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**

　　隨著社群媒體演算法的改變，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，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，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，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。而本研究欲探討在商家的直播當中消費者的參與度，經由K-means、Hierarchical、DBSCAN研究不同時段及不同參與度的指標是否有群聚的效果，並透過Purity進行評估，研究顯示K-means的Purity為0.402，Hierarchical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

1. **緒論**

1.1動機

1.1.1 Mini Twenty Newsgroups Dataset

1.1.2 Facebook Live Sellers in Thailand Dataset

1.2目的

1.2.1 Mini Twenty Newsgroups Dataset

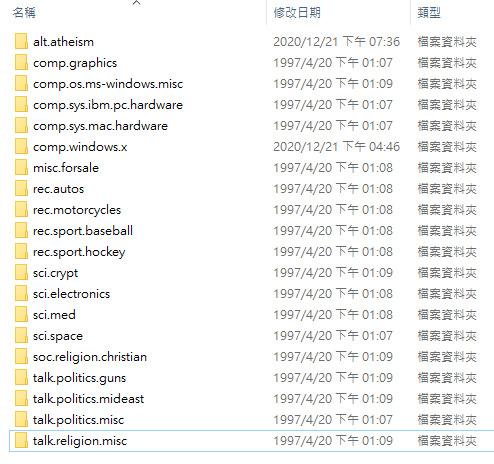
1.2.2 Facebook Live Sellers in Thailand Dataset

1. **資料集**

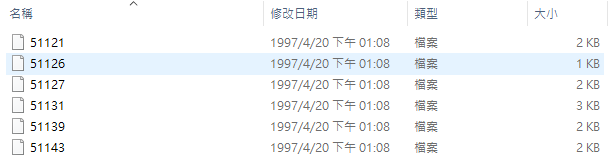
2.1真實資料集

2.1.1 Mini Twenty Newsgroups Dataset說明

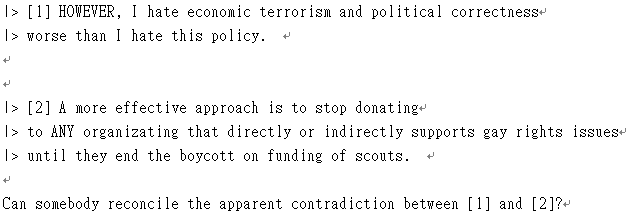
此資料集建立於1999年9月9日，有二十個新聞分類的資料夾，其中每個資料夾各有100個新聞文本，總共有2000筆資料，其內容為英文文字。



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



圖二 Mini Twenty Newsgroups Dataset的新聞文本檔案



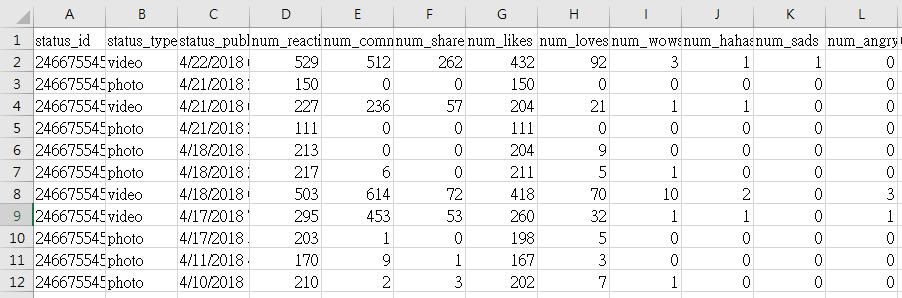
圖三 Mini Twenty Newsgroups Dataset的新聞文本內容

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

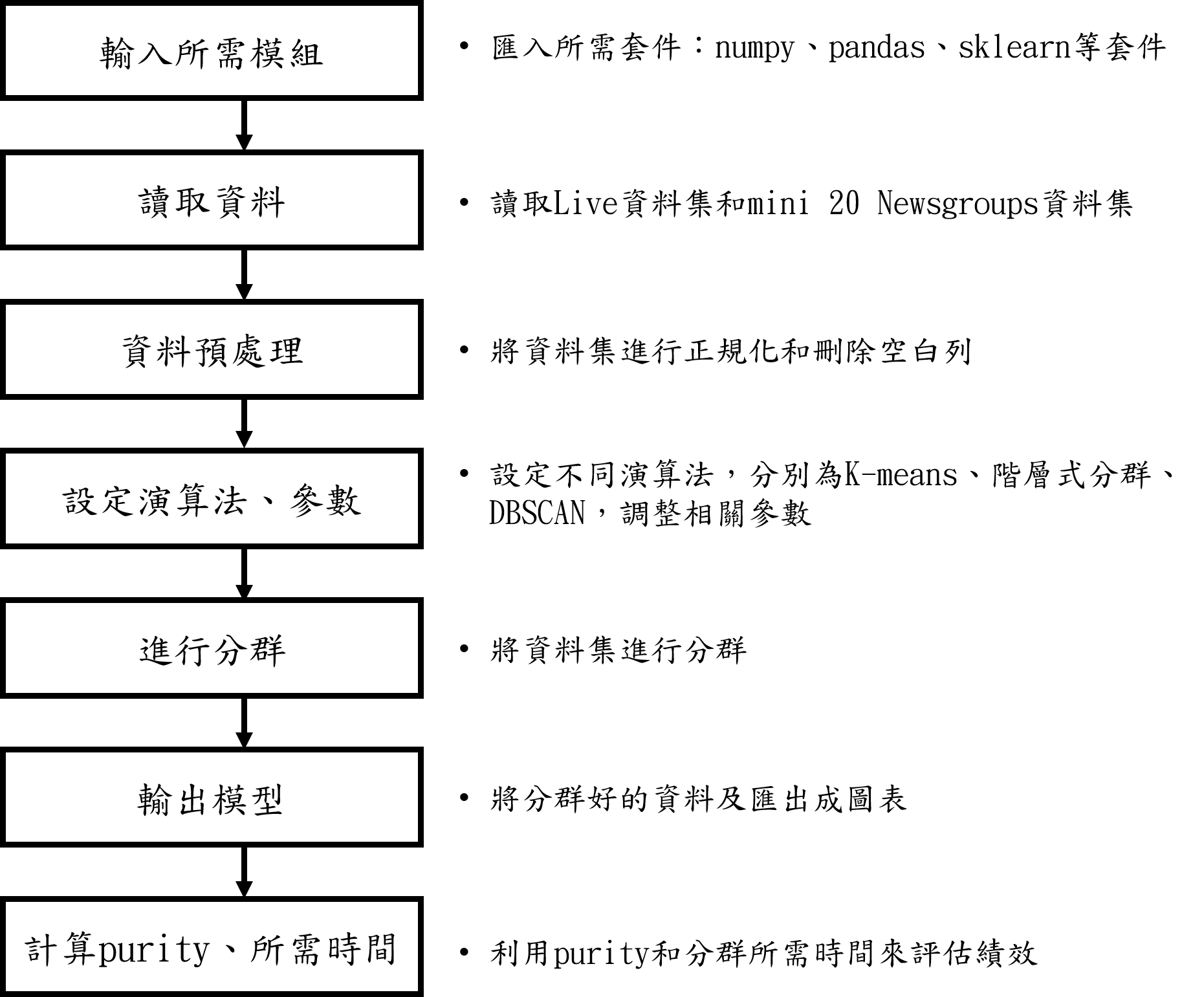
|  |  |
| --- | --- |
| 欄位名稱 | 欄位說明 |
| status\_id | 文章的索引 |
| status\_type | 文章的種類 |
| status\_published | 文章張貼時間 |
| num\_reactions | 文章被反應的次數 |
| num\_comments | 文章被留言的次數 |
| num\_shares | 文章被分享的次數 |
| num\_likes | 文章被按喜歡的次數 |
| num\_loves | 文章被按愛心的次數 |
| num\_wows | 文章被按驚訝的次數 |
| num\_hahas | 文章被按大笑的次數 |
| num\_sads | 文章被按傷心的次數 |
| num\_angrys | 文章被按生氣的次數 |



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

1. **方法**

3.1實作說明



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。 |

1. **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。

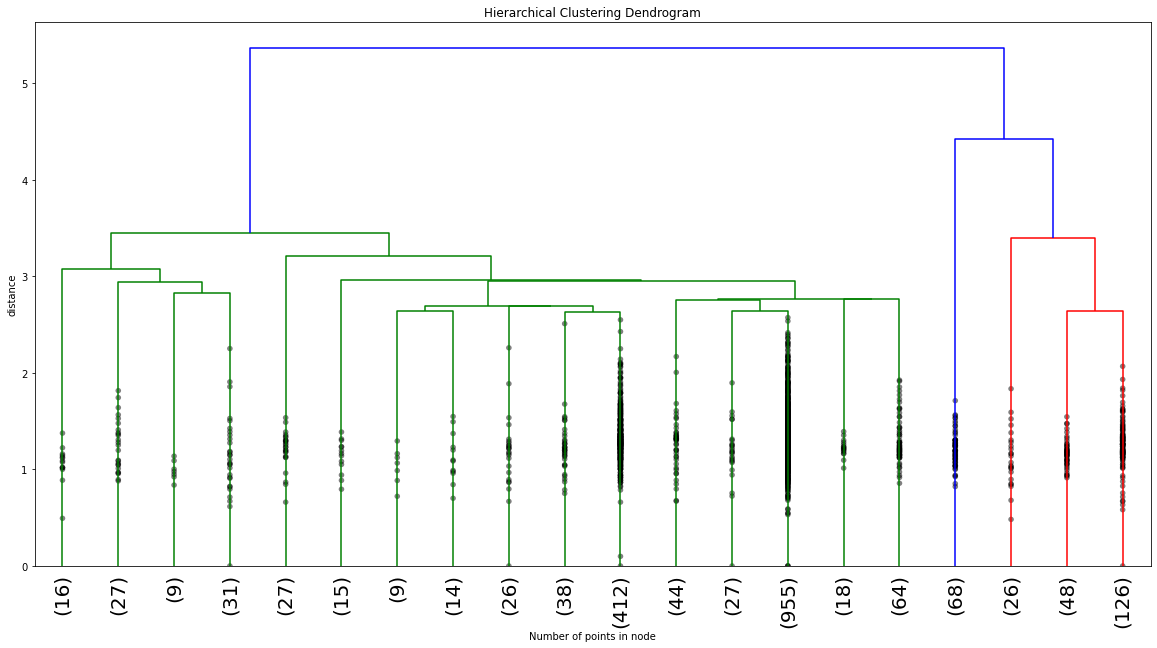
4.3實驗結果

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 |



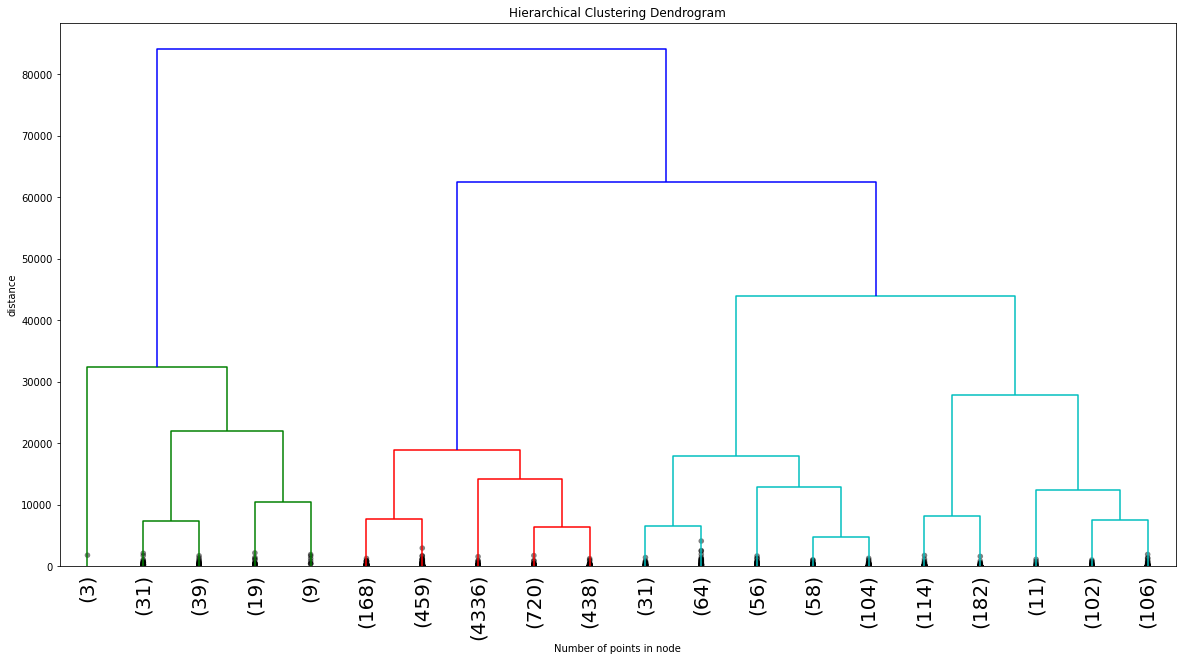
圖五 Mini Twenty Newsgroups Dataset階層式分群的階層樹

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 |



圖五 Facebook Live Sellers in Thailand Dataset階層式分群的階層樹

1. **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。

1. **參考資料**
2. 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

1. 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

1. 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

1. 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>

1. UCI-Facebook Live Sellers in Thailand Data Set資料集

<https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups>

1. 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>

1. 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>

1. 層次聚類

<https://www.itread01.com/content/1541618366.html>

1. sklearn—CountVectorizer詳解

<https://www.itread01.com/content/1547571635.html>

1. Python 路徑檢查

<https://blog.gtwang.org/programming/python-howto-check-whether-file-folder-exists/>

1. Python 純度指標

<https://cloud.tencent.com/developer/ask/189986>

1. 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>

1. 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>

1. 層次聚類

<https://www.itread01.com/content/1541618366.html>

1. sklearn—CountVectorizer詳解

<https://www.itread01.com/content/1547571635.html>

1. Python 路徑檢查

<https://blog.gtwang.org/programming/python-howto-check-whether-file-folder-exists/>

1. Python 純度指標

<https://cloud.tencent.com/developer/ask/189986>

1. K-means分群法

<http://mirlab.org/jang/books/dcpr/kMeans.asp?title=3-3%20K-means%20%A4%C0%B8s%AAk>

1. 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>

1. 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>

1. 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