2022 電子商務技術作業四

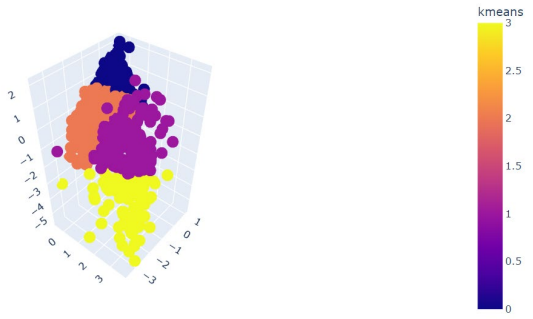
一. **請用 python 實作以下問題，截圖程式碼與執行結果加上說明文字**
以下問題皆使用 wu_song.csv 資料集

1. 載入資料並刪除除了"energy", "speechiness", "acousticness", "instrumentalness", "loudness","tempo","danceability",'valence', "liveness" 以外之欄位(使用 pandas dataframe) (2%)

2. 將剩下的欄位做特徵篩選的動作，並使用 kmeans silhouette analysis 的方法找出在哪三個欄位的情況下(需考慮所有組合)，分 X 群會有最高的 silhouette score。請找出 X 與 silhouette score 還有是哪三個欄位。(20%)
請解釋 silhouette 分析法 與 elbow 轉折判斷法的差別(3%)
注意事項：
每次丟進去 fit 時都要將資料標準化。(StandardScaler)
random_state 皆設為 15
X 的範圍落在 2~12 之間

以下題目資料都須經過標準化才做分群喔!

3. 使用剛剛找出來的欄位用 k-means 做分群。超參數設定為 n_cluster=4 , random_state=15。
並使用 plotly 繪製出 3d 圖形如以下所示(15%)：
注意要有欄位名稱，也就是剛剛找出來的那三個
(下圖沒有是因為放了你們就知道是哪三個欄位了)



4. 使用剛剛找出來的欄位用 Meanshift 做分群(15%)
請找出最佳的 estimate_bandwidth.超參數設定為 random_state=15, quantile=0.32, n_samples=1000

使用剛剛找出的 estimate_bandwidth 做分群並繪製如第三題的圖

5. 使用剛剛找出來的欄位用 k-prototypes 做分群。超參數設定為 n_cluster=4 random_state=15,init='Huang',verbose=0。
   並使用 plotly 繪製出 3d 圖形如第三題的圖(15%)

6. 使用剛剛找出來的欄位用 k-modes 做分群。超參數設定為 n_cluster=4 random_state=15,init='Huang',verbose=0。
   並使用 plotly 繪製出 3d 圖形如第三題的圖(15%)

7. 請比較說明上述四種分群法的差異(5%)

二. 使用 weka 做分群

使用 weka 中的 simplekmeans 將 wu_songs.arff 做分群。請使用你用 python 找到的最佳三個欄位做分群，並截圖結果圖表。Cluster 數量設為 4, seed 設為 15。(10%)

繳交期限：2022/4/13

請繳交 pdf 與 ipynb 檔，檔名為 ECT_HW4_學號

遲交一天扣 5%，最多扣 50%

- 補充：

所需套件：plotly 安裝方式：pip install plotly