國立雲林科技大學

資料探勘專案作業三

Python 軟體實作群聚分析

指導教授：許中川　教授

學生：M10921002 宋沂芸

M10921032 林師弘

M10921036 童湘庭

M10921038 張珮柔

摘要

隨著網路的發展，一打開手機、電腦並連上網路，新聞文章隨時皆可被閱覽，而面對大量的新聞內容很難使用人工一一進行新聞的分類，因此，本研究欲探討哪些新聞類別會有群聚的效果，故選用 mini 20 Newsgroups 資料集，使用 K-means、Hierarchical clustering 及 DBSCAN 研究新聞類別的分群， 且以 Purity 進行評估，而研究顯示 K-means 的 Purity 為 0.30005，Hierarchical clustering 的 Purity 為 0.07，DBSCAN 的 Purity 為 0.05，其中 K-means 的群聚效果最佳。

近年來越來越多的賣家透過臉書直播功能來吸引顧客並銷售商品，因商品可被無死角的檢閱，使得消費者在觀看直播後購買的意願增加，因此，欲探討在商家的直播當中消費者在不同時段及不同參與度的指標是否有群聚的效果，故選用 Facebook Live Sellers in Thailand 資料集，使用 K-means、Hierarchical clustering, 及 DBSCAN 研究消費者參與度指標是否有群聚性，並透過 Purity 進行評估，研究顯示 K-means 的 Purity 為 0.402，Hierarchical clustering 的 Purity 為 0.394，DBSCAN 的 Purity 為 0.0062，其中 K-means 的群聚效果最佳。

關鍵字：機器學習、K-means、Hierarchical、DBSCAN

一、 緒論

1.1 動機

1.1.1 Mini Twenty Newsgroups Dataset

在資訊爆炸的時代，我們只要打開手機、電腦並連上網路，大量的資料將顯現在我們面前，只要你想查詢甚麼資料，在網路的搜尋引擎幾乎能找到你想要的資料，而又因現代人經常瀏覽社群網站，如 Facebook、Instagram 及 Line 等等，一打開就會有朋友的動態、照片、連結等，甚至是各種新聞，而以新聞來看，目前新聞是處於一個娛樂化的時代[22]，雖然新聞已不再像以前具有高度價值，但裡頭的內容卻是更加的多元，寵物、美妝、旅遊及遊戲動漫各式各樣的新聞文章充斥在我們的生活裡，但是當你打開新聞網頁不見得會跳出自己想看的新聞類別，有可能是些不必要的新聞，如哪位藝人跟誰離婚、誰的車違停，這些跟我們自身生活不相關的新聞內容對於我們來說反而是增加了資訊的負荷，必須多花一點心力去過濾掉這些對自身毫無意義的新聞，因此本研究欲了解對於閱聽人來說，哪些新聞類別會有群聚的效果，以利將來作為個人化新聞推薦之基礎。

1.1.2 Facebook Live Sellers in Thailand Dataset

隨著社群媒體演算法的改變[1]，以 Facebook 臉書來說，當我們一打開社群的首頁時，除了能看見好友動態外，更多的是臉書推薦的內容，透過演算法，臉書能蒐集每一個人平常喜歡點閱的內容並進行分析，藉此將符合其興趣的內容推薦給使用者，也能對使用者進行精準行銷。而近幾年來臉書也開放所有的使用者都能使用直播功能，有了直播，就能即時拍攝並同步讓其他人觀看他人當下的畫面，也因此縮短了人與人之間的距離，更讓腦筋動得快的賣家看到了新的商機[2]，利用直播賣起商品，因臉書的病毒式傳播，讓直播被更多的人看見，而直播可以 360 度真實地呈現商品，賣家也會放大商品優點，且直播的商品通常會比實體店面的價格來的低，許多消費者就會在直播的留言區留下+1 購買[3]，當看見許多人留言購買時，也會覺得這商品是值得信任可被購買的[4]，因此透過直播賣商品的商機愈來愈不容小覷，越來越多的商家也開始加入直播的行列，而本研究欲了在商家的直播當中消費者的參與度得群聚效果。

1.2 目的

1.2.1 Mini Twenty Newsgroups Dataset

本研究欲探討哪些新聞類別會有群聚的效果，故利用 mini 20 Newsgroups 資料集，並透過 K-means、Hierarchical clustering 及 DBSCAN 研究新聞類別的分群，且以 Purity 進行評估，藉此更加快速了解新聞文章時各是屬於哪些新聞類別。

1.2.2 Facebook Live Sellers in Thailand Dataset

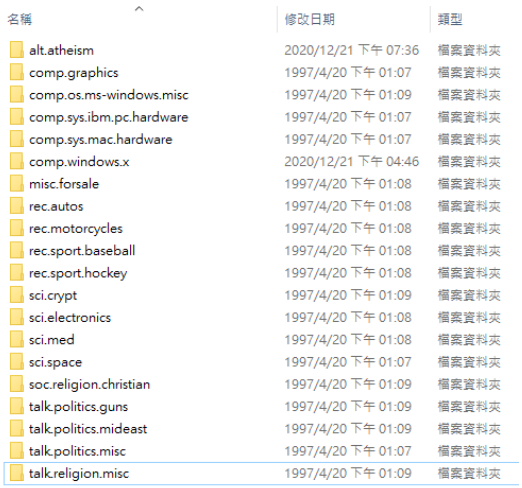
本研究欲探討在商家的直播當中消費者在不同時段及不同參與度的指標是否有群聚的效果，故選用 Facebook Live Sellers in Thailand 資料集，並使用 K-means、Hierarchical clustering 及 DBSCAN 進行分群，且以 Purity 進行評估，藉此更清楚消費者參與度指標是否有群聚性。

二、 資料集

2.1 真實資料集

2.1.1 Mini Twenty Newsgroups Dataset 說明

此資料集建立於 1999 年 9 月 9 日，有二十個新聞分類的資料夾，其中每個資料夾各有 100 個新聞文本，總共有 2000 筆資料，其內容為英文文字。



名稱	修改日期	類型
alt.atheism	2020/12/21 下午 07:36	檔案資料夾
comp.graphics	1997/4/20 下午 01:07	檔案資料夾
comp.os.ms-windows.misc	1997/4/20 下午 01:09	檔案資料夾
comp.sys.ibm.pc.hardware	1997/4/20 下午 01:07	檔案資料夾
comp.sys.mac.hardware	1997/4/20 下午 01:07	檔案資料夾
comp.windows.x	2020/12/21 下午 04:46	檔案資料夾
misc.forsale	1997/4/20 下午 01:08	檔案資料夾
rec.autos	1997/4/20 下午 01:08	檔案資料夾
rec.motorcycles	1997/4/20 下午 01:08	檔案資料夾
rec.sport.baseball	1997/4/20 下午 01:08	檔案資料夾
rec.sport.hockey	1997/4/20 下午 01:08	檔案資料夾
sci.crypt	1997/4/20 下午 01:09	檔案資料夾
sci.electronics	1997/4/20 下午 01:08	檔案資料夾
sci.med	1997/4/20 下午 01:08	檔案資料夾
sci.space	1997/4/20 下午 01:07	檔案資料夾
soc.religion.christian	1997/4/20 下午 01:09	檔案資料夾
talk.politics.guns	1997/4/20 下午 01:09	檔案資料夾
talk.politics.mideast	1997/4/20 下午 01:09	檔案資料夾
talk.politics.misc	1997/4/20 下午 01:07	檔案資料夾
talk.religion.misc	1997/4/20 下午 01:09	檔案資料夾

圖一 Mini Twenty Newsgroups Dataset 的二十個新聞分類資料夾

名稱	修改日期	類型	大小
51121	1997/4/20 下午 01:08	檔案	2 KB
51126	1997/4/20 下午 01:08	檔案	1 KB
51127	1997/4/20 下午 01:08	檔案	2 KB
51131	1997/4/20 下午 01:08	檔案	3 KB
51139	1997/4/20 下午 01:08	檔案	2 KB
51143	1997/4/20 下午 01:08	檔案	2 KB

圖二 Mini Twenty Newsgroups Dataset 的新聞文本檔案

```
|> [1] HOWEVER, I hate economic terrorism and political correctness
|> worse than I hate this policy.
|> [2] A more effective approach is to stop donating
|> to ANY organizating that directly or indirectly supports gay rights issues
|> until they end the boycott on funding of scouts.
Can somebody reconcile the apparent contradiction between [1] and [2]?
```

圖三 Mini Twenty Newsgroups Dataset 的新聞文本內容

2.1.2 Facebook Live Sellers in Thailand Dataset 說明

此資料集建立於 2019 年 4 月 22 日共有 7,051 筆資料，12 個欄位。

表一 Facebook Live Sellers in Thailand Dataset 欄位資料說明彙總表

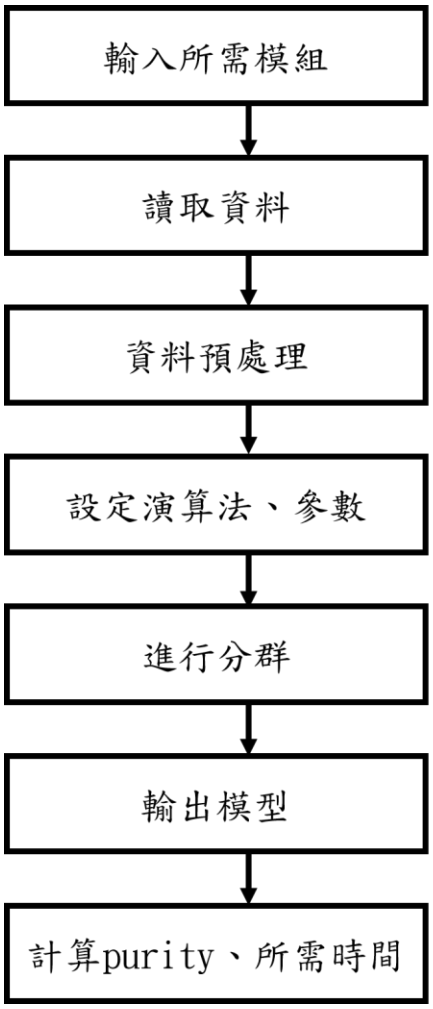
欄位名稱	欄位說明
status_id	文章的索引
status_type	文章的種類
status_published	文章張貼時間
num_reactions	文章被反應的次數
num_comments	文章被留言的次數
num_shares	文章被分享的次數
num_likes	文章被按喜歡的次數
num_loves	文章被按愛心的次數
num_wows	文章被按驚訝的次數
num_hahas	文章被按大笑的次數
num_sads	文章被按傷心的次數
num_angrys	文章被按生氣的次數

	A	B	C	D	E	F	G	H	I	J	K	L
1	status_id	status_type	status_publ	num_reacti	num_comn	num_share	num_likes	num_loves	num_wows	num_hahas	num_sads	num_angry
2	246675545	video	4/22/2018	529	512	262	432	92	3	1	1	0
3	246675545	photo	4/21/2018	150	0	0	150	0	0	0	0	0
4	246675545	video	4/21/2018	227	236	57	204	21	1	1	0	0
5	246675545	photo	4/21/2018	111	0	0	111	0	0	0	0	0
6	246675545	photo	4/18/2018	213	0	0	204	9	0	0	0	0
7	246675545	photo	4/18/2018	217	6	0	211	5	1	0	0	0
8	246675545	video	4/18/2018	503	614	72	418	70	10	2	0	3
9	246675545	video	4/17/2018	295	453	53	260	32	1	1	0	1
10	246675545	photo	4/17/2018	203	1	0	198	5	0	0	0	0
11	246675545	photo	4/11/2018	170	9	1	167	3	0	0	0	0
12	246675545	photo	4/10/2018	210	2	3	202	7	1	0	0	0

圖四 Facebook Live Sellers in Thailand Dataset 資料集部份內容

三、 方法

3.1 實作說明



- 匯入所需套件：numpy、pandas、sklearn等套件
- 讀取Live資料集和mini 20 Newsgroups資料集
- 將資料集進行正規化和刪除空白列
- 設定不同演算法，分別為K-means、階層式分群、DBSCAN，調整相關參數
- 將資料集進行分群
- 將分群好的資料及匯出成圖表
- 利用purity和分群所需時間來評估績效

圖五 實作說明圖

3.2 操作說明

分群方法大多都屬於非監督式學習，在學習過程時沒有特定標準答案，本組所使用的方法為 K-means、DBSCAN 以及階層式分群，下面會進行較詳細的介紹。

表二 演算法說明表

方法	說明
K-means	先設定將資料分成 K 群，在初次分群時會隨機給予 K 個群心。而後計算全心到資料點的直線距離，計算完成後將資料點分給最小距離的群心，並且更新群心，如此重複直到不再會有太大變動則停止。
DBSCAN	DBSCAN 基於密度來進行分群的演算法，會將相鄰的資料點分為一群，並標註低密度區域的邊界點，此演算法將雜訊點進行剔除。
階層式分群	利用階層架構的方式進行分群，央資料一層一層進行分裂或聚集，最後產生樹狀結構。

四、 實驗

4.1 前置處理

4.1.1 Mini Twenty Newsgroups Dataset

本組先讀取資料夾中的所有文件檔案，而後一一將檔案打開存取治所設定的變數中。後續計算文字出現的次數並且將文字轉換為向量，方便演算法的進行。

4.1.2 Facebook Live Sellers in Thailand Dataset

將 CSV 檔案中的空白欄位「column_1」、「column_2」、「column_3」和「column_4」刪除。由於特徵「id」不具有統計分析意義，因此本組將此欄位進行刪除。而後本組將特徵「status_type」進行正規化同時也轉換資料型態，以利後續演算法的進行。

4.2 實驗設計

4.2.1 Mini Twenty Newsgroups Dataset

本組利用三種方法來進行評估，並且計算分群所需時間和 Purity 來評估績效。K-means 將分群數設定為 20 群，並且得到分群結果。階層式分群將

分群設定為 20 群並且計算距離。DBSCAN 參數設定 eps 為 1，min_samples 為 20。

4.2.2 Facebook Live Sellers in Thailand Dataset

本組利用三種方法來進行評估，並且計算分群所需時間和 Purity 來評估績效。K-means 將分群數設定為 20 群，計算分成幾群會達到最佳效果。DBSCAN 參數設定 eps 為 1，min_samples 為 20。

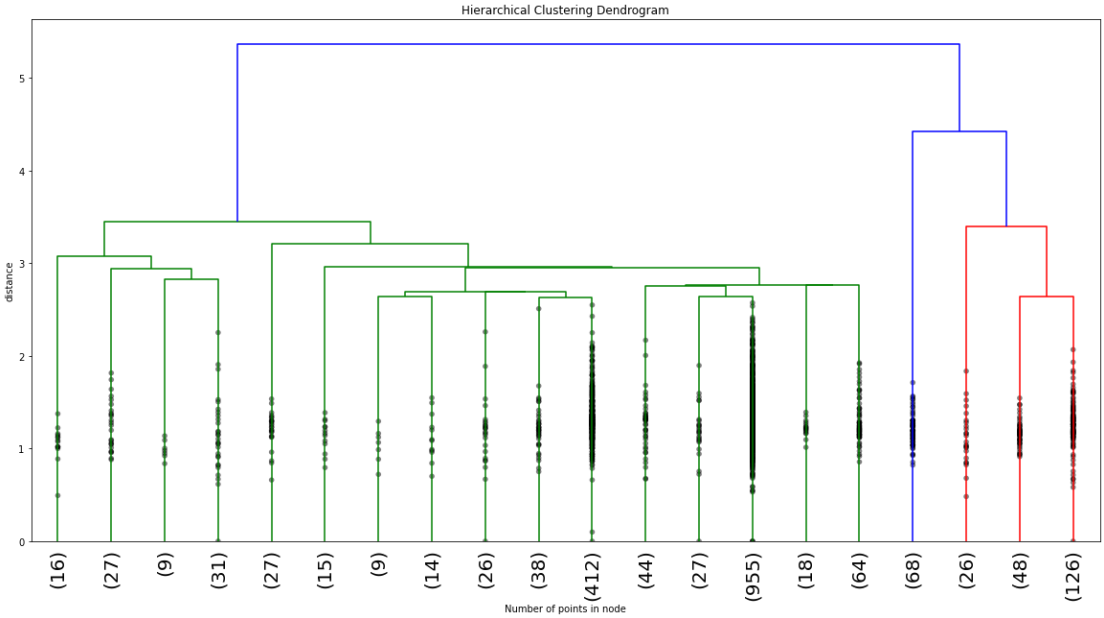
4.3 實驗結果

4.3.1 Mini Twenty Newsgroups Dataset

本研究使用 K-means、DBSCAN、階層式分群方法來進行，由下表可得知 K-means 的效果最好，運算時間為 117.7 秒，其次為階層式分群，純度為 0.394，運算時間為 102.8 秒，樹狀結構圖如下圖所示。

表三 Mini Twenty Newsgroups Dataset 績效評估彙總表

方法	purity	運算時間(秒)
K-means	0.30005	117.7
DBSCAN	0.05	0.5787
階層式分群	0.07	102.8



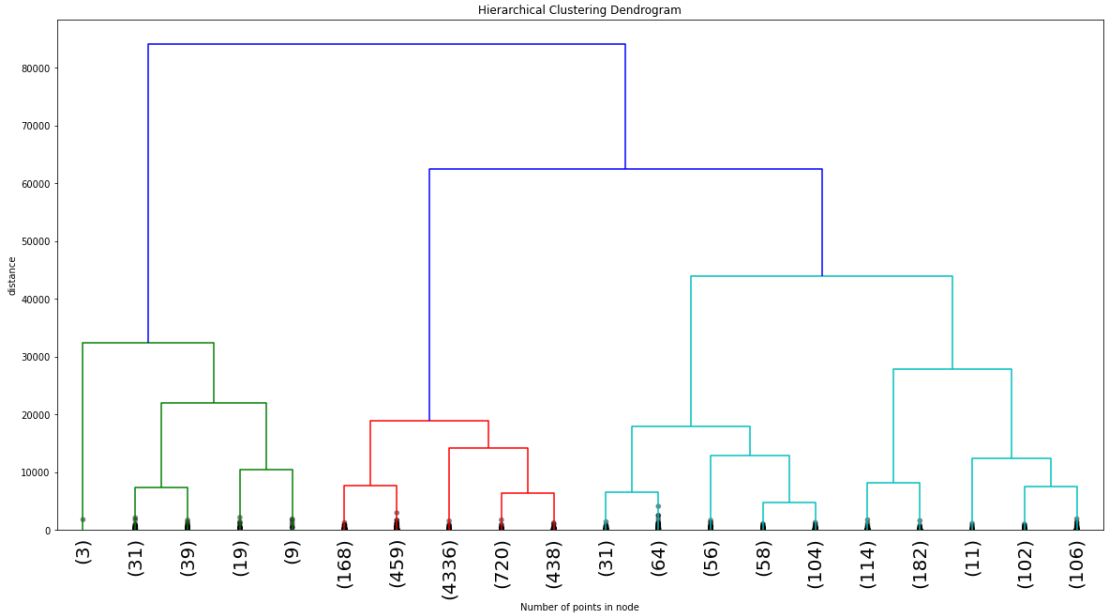
圖六 Mini Twenty Newsgroups Dataset 階層式分群的階層樹

4.3.2 Facebook Live Sellers in Thailand Dataset

本研究使用 K-means、DBSCAN、階層式分群方法來進行，由下表可得知 K-means 分為 20 群的純度最高，運算時間為 2.169 秒，其次為階層式分群，純度為 0.394，運算時間為 2.8 秒，樹狀結構如下圖所示。

表四 Facebook Live Sellers in Thailand Dataset 績效評估彙總表

方法	purity	運算時間(秒)
K-means	0.402	2.169
DBSCAN	0.0062	0.023523
階層式分群	0.394	2.8



圖七 Facebook Live Sellers in Thailand Dataset 階層式分群的階層樹

五、 結論

5.1 Mini Twenty Newsgroups Dataset

本研究利用 K-means、DBSCAN 以及階層式分群以上三種方法，對 mini 20 Newsgroups 資料集進行分群。研究顯果顯示 K-means 績效最好，第二則是階層式分群，由研究結果可知若有研究需要進行分群可以優先長式 K-means 方法。

5.2 Facebook Live Sellers in Thailand Dataset

本研究利用三種方法對資料集進行分群，經過 Live 資料集發現 K-means 對於分群的績效最高，運算時間也非常短暫，其次是階層式分群，雖然階層式分群與 K-means 的 purity 績效指標差異微小，但階層式分群所花費時間較長，因此未來若有相關分群方法建議優先選擇 K-means。

六、 參考資料

[1] 社群媒體千變萬化的演算法，比另一半的心更難猜
https://www.inside.com.tw/article/21681-social-media-Algorithm

[2] 只是社交平台的臉書「直播拍賣」正夯：一晚進帳破百萬，賣得最好的都是「高單價」的
https://www.thenewslens.com/article/104708

[3] 臉書直播賣商品正夯，如何擺脫數位行銷框架銷售商品？
http://imarketing.iwant-in.net/?p=5454

[4] 臉書直播趨勢分析：人氣最高的不是美妝，而是賣運動鞋
https://www.thenewslens.com/article/83459

[5] UCI-Facebook Live Sellers in Thailand Data Set 資料集
https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups

[6] 階層式分群介紹
https://chih-sheng-huang821.medium.com/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-%E9%9B%86%E7%BE%A4%E5%88%86%E6%9E%90-k-means-clustering-e608a7fe1b43

[7] 集群分析介紹
https://chih-sheng-huang821.medium.com/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-%E9%9B%86%E7%BE%A4%E5%88%86%E6%9E%90-k-means-clustering-e608a7fe1b43

[8] 層次聚類
https://www.itread01.com/content/1541618366.html

[9] sklearn—CountVectorizer 詳解
https://www.itread01.com/content/1547571635.html

[10] Python 路徑檢查
https://blog.gtwang.org/programming/python-howto-check-whether-file-folder-exists/

[11] Python 純度指標
https://cloud.tencent.com/developer/ask/189986

[12] 階層式分群介紹
https://chih-sheng-huang821.medium.com/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-%E9%9B%86%E7%BE%A4%E5%88%86%E6%9E%90-k-means-clustering-e608a7fe1b43

[13] 集群分析介紹

https://chih-sheng-huang821.medium.com/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-%E9%9B%86%E7%BE%A4%E5%88%86%E6%9E%90-k-means-clustering-e608a7fe1b43

[14] 層次聚類
https://www.itread01.com/content/1541618366.html

[15] sklearn—CountVectorizer 詳解
https://www.itread01.com/content/1547571635.html

[16] Python 路徑檢查
https://blog.gtwang.org/programming/python-howto-check-whether-file-folder-exists/

[17] Python 純度指標
https://cloud.tencent.com/developer/ask/189986

[18] K-means 分群法
http://mirlab.org/jang/books/dcpr/kMeans.asp?title=3-3%20K-means%20%A4%C0%B8s%AAk

[19] K-means 和 K-means++的演算法原理及 sklearn 庫中參數解釋、選擇
https://blog.csdn.net/github_39261590/article/details/76910689

[20] 機器學習（7），分群/聚類：階層式分群
https://mropengate.blogspot.com/2015/06/ai-ch17-6-clustering-hierarchical.html

[21] 淺談聚合式階層分群法與熱圖
https://yourgene.pixnet.net/blog/post/117264518-%E6%B7%BA%E8%AB%87%E8%81%9A%E5%90%88%E5%BC%8F%E9%9A%8E%E5%B1%A4%E5%88%86%E7%BE%A4%E6%B3%95%E8%88%87%E7%86%B1%E5%9C%96

[22] 新聞娛樂化衝擊 媒體喪失社會公益性
http://shuj.shu.edu.tw/blog/2019/12/23/%E6%96%B0%E8%81%9E%E5%A8%9B%E6%A8%82%E5%8C%96%E8%A1%9D%E6%93%8A-%E5%AA%92%E9%AB%94%E5%96%AA%E5%A4%B1%E7%A4%BE%E6%9C%83%E5%85%AC%E7%9B%8A%E6%80%A7/