資料探勘 專案作業三

* 利用Python軟體實作群聚分析

1. 使用mini 20 Newsgroups資料集：[mini\_newsgroups.tar.gz](http://kdd.ics.uci.edu/databases/20newsgroups/mini_newsgroups.tar.gz)。
2. 分別使用K-means、階層式分群、DBSCAN，將資料分成20群，並比較分群所花費時間，自選評估指標比較分群結果品質，包括使用Entropy指標。
3. 請劃出階層式分群的階層樹(Dendrogram)。

* 專案程式直接上傳GitLab ( <http://140.125.84.84/> )，專案名稱：hw3+組長學號，並新增助教為專案成員，以利下載評分。
* 請遵循學術論文寫作格式撰寫專案報告，內容至少必須包括下列要項：

**摘要** (學術論文摘要通常不分段，內容簡要地涵蓋動機、目的、方法與實驗結果)

**一、緒論**

**1.1動機** (探勘所選用的資料集之動機)

**1.2目的** (分析該資料集之目的)

**二、資料集**

**2.1 真實資料集** (簡介包括名稱、筆數、欄位、型態…、並以資料表格呈現部份資料內容)**三、方法**

**3.1實作說明** (簡要說明實作過程)

**3.2操作說明** (說明如何操作繳交的程式)

**四、實驗**

(分別探討實驗設計與實驗結果)

**4.1前置處理** (說明對資料集做了甚麼前置處理 )

**4.2實驗設計** (實驗如何進行、參數如何設定以及使用的評估指標等）

(盡可能運用課堂上學過的評估指標)

**4.3實驗結果** (以圖表呈現結果、分析結果及解讀)

(應分析績效好壞之原因、比較不同分群方法之執行時間、實驗結果以表格或圖形統整)

**五、結論**

**六、參考文獻**

* 評分標準
  + 內容(難度、豐富性、方法正確性)：60%，文件格式：20%，寫作技巧：20%
* **任何複製剪貼網站資料或學長姐以前作業，皆屬抄襲，嚴重者以零分計。引用網站或他人資料，應註明引用來源。**
* 請遵循投稿論文寫作格式。使用單欄式，無需製作論文目錄，但每頁需標註頁碼。
* 紙本於上課前繳交或直接放入許老師的系辦信箱。電子檔繳交到網路學園。缺紙本或電子檔都視為未完成繳交作業。電子檔命名：HW3-組長學號。
* 繳交紙本時，在左上角訂一訂書針即可，無須其他任何裝訂。
* 繳交日期：108/12/12（四）；每遲交一天，扣該次作業成績10分。