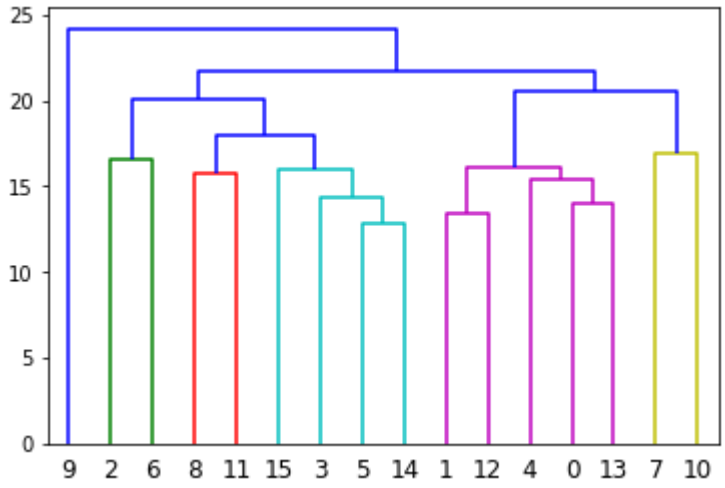
Oscar Eduardo Flores Hernández

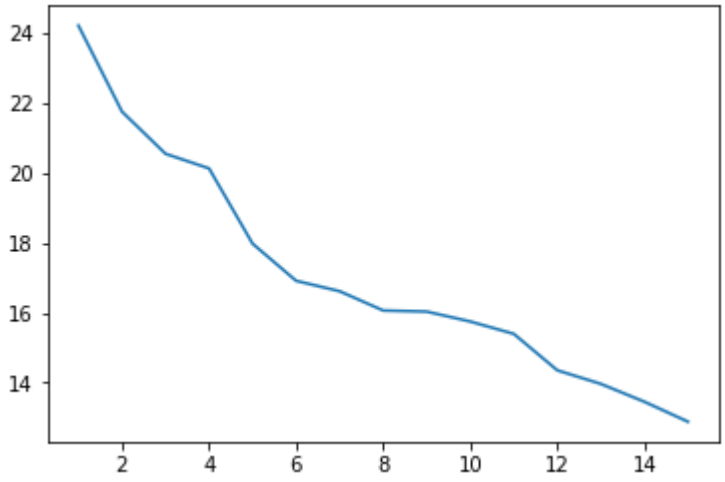
IF715029

Hierarchical clustering.

Se analiza la base de datos no estandarizada (debido a que el rango en el cuál se obtienen los datos va de 0 a 5 en todos los casos) con hierarchical clustering en método completo y utilizando la métrica de distancia euclidiana (debido a que son datos ordinales, dónde una calificación de 1 es menor a 5). Los resultados obtenidos se pueden representar en el diagrama siguiente:



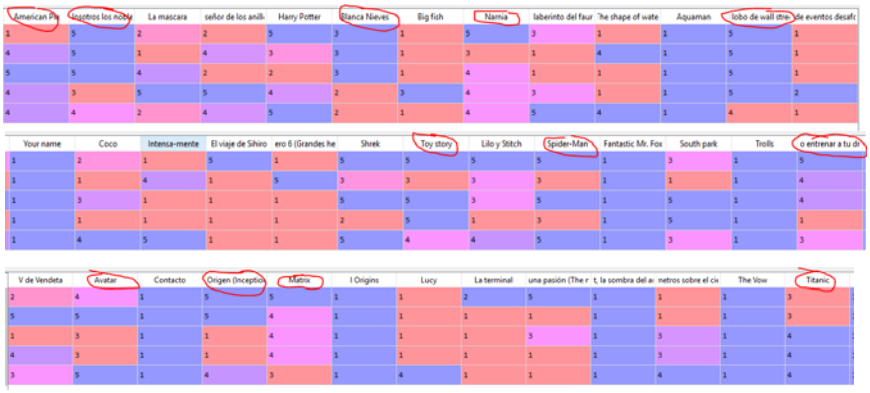
Cada uno de los números de abajo es el 'ID' de los usuarios clasificados. Se puede apreciar que existen dos grandes grupos y un compañero 'radical'; el 9 se encuentra alejado de todos, del 2 al 14 (en orden del dendograma) representarían el primer grupo y del 1 al 10 el segundo grupo. Si se desearan más particiones podríamos llegar a aquellas que se encuentran de colores, en ese punto tendríamos 6 grupos similares, dónde el más grande es el morado.



Se busca confirmar que los grupos están correctamente formados, para ello hacemos una gráfica de codos y encontramos que existe un posible codo en 3 y 6 (como lo habíamos notado visualmente antes de hacer algún calculo). A pesar de la existencia de los codos; la diferencia entre cada uno de los usuarios es bastante grande.

En la imagen a la derecha se analiza la pertenencia del usuario en los grupos. En este caso Oscar Flores (usuario 1) pertenece al grupo 4, por lo tanto, usuarios 0, 4, 12 y 13 deberían tener gustos similares.

Una vez que conocemos los usuarios similares observamos las películas de los 5.



Se puede notar que existen algunas películas en dónde la mayoría coincide. No obstante, hay otras en las que ninguno entra en acuerdo con los demás usuarios, se cree que esta disparidad ocurre debido a que algún usuario no vio alguna película (y la calificó con 1), mientras que los demás le dieron puntuaciones altas o porque estamos forzando al sistema a hacer grupos en dónde no existen usuarios realmente similares (muy pocos datos).

	0
0	4
1	4
2	1
3	3
4	4
5	3
6	1
7	5
8	2
9	6
10	5
11	2
12	4
13	4
14	3
15	3

Código:

```
import pandas as pd
import numpy as np
import scipy.cluster.hierarchy as hi
import matplotlib.pyplot as plt

#%% Seleccionar datos
data = pd.read_excel('../data/Test de películas(1-16).xlsx')

csel = np.arange(6,243,3)
cnames = list(data.columns.values[csel])
datan = data[cnames]

#%% Dendograma
Z = hi.linkage(datan,metric='euclidean',method='complete')

#%% Grafica de dendograma
hi.dendrogram(Z)

#%% Grafica de codo
last = Z[:,2]
last = last[::-1]
plt.plot(np.arange(len(last))+1,last)

#%% Usuario 1: Oscar Flores
sim = hi.fcluster(Z,6,criterion='maxclust') #pertenezco al grupo 4
pel_sim = datan[sim==4]
```