
結論

導論 研究背景

2018 年台灣九合一大選期間，我們或多或少曾在社群媒體上看見朋友對候選人的正反意見。當你留言或分享一則貼文，甚至只是按讚，這則訊息便會透過你的帳號傳遞到你的朋友群中。這可以廣泛應用在商業上，像是近年來新興職業YOUTUBER、KOL，都是透過網路聲量來達到產品宣傳的實質效益。

2016 年的美國總統大選期間，川普的競選團隊更是透過數據分析公司「劍橋分析」（Cambridge Analytica）設計的演算法來分析社群媒體用戶，進而投放具有特定立場的內容，以此左右選民投票傾向，藉此讓有利於自己的資訊廣為流傳。

導論 研究動機與目的

希望藉由這次的分析，去了解貼文長度、內文正面詞比例、貼文所屬之類別等，是否會影響目標客群對網路貼文的分享量。因此我們使用資料視覺化與維度縮減，去探討什麼因素會提高民眾分享貼文的意願。

2

資料描述
DESCRIPTION

- 資料來源
- 變數描述
- 五數綜合
- 資料預處理
- 探索性資料分析

資料描述 資料來源

資料來源：UCI, Online News Popularity Data Set

http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity

資料中的文章來自馬沙布爾公司（Mashable），其為網路新聞部落格，取得日期為 2015 年 1 月 8 日，其中有些變數為作者利用隨機森林等方法估計而得。

觀察值個數：39797

類別變數個數：15

連續變數個數：46

其中變數「url」與變數「timedelta」分別為該文章之網址與該文章發佈到收錄此資料集之時間差，本研究不考慮這兩個變數。

資料描述 變數描述

Feature	Description	type
n_tokens_title	Number of words in the title	num, > 0
n_tokens_content	Number of words in the content	num, > 0
n_unique_tokens	Rate of unique words in the content	num, in [0, 1]
n_non_stop_words	Rate of non-stop words in the content	num, in [0, 1]
n_non_stop_unique_tokens	Rate of unique non-stop words in the content	num, in [0, 1]
average_tokens_length	Average length of the words in the content	num, > 0

Words

資料描述 變數描述

Feature	Description	type
num_hrefs	Number of links	num, > 0
num_self_hrefs	Number of links to other articles published by Mashable	num, > 0
self_reference_min_shares	Min. shares of referenced articles in Mashable	num, > 0
self_reference_max_shares	Max. shares of referenced articles in Mashable	num, > 0
self_reference_avg_sharess	Avg. shares of referenced articles in Mashable	num, > 0

Links

資料描述 變數描述

Feature	Description	type
num_imgs	Number of images	num, > 0
num_videos	Number of videos	num, > 0

Media

資料描述 變數描述

Feature	Description	type
num_keywords	Number of keywords in the metadata	num, > 0
kw_min_min	Worst keyword (min. shares)	num, > 0
kw_max_min	Worst keyword (max. shares)	num, > 0
kw_avg_min	Worst keyword (avg. shares)	num, > 0
kw_min_max	Best keyword (min. shares)	num, > 0

Keywords

資料描述 變數描述

Feature	Description	type
kw_max_max	Best keyword (max. shares)	num, > 0
kw_avg_max	Best keyword (avg. shares)	num, > 0
kw_min_avg	Avg. keyword (min. shares)	num, > 0
kw_max_avg	Avg. keyword (max. shares)	num, > 0
kw_avg_avg	Avg. keyword (avg. shares)	num, > 0

Keywords

資料描述 變數描述

Feature	Description	type
data_channel_is_lifestyle	Is data channel 'Lifestyle'?	factor, 0 or 1
data_channel_is_entertainment	Is data channel 'Entertainment'?	factor, 0 or 1
data_channel_is_bus	Is data channel 'Business'?	factor, 0 or 1
data_channel_is_socmed	Is data channel 'Social Media'?	factor, 0 or 1
data_channel_is_tech	Is data channel 'Tech'?	factor, 0 or 1
data_channel_is_world	Is data channel 'World'?	factor, 0 or 1

Category

資料描述 變數描述

Feature	Description	type
weekday_is_monday	Was the article published on a Monday?	factor, 0 or 1
weekday_is_tuesday	Was the article published on a Tuesday?	factor, 0 or 1
weekday_is_wednesday	Was the article published on a Wednesday?	factor, 0 or 1
weekday_is_thursday	Was the article published on a Thursday?	factor, 0 or 1

Category

資料描述 變數描述

Feature	Description	type
weekday_is_friday	Was the article published on a Friday?	factor, 0 or 1
weekday_is_saturday	Was the article published on a Saturday?	factor, 0 or 1
weekday_is_sunday	Was the article published on a Sunday?	factor, 0 or 1
is_weekend	Was the article published on the weekend?	factor, 0 or 1

Category

資料描述 變數描述

Feature	Description	type
LDA_00	Closeness to LDA topic 0	num, in [0, 1]
LDA_01	Closeness to LDA topic 1	num, in [0, 1]
LDA_02	Closeness to LDA topic 2	num, in [0, 1]
LDA_03	Closeness to LDA topic 3	num, in [0, 1]
LDA_04	Closeness to LDA topic 4	num, in [0, 1]

Natural Language Processing

資料描述 變數描述

Feature	Description	type
global_subjectivity	Text subjectivity	num, in [0, 1]
global_sentiment_polarity	Text sentiment polarity	num, in [-1, 1]
global_rate_positive_words	Rate of positive words in the content	num, in [0, 1]
global_rate_negative_words	Rate of negative words in the content	num, in [0, 1]
rate_positive_words	Rate of positive words among non-neutral tokens	num, in [0, 1]
rate_negative_words	Rate of negative words among non-neutral tokens	num, in [0, 1]

Natural Language Processing

資料描述 變數描述

Feature	Description	type
avg_positive_polarity	Avg. polarity of positive words	num, in [0, 1]
min_positive_polarity	Min. polarity of positive words	num, in [0, 1]
max_positive_polarity	Max. polarity of positive words	num, in [0, 1]
avg_negative_polarity	Avg. polarity of negative words	num, in [0, 1]
min_negative_polarity	Min. polarity of negative words	num, in [0, 1]
max_negative_polarity	Max. polarity of negative words	num, in [0, 1]

Natural Language Processing

資料描述 變數描述

Feature	Description	type
title_subjectivity	Title subjectivity	num, in [0, 1]
title_sentiment_polarity	Title polarity	num, in [-1, 1]
abs_title_subjectivity	Absolute subjectivity level	num, in [0, 1]
abs_title_sentiment_polarity	Absolute polarity level	num, in [0, 1]

Natural Language Processing

資料描述 變數描述

Target

Feature	Description	type
shares	Number of shares	num, > 0

資料描述 五數綜合

```
n_tokens_title
 Min.    : 2.0
 1st Qu.: 9.0
 Median :10.0
 Mean    :10.4
 3rd Qu.:12.0
 Max.    :23.0
```

```
n_tokens_content
 Min.    :   0.0
 1st Qu.: 246.0
 Median : 409.0
 Mean    : 546.5
 3rd Qu.: 716.0
 Max.    :8474.0
```

```
n_unique_tokens
Min.    :  0.0000
1st Qu.:  0.4709
Median :  0.5392
Mean    :  0.5482
3rd Qu.:  0.6087
Max.    :701.0000
```

```
n_non_stop_words
Min.    :   0.0000
1st Qu.:   1.0000
Median :   1.0000
Mean    :   0.9965
3rd Qu.:   1.0000
Max.    :1042.0000
```

```
n_non_stop_unique
     _tokens
 Min.    :  0.0000
 1st Qu.:  0.6257
 Median :  0.6905
 Mean    :  0.6892
 3rd Qu.:  0.7546
 Max.    :650.0000
```

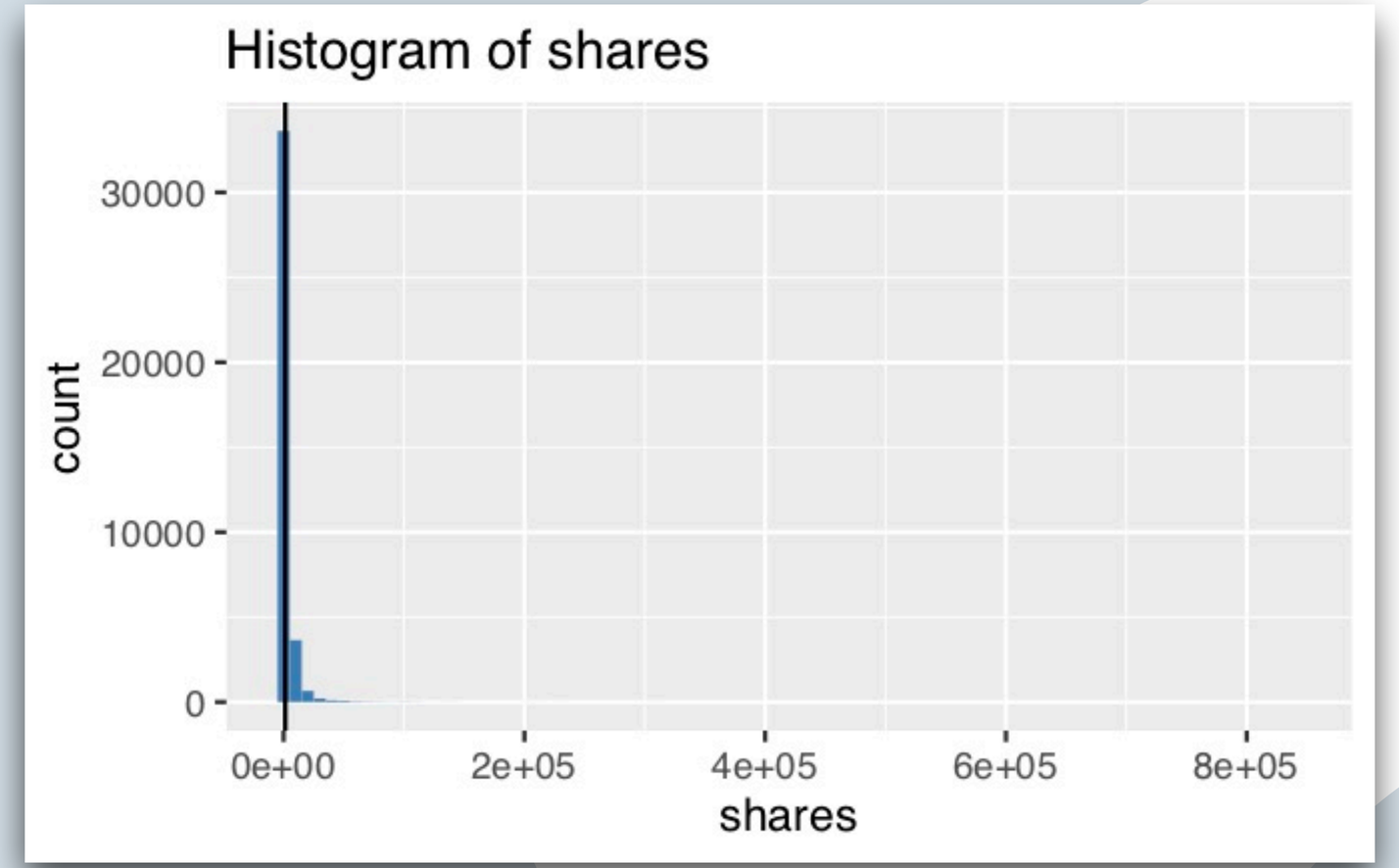
```
    num_hrefs
Min.    :  0.00
1st Qu.:  4.00
Median :  8.00
Mean    : 10.88
3rd Qu.: 14.00
Max.    :304.00
```

```
 num_self_hrefs
Min.    :  0.000
1st Qu.:  1.000
Median :  3.000
Mean    :  3.294
3rd Qu.:  4.000
Max.    :116.000
```

資料描述 資料預處理

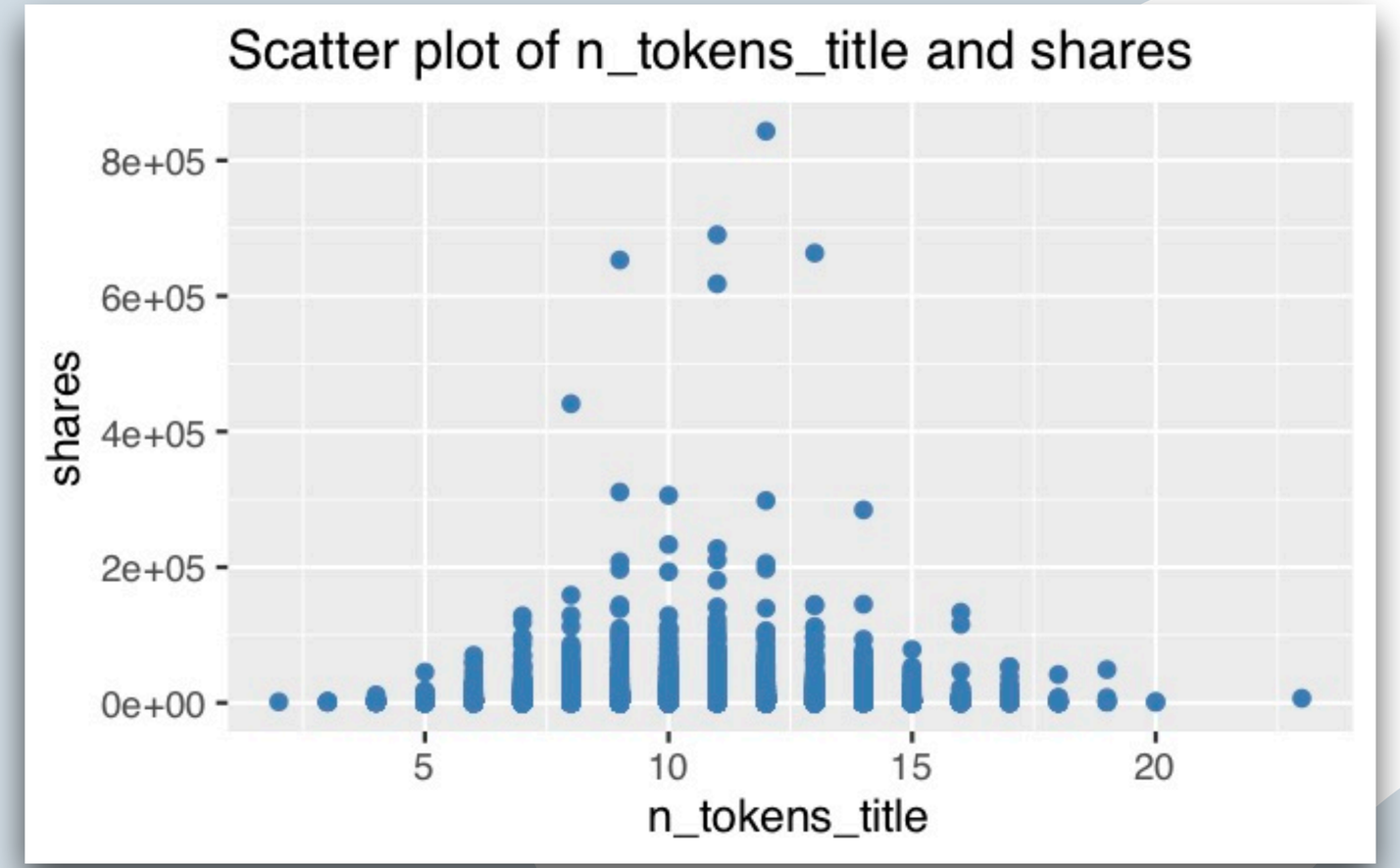
- 刪除異常值
- 將 channel 變數都為 0 的資料設定成其他類別（other）
- 將類別變數刪除

資料描述 探索性資料分析



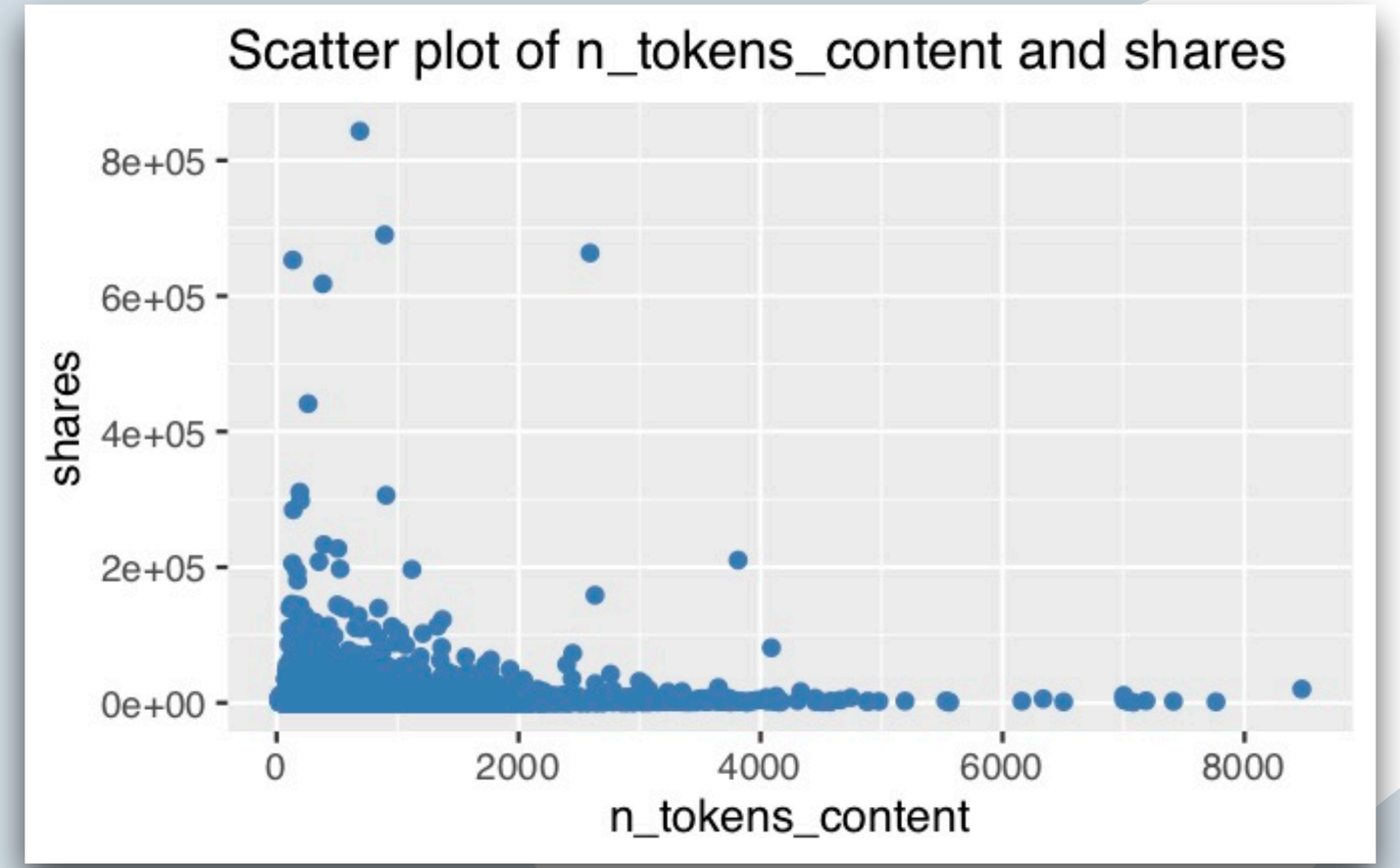
此現為分享量之中位數，文章大多集中在 1400 個分享數，但也有少數文章有相當高的分享量。

資料描述 探索性資料分析



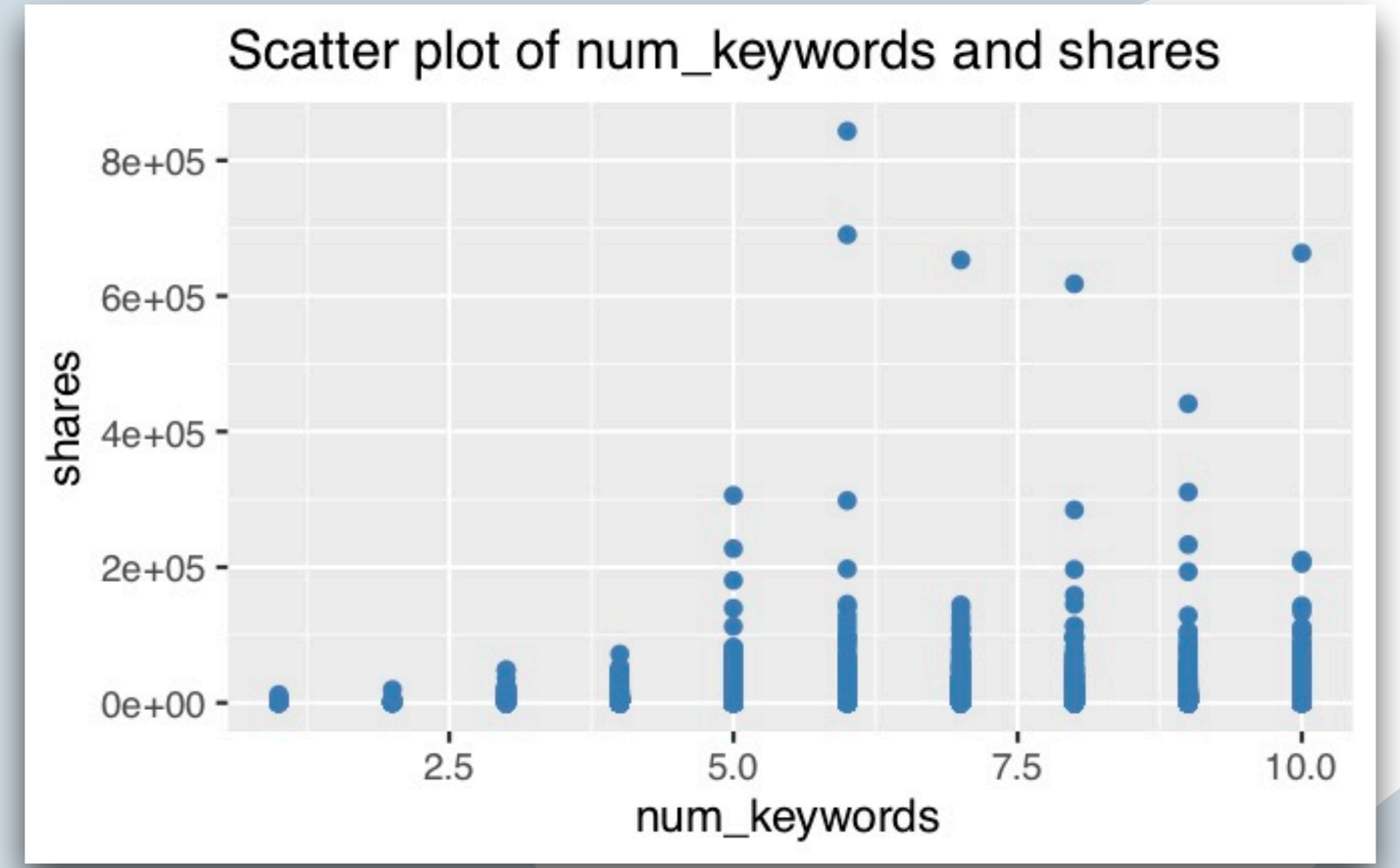
標題字數太多或太少都傾向獲得較少的分享數。

資料描述 探索性資料分析



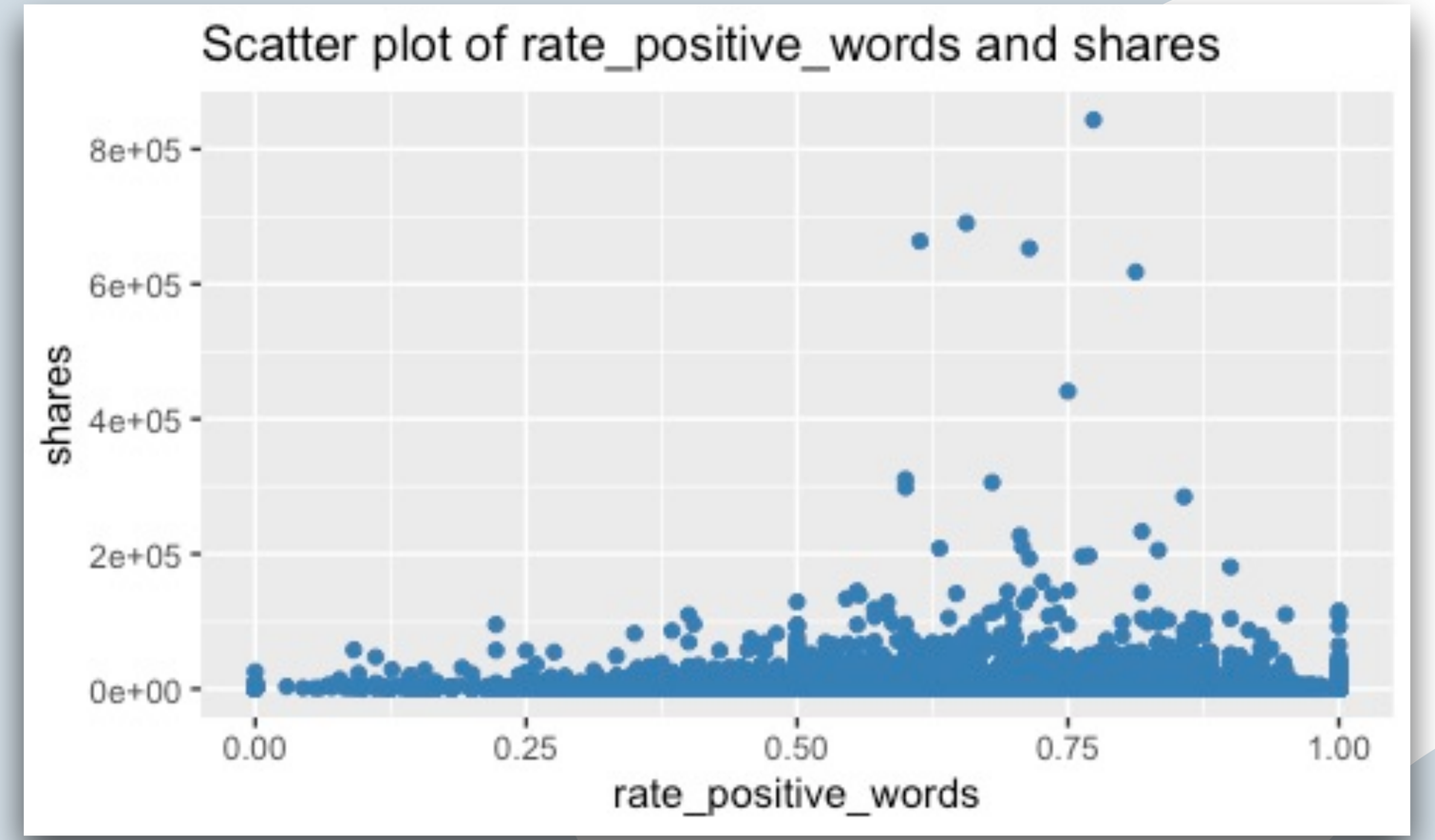
內容字數太多傾向獲得較少的分享數。

資料描述 探索性資料分析



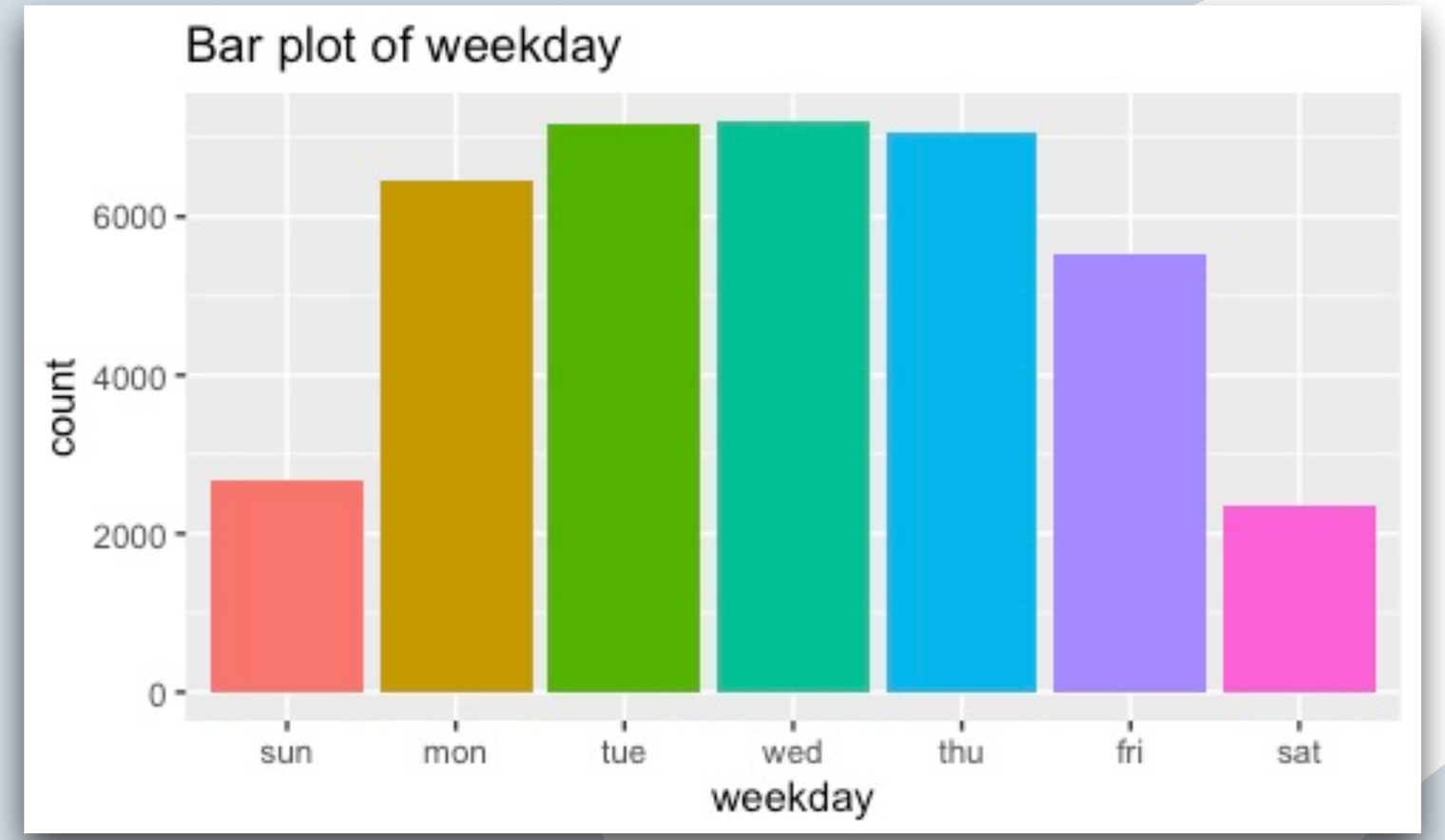
關鍵字數多傾向獲得較多的分享數。

資料描述 探索性資料分析



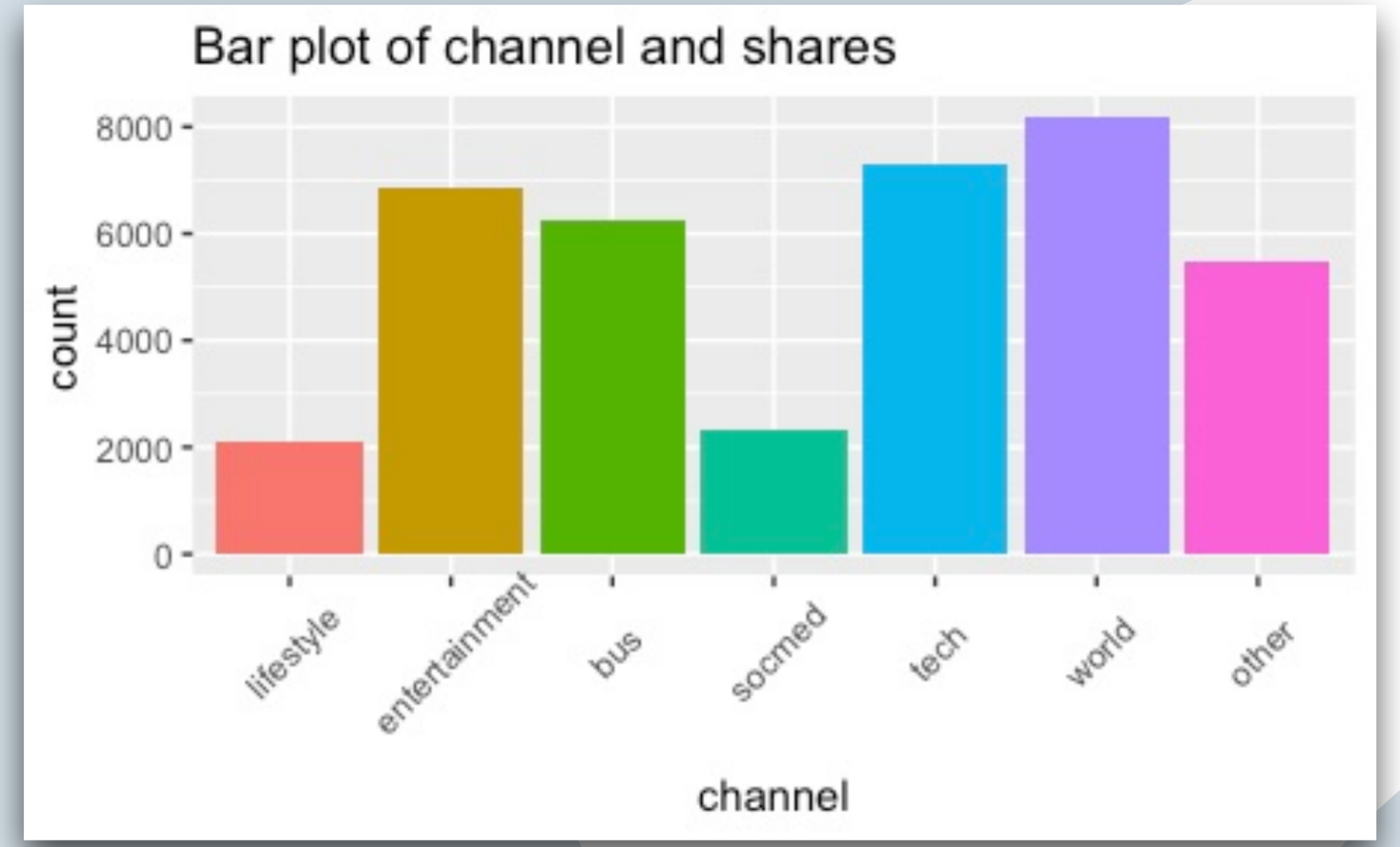
正面詞彙多傾向獲得較多的分享數。

資料描述 探索性資料分析



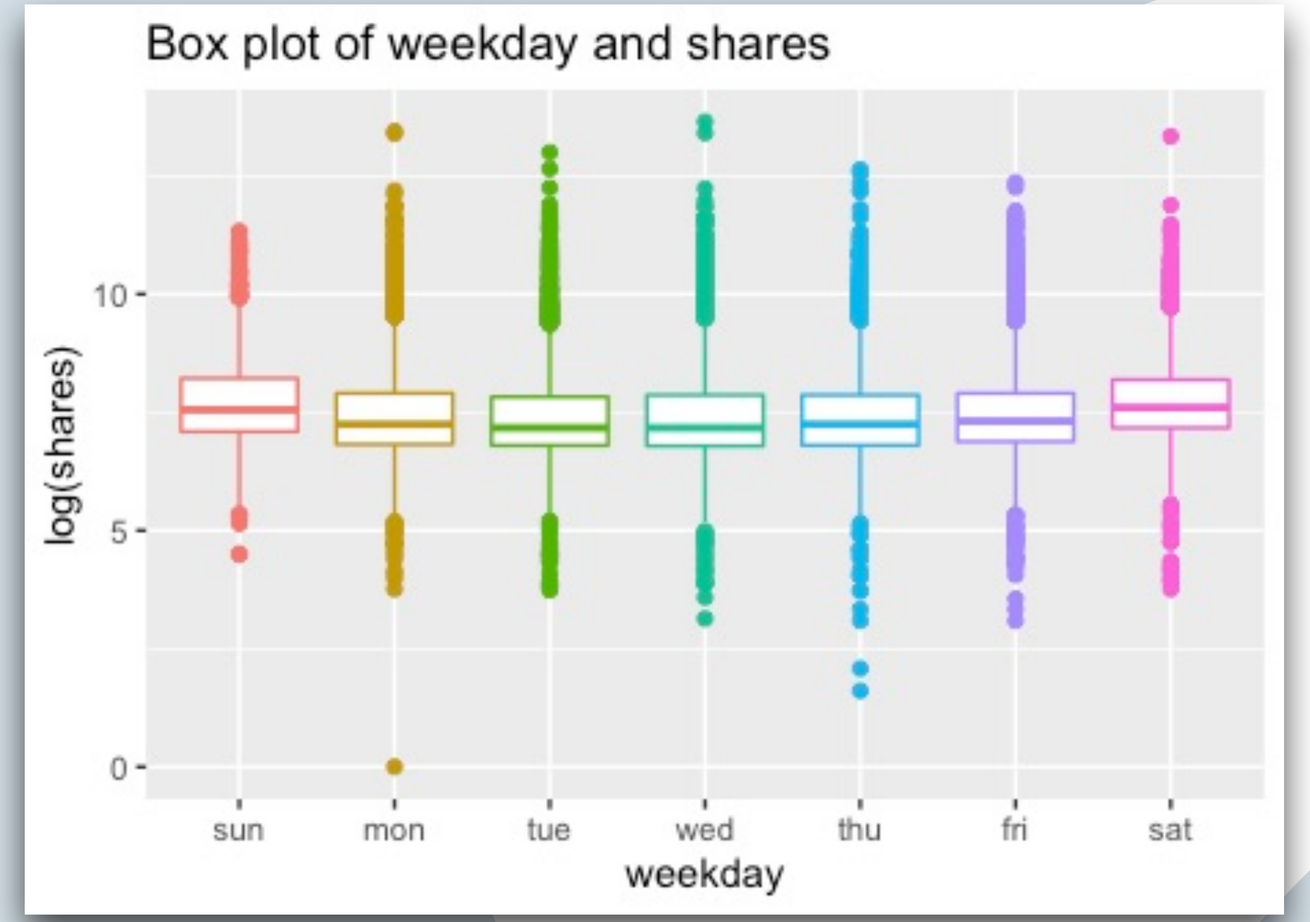
假日的文章數明顯比平日的文章數少，原因可能是假日沒有收集足夠的資料或是此網站假日發佈之文章較少。

資料描述 探索性資料分析



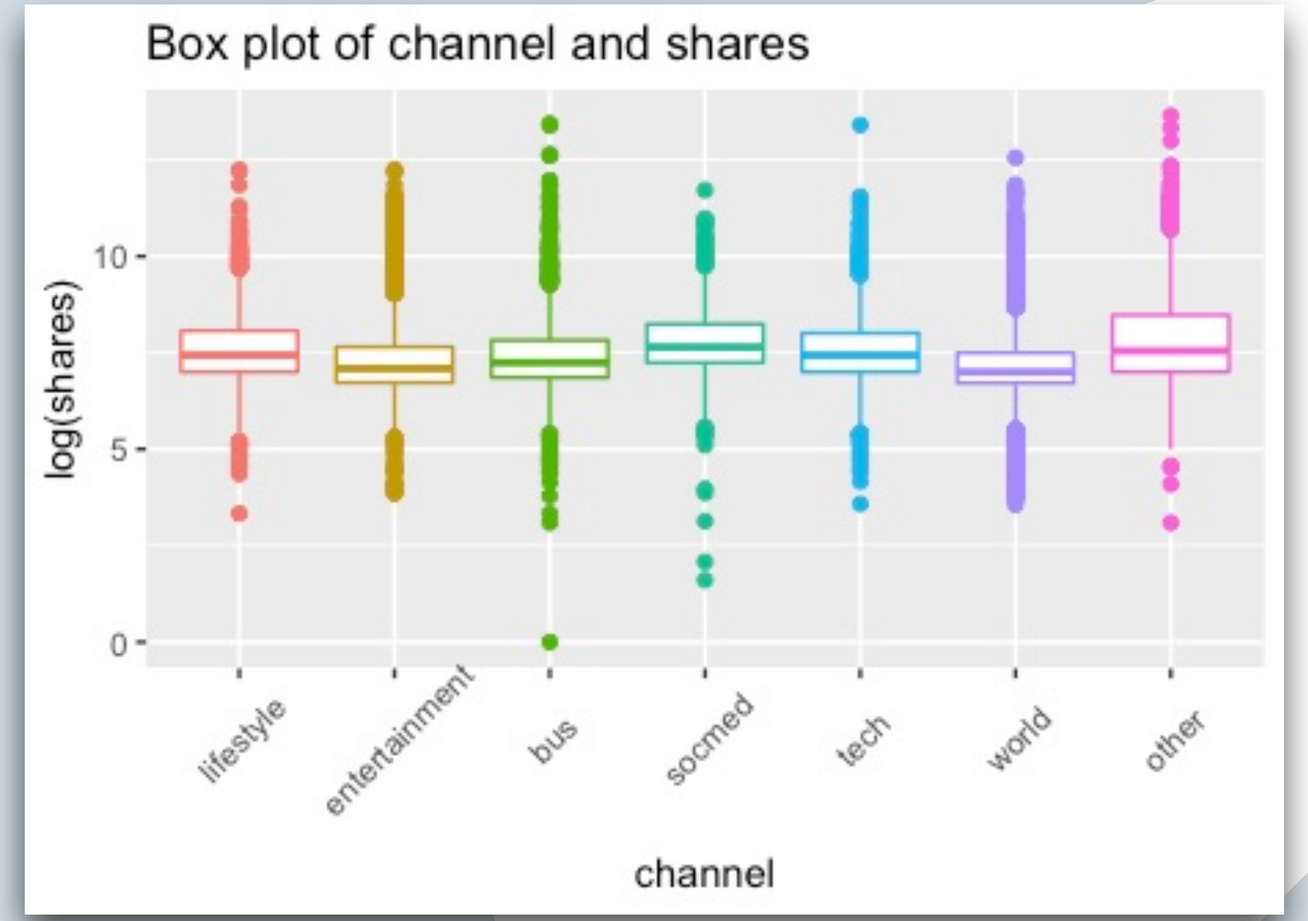
生活風格與社交媒體類別的文章數較其他類別少，最多的則為世界類別。

資料描述 探索性資料分析



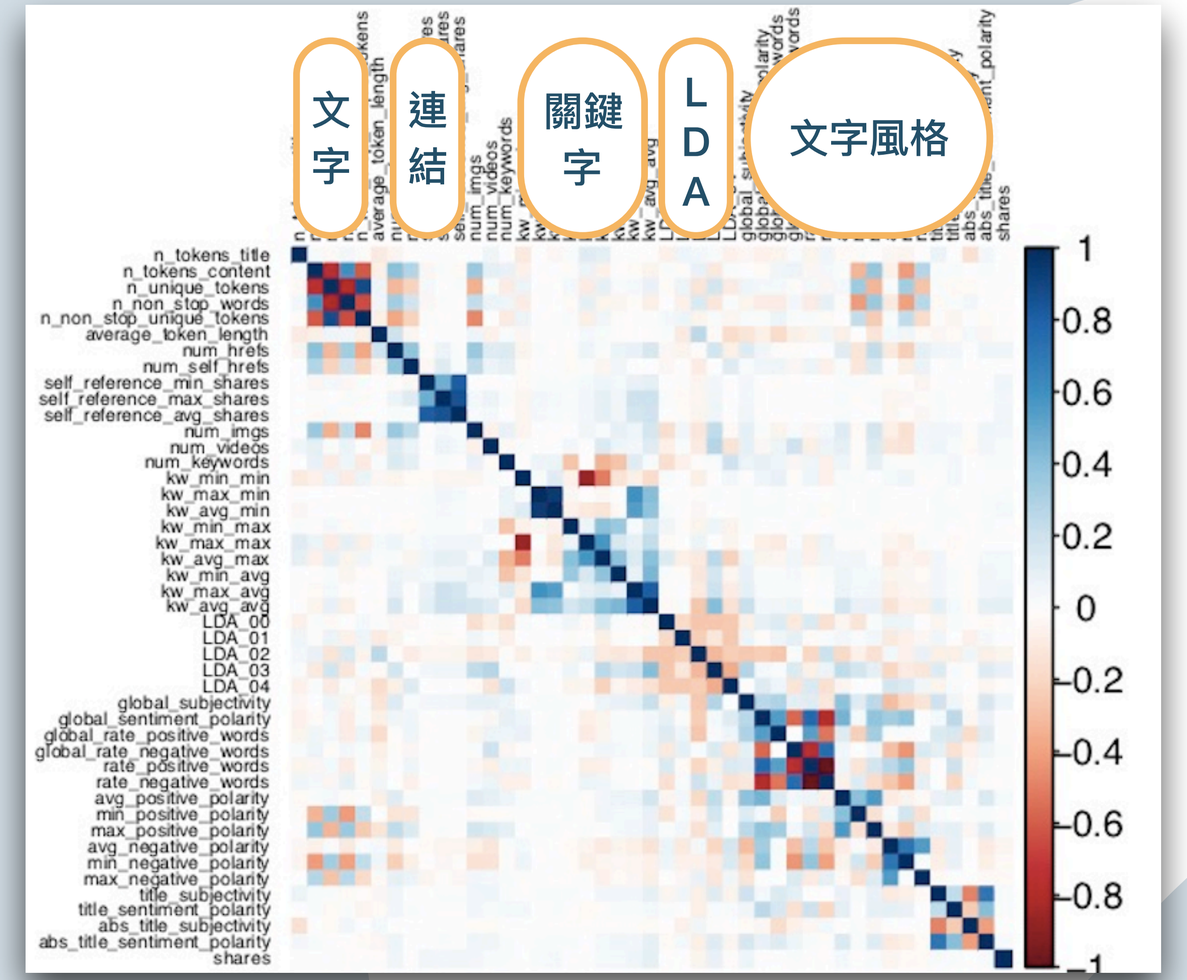
不同的發佈星期對於分享量的分佈相似。

資料描述 探索性資料分析



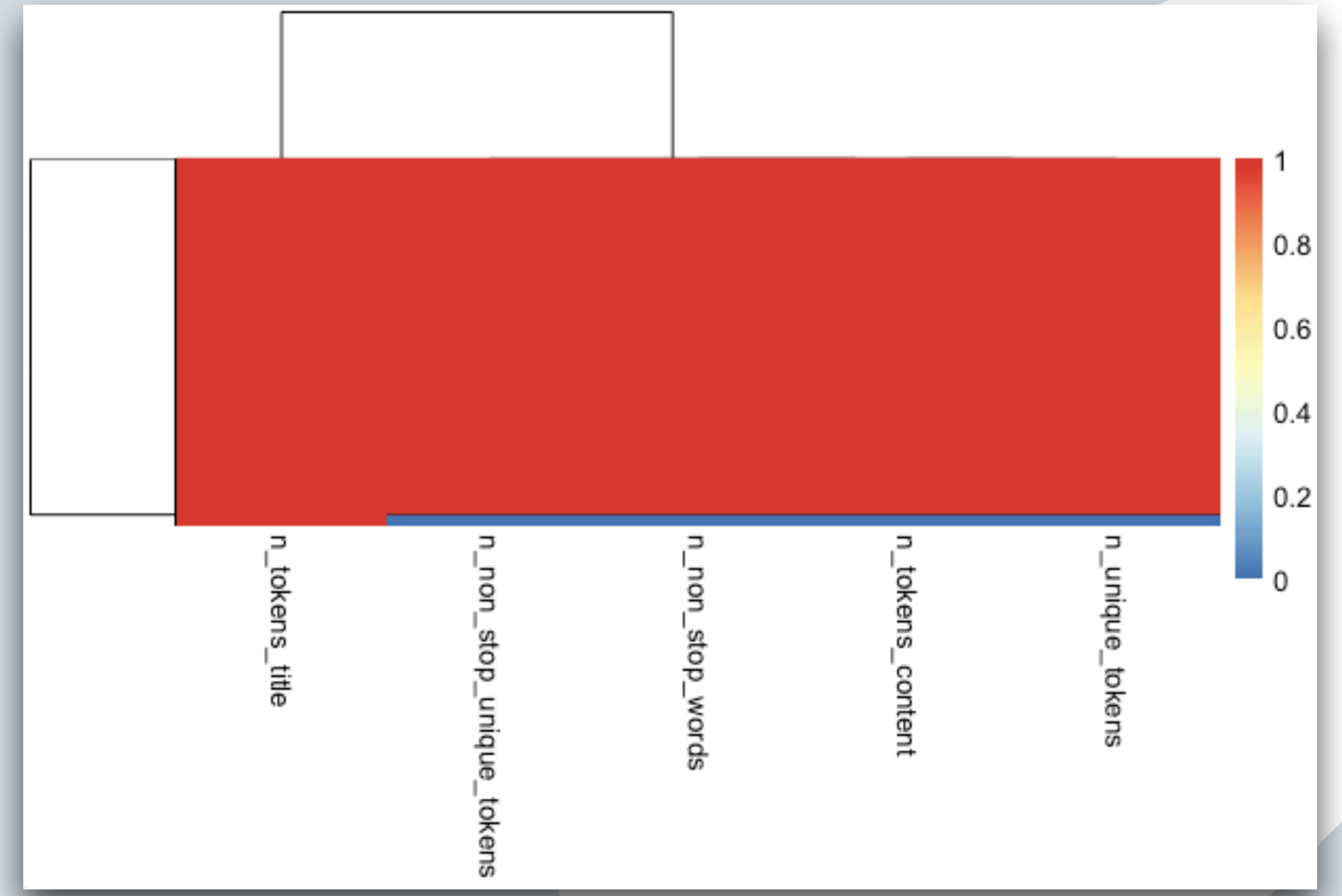
不同的文章類別對於分享量的分佈相似。

資料描述 探索性資料分析



變數之間大致分為五類，類別裡有較高程度的相關性。

資料描述 探索性資料分析



文章內容字數為 0 之資料，其他關於文字個數之變數也為 0。

- 資料標準化
- 變數分佈
- 資料轉換
- 分割資料集

3

資料處理
PROCESSING

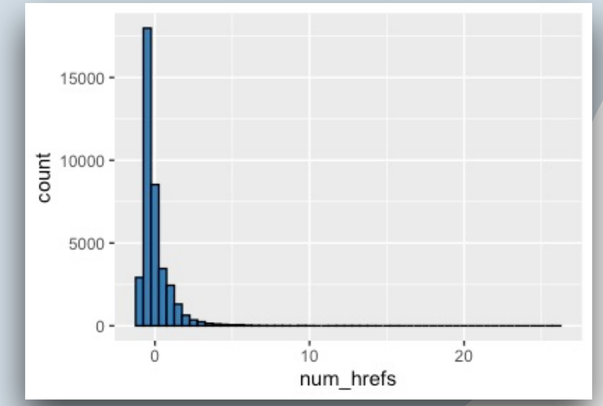
資料處理 資料標準化

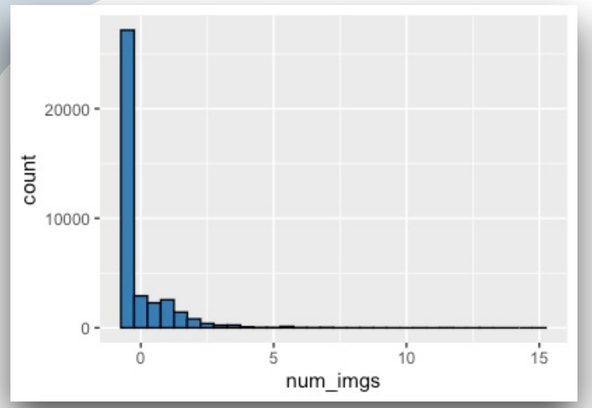
由 summary 可看出變數的範圍差距很大，因此，將連續型變數進行標準化。

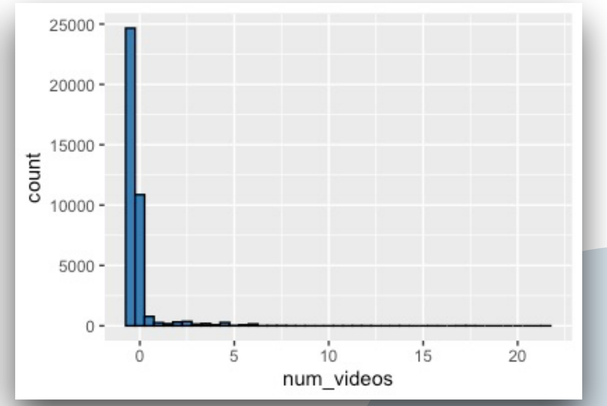
$$\frac{x - \mathbf{mean}(x)}{\mathbf{sd}(x)}$$

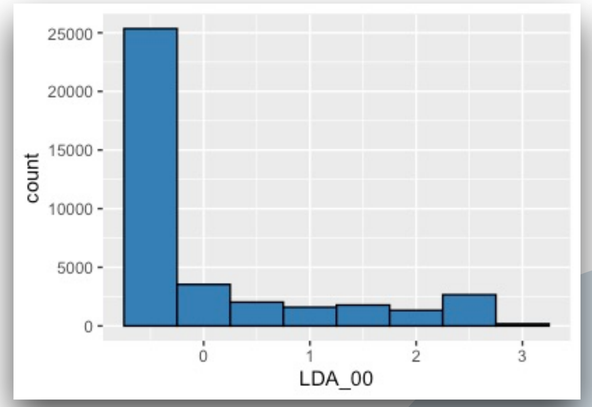
資料處理 變數分佈

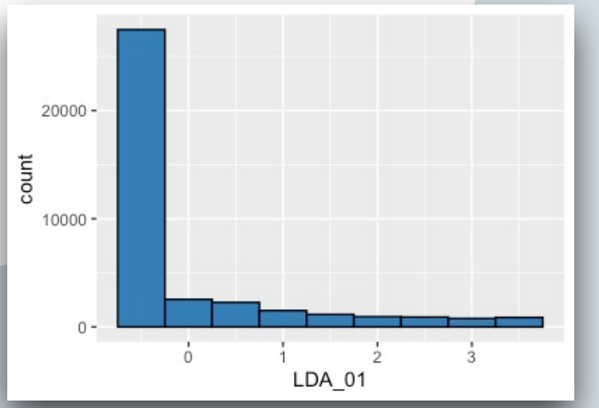
右偏


count
15000
10000
5000
0
0
10
20
num_hrefs


count
20000
10000
0
0
5
10
15
num_imgs

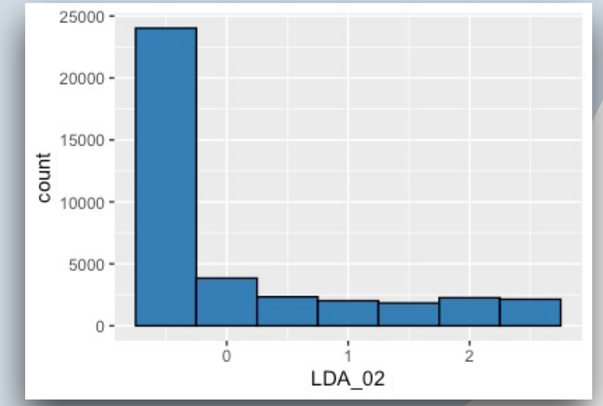

count
25000
20000
15000
10000
5000
0
0
5
10
15
20
num_videos

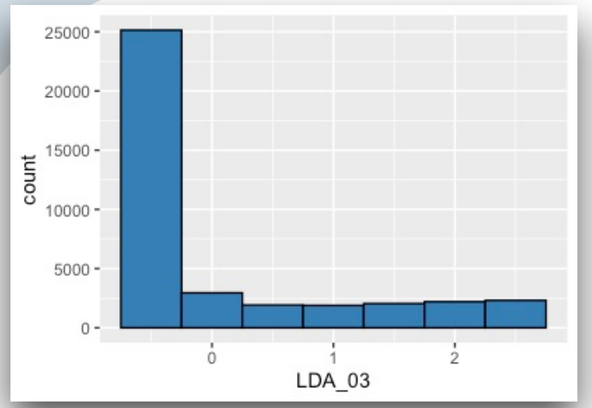

count
25000
20000
15000
10000
5000
0
0
1
2
3
LDA_00

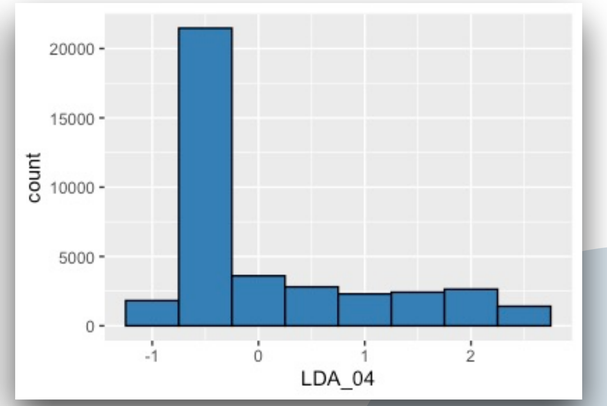

count
20000
10000
0
0
1
2
3
LDA_01

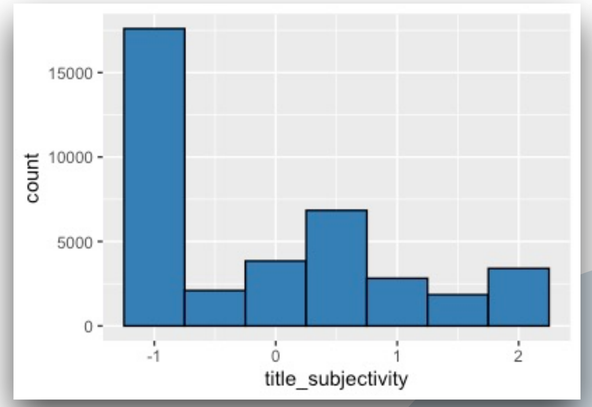
資料處理 變數分佈

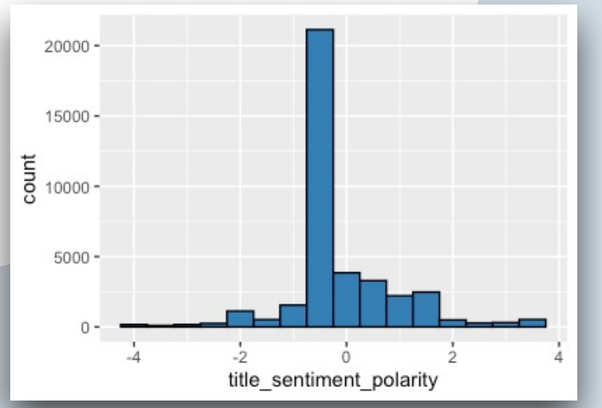
右偏


25000
20000
15000
10000
5000
0
count
0
1
2
LDA_02


25000
20000
15000
10000
5000
0
count
0
1
2
LDA_03


20000
15000
10000
5000
0
count
-1
0
1
2
LDA_04


15000
10000
5000
0
count
-1
0
1
2
title_subjectivity

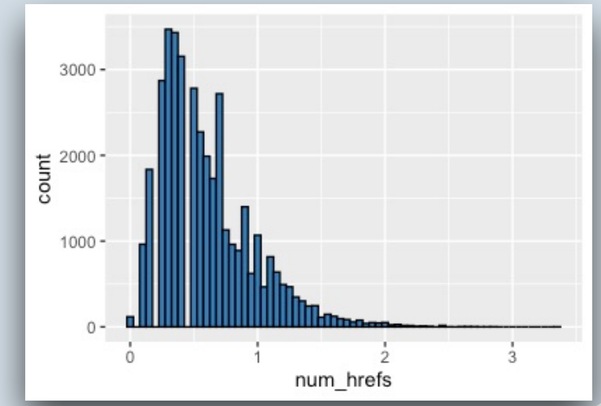

20000
15000
10000
5000
0
count
-4
-2
0
2
4
title_sentiment_polarity

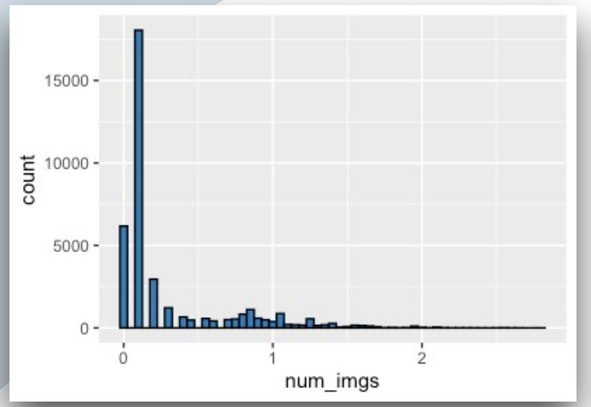
資料處理 資料轉換

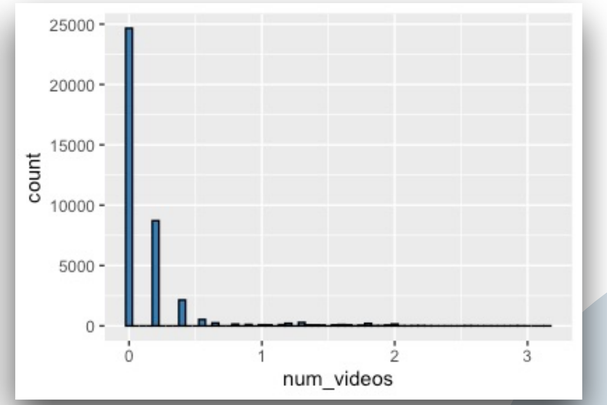
由直方圖，可看出這些變數有明顯右偏之現象，因此，將該變數進行對數轉換，減少極端值對資料的影響。

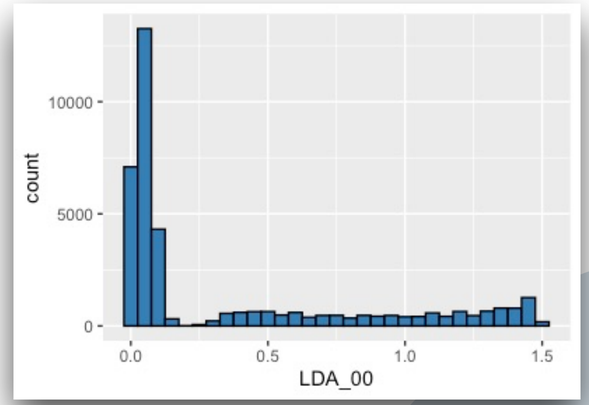
$$\log(x + 1 - \min(x))$$

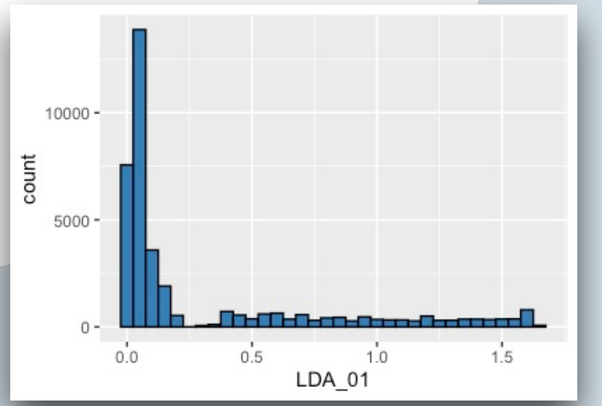
資料處理 變數分佈


count
3000
2000
1000
0
0
1
2
3
num_hrefs

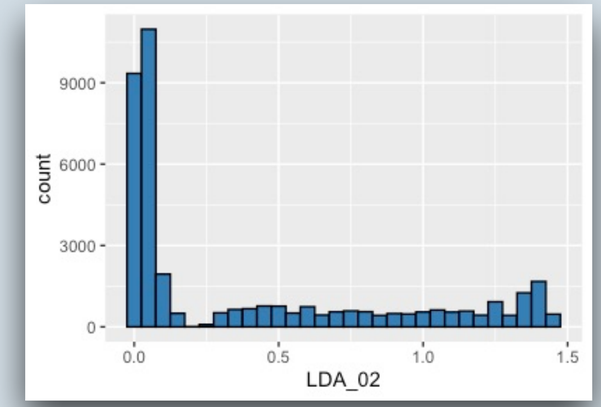

count
15000
10000
5000
0
0
1
2
num_imgs

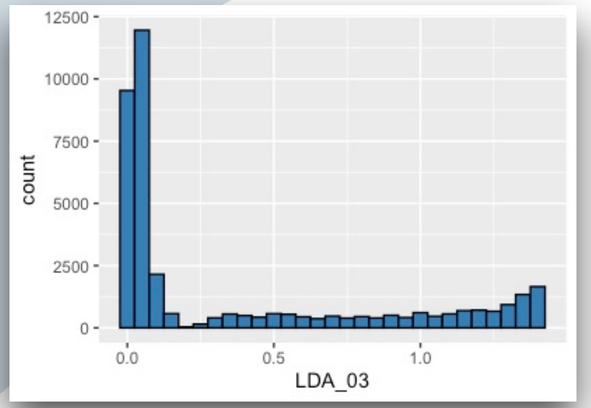

count
25000
20000
15000
10000
5000
0
0
1
2
3
num_videos

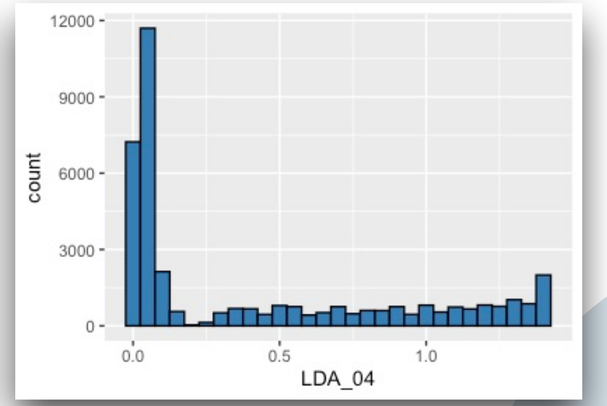

count
10000
5000
0
0.0
0.5
1.0
1.5
LDA_00

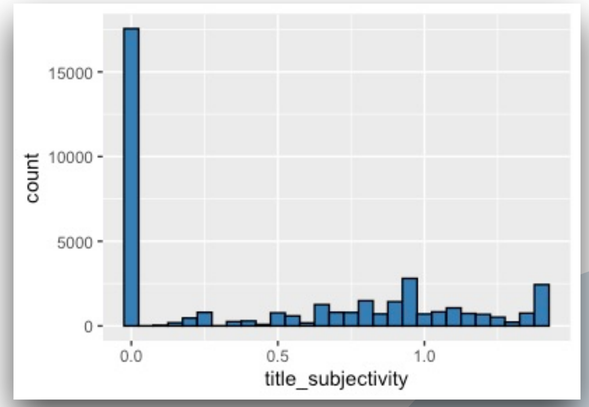

count
10000
5000
0
0.0
0.5
1.0
1.5
LDA_01

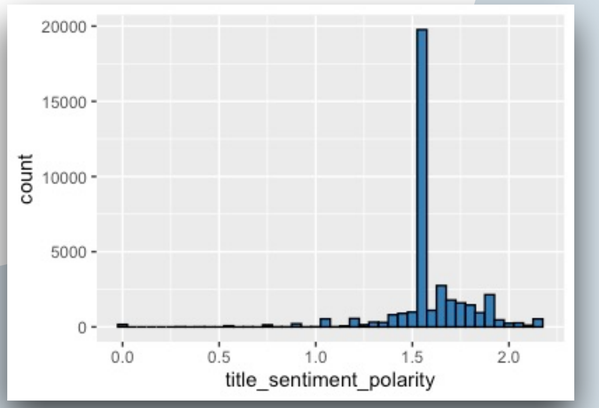
資料處理 變數分佈


count
9000
6000
3000
0
0.0
0.5
1.0
1.5
LDA_02


count
12500
10000
7500
5000
2500
0
0.0
0.5
1.0
LDA_03

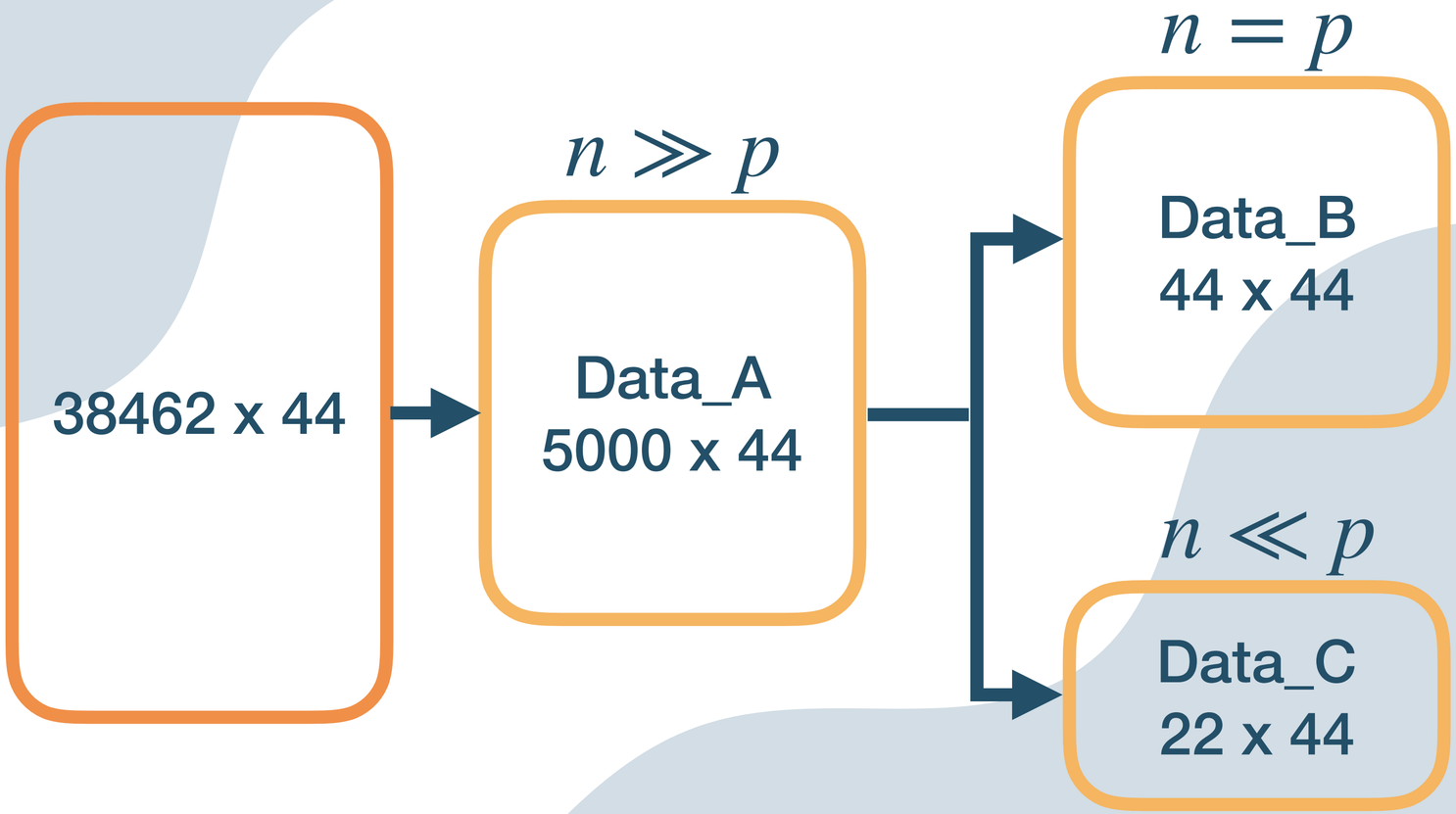

count
12000
9000
6000
3000
0
0.0
0.5
1.0
LDA_04


count
15000
10000
5000
0
0.0
0.5
1.0
title_subjectivity


count
20000
15000
10000
5000
0
0.0
0.5
1.0
1.5
2.0
title_sentiment_polarity

資料處理 分割資料集

由於維度縮減分析與分群分析的計算量較大，因此，在分析之前，先對資料做隨機抽樣，並分成長型、正方形以及寬型之資料。



- 維度縮減
- LCMC評估
- 變數之重要程度
- 分群分析

4 資料分析
ANALYSIS

資料分析 維度縮減

LDA

LDA 在尋找最佳解釋資料的變量線性組合，試圖找出一個線性方程式，將資料分開。

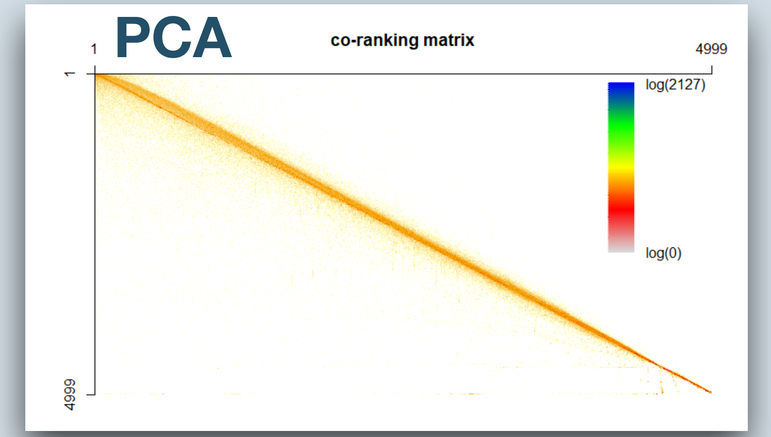
PCA

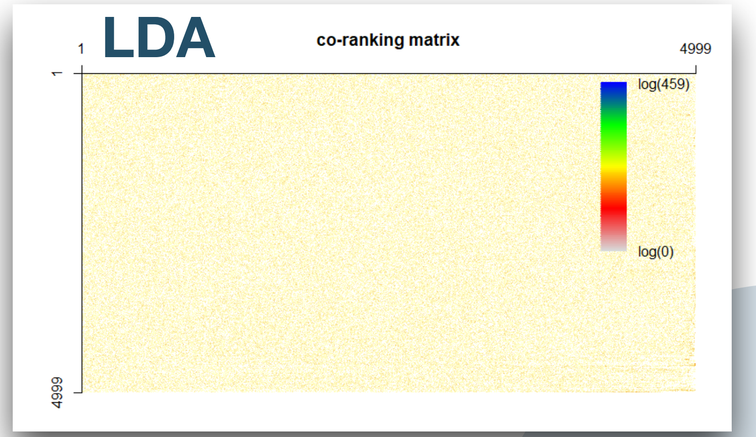
PCA 透過選取較大的特徵值達成維度縮減，並保留資料的最大變異。

SVD

SVD 透過奇異值分解，隱含的縮減維度，變數個數不會減少，但可以減少計算量。

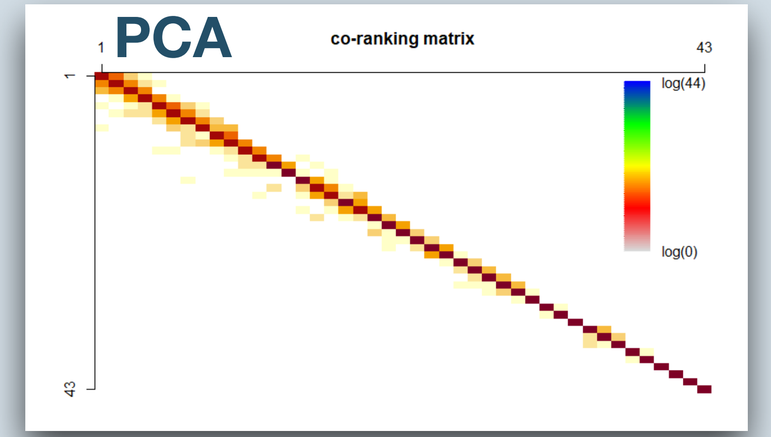
資料分析 LCMC 評估

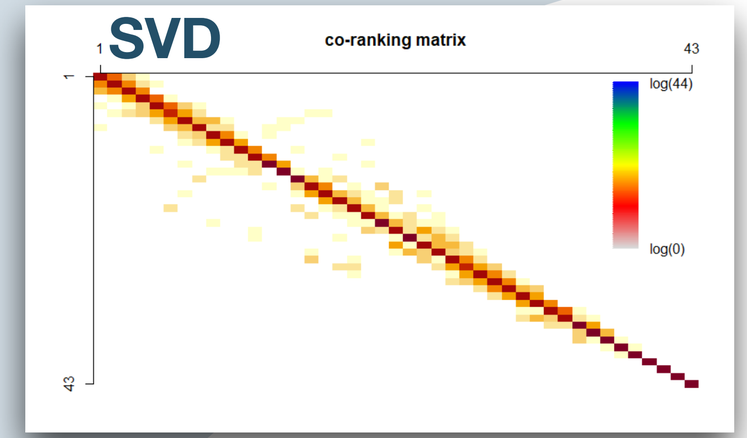


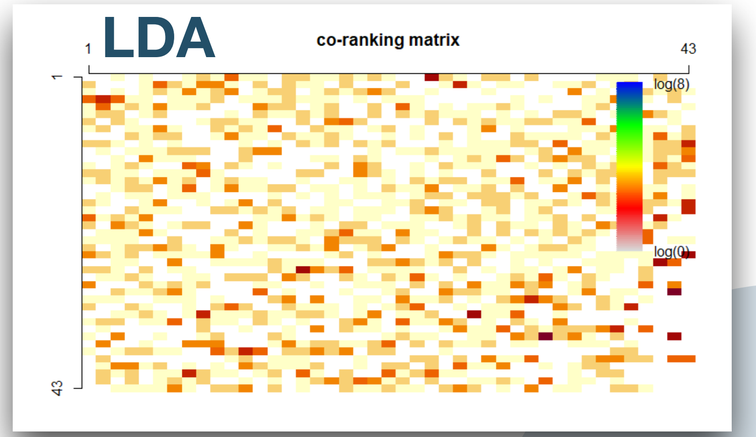


對於觀察值筆數大於變數個數的情況，透過 co-ranking matrix 明顯可看出 PCA 的降維方法使資料的結構改變較少。

資料分析 LCMC 評估

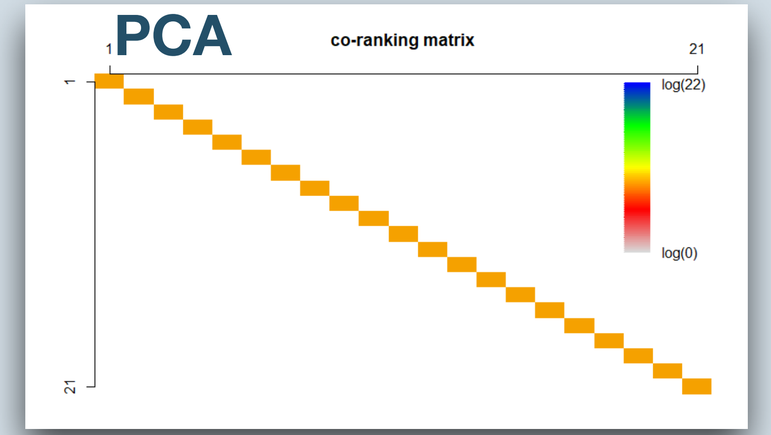


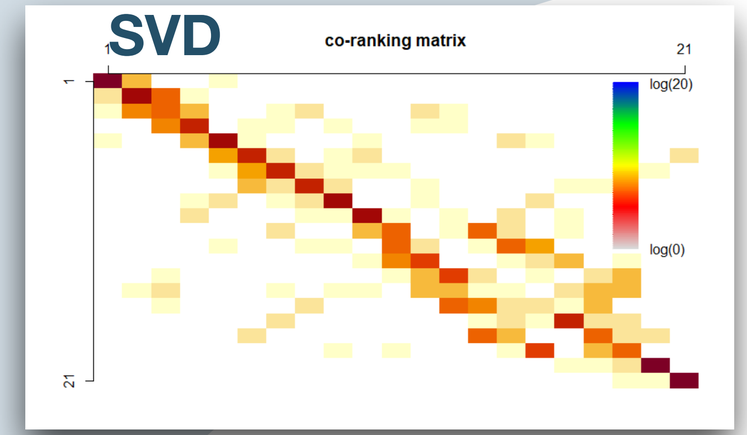


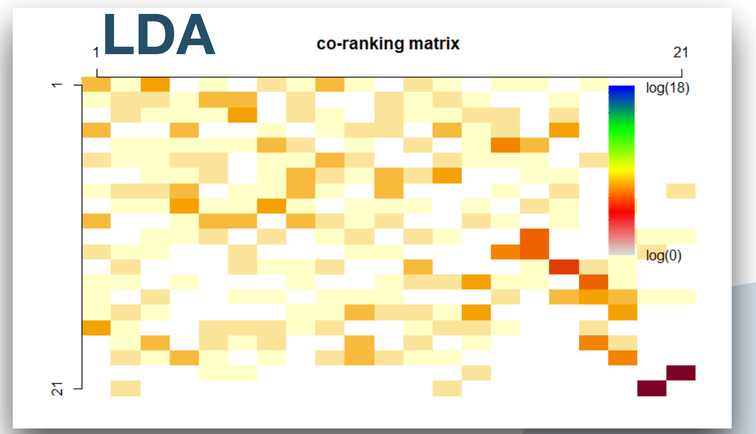


對於觀察值筆數等於變數個數的情況，透過 co-ranking matrix 明顯可看出 PCA 的降維方法使資料的結構改變較少。

資料分析 LCMC 評估



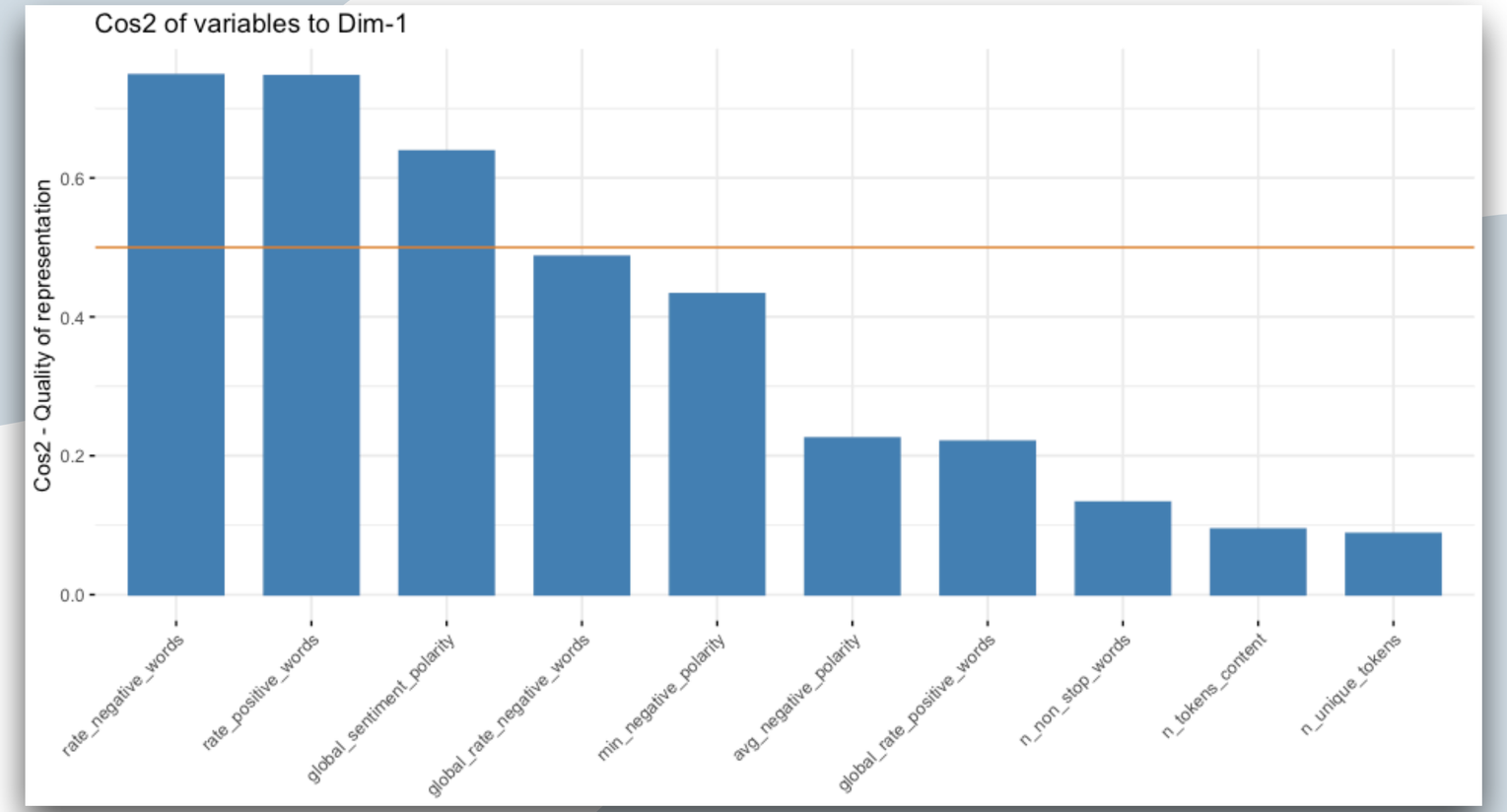




對於觀察值筆數小於變數個數的情況，透過 co-ranking matrix 明顯可看出 PCA 的降維方法使資料的結構改變較少。

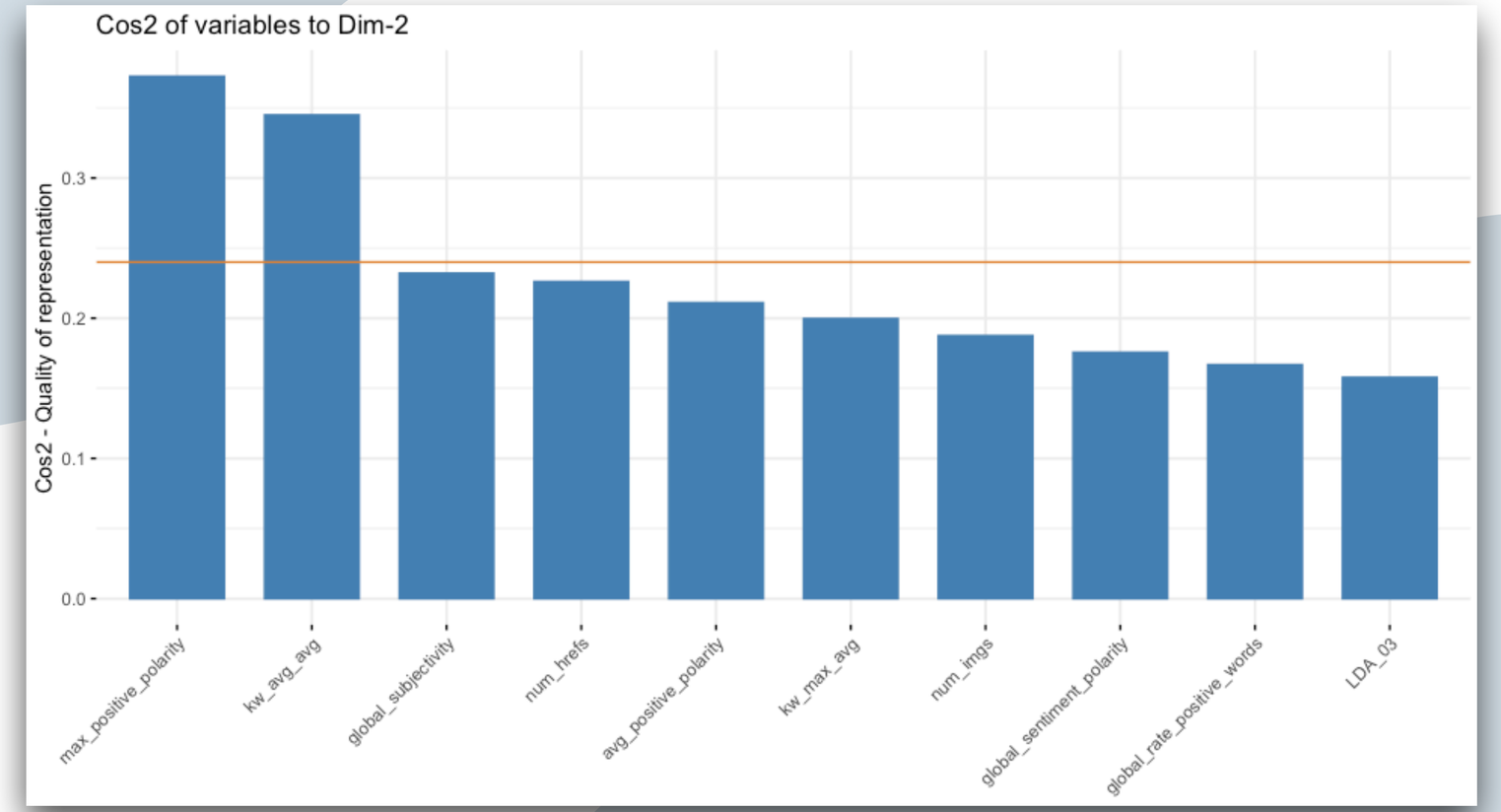
資料分析 變數之重要程度

對第一主成份來說，最重要的幾個變數都為經自然語言處理而統計出的變數。



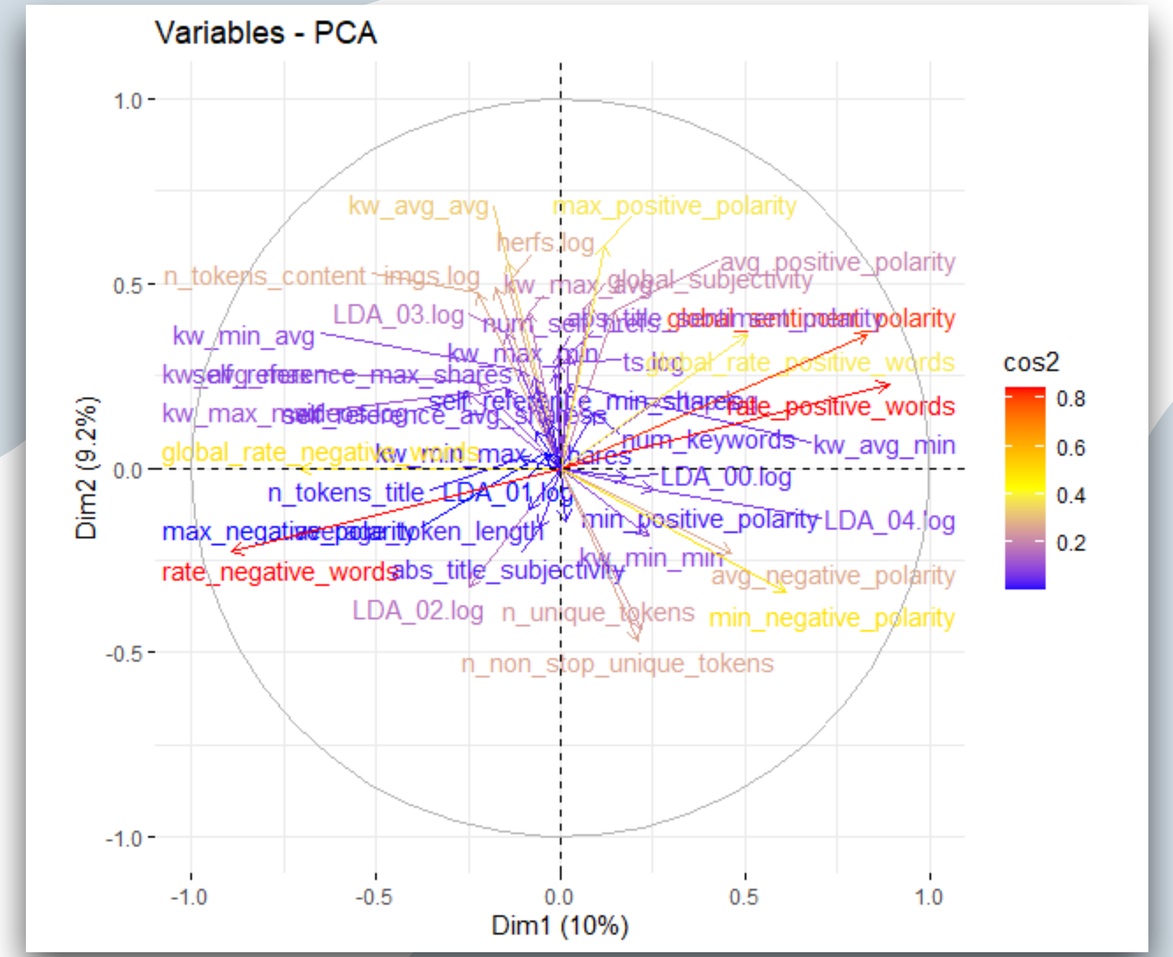
資料分析 變數之重要程度

對第二主成份來說，最重要的變數都為描述關鍵字之變數與經自然語言處理而統計出之變數。

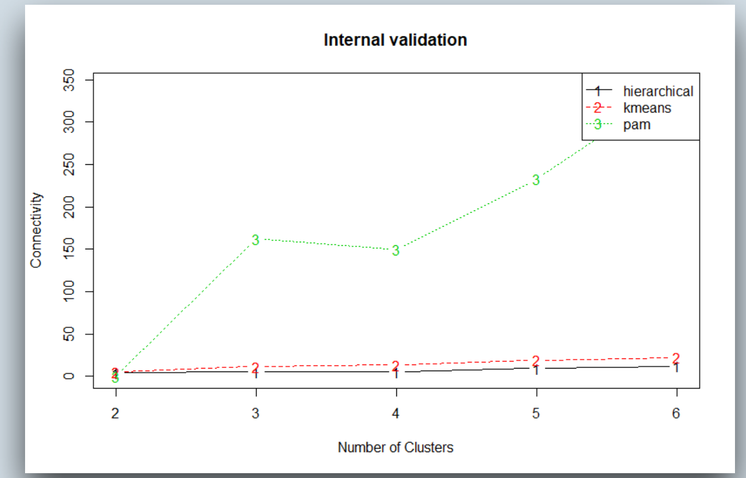


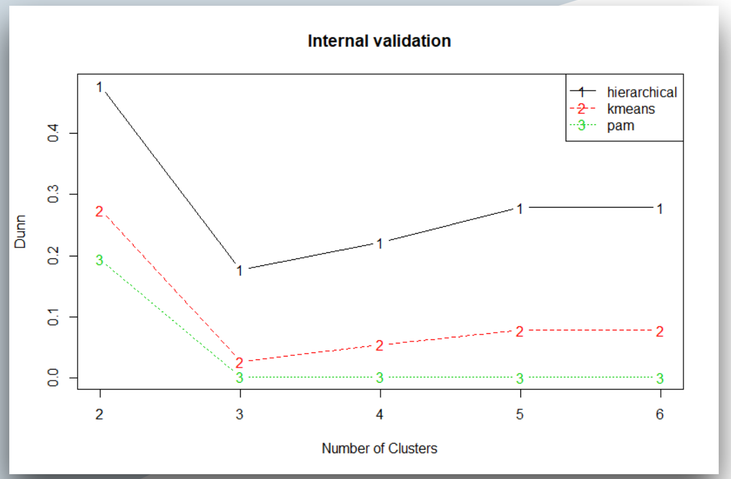
資料分析 變數之重要程度

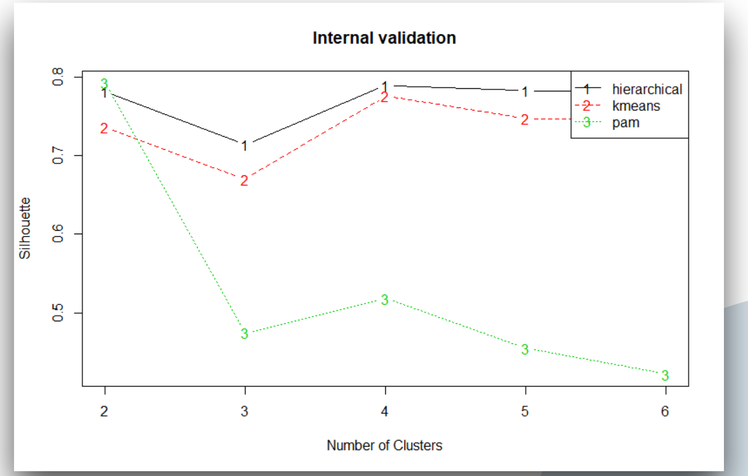
箭頭越長表示貢獻度越大，因此可看出經自然語言處理而統計出之變數貢獻度較大。



資料分析 分群分析







由分群的三個指標可看出利用階層式分群將資料分成兩群會是較好的結果。

- 分析結果
- 延伸討論

5 結論 CONCLUSION

結論 分析結果

- 標題字數太多或太少都傾向獲得較少的分享量
- 關鍵字字數多傾向獲得較高的分享量
- 正面詞彙多傾向獲得較高的分享量
- 經自然語言處理而統計出的變數較能解釋分享量的變異程度，尤其是正面詞彙的比率

結論 延伸討論

新聞是向大眾傳達時事的工具，因此社會中的各個事件都可能影響到一則新聞的熱門程度，未來若加上該篇新聞發佈時的重大時事議題，或許能使分析更加完善。

- 致謝詞
- 組員分工
- 組員設備概要
- 參考資料
- R 套件

A 附錄 APPENDIX

附錄 致謝詞

- 感謝吳漢銘老師這學期的教導
- 感謝各位同學的聆聽
- 感謝組員們的配合

附錄 組員分工

陳逸瑄

探索性資料分析、資料處理、
維度縮減、結論、製作簡報、
製作影片、程式碼彙整
期望分數：90

林政寬

探索性資料分析、資料處理、
維度縮減、分群分析、結論
期望分數：90

簡亦萱

導論、探索性資料分析、資料
處理
期望分數：90

附錄 組員設備概要

陳逸瑄

作業系統：macOS 10.14.6

處理器：1.6 GHz Intel Core i5

記憶體：8 GB 1600 MHz DDR3

林政寬

作業系統：Windows 10

處理器：Intel(R) Core(TM) i5-6198DU CPU

記憶體：20 GB

簡亦萱

作業系統：Windows 10

處理器：Intel(R) Core(TM) i5-8250U CPU

記憶體：8 GB

附錄 參考資料

- https://medium.com/@syedsadiqalinaqvi/predicting-popularity-of-online-news-articles-a-data-scientists-report-fac298466e7
- https://www.kaggle.com/kerneler/starter-uci-online-news-popularity-30f849b5-1
- https://rstudio-pubs-static.s3.amazonaws.com/122671_778c16d46da6489c9f88cd7c12b20ed3.html
- https://www.kaggle.com/thehapyone/exploratory-analysis-for-online-news-popularity

附錄 R 套件

filter {dplyr}, sample_n {dplyr}, select {dplyr}, mutate {dplyr},
ggplot {ggplot2}, geom_historgam {ggplot2}, geom_point {ggplot2},
geom_bar {ggplot2}, geom_boxplot {ggplot2}, geom_vline {ggplot2},
geom_hline {ggplot2}, pheatmap {pheatmap}, corrplot {corrplot},
PCA {FactoMineR}, fviz_cos2 {factoextra},
fviz_pca_var {factoextra},
coranking {coRanking},
imageplot {coRanking}, LCMC {coRanking},
lda {MASS}, clValid {clValid}

THANKS