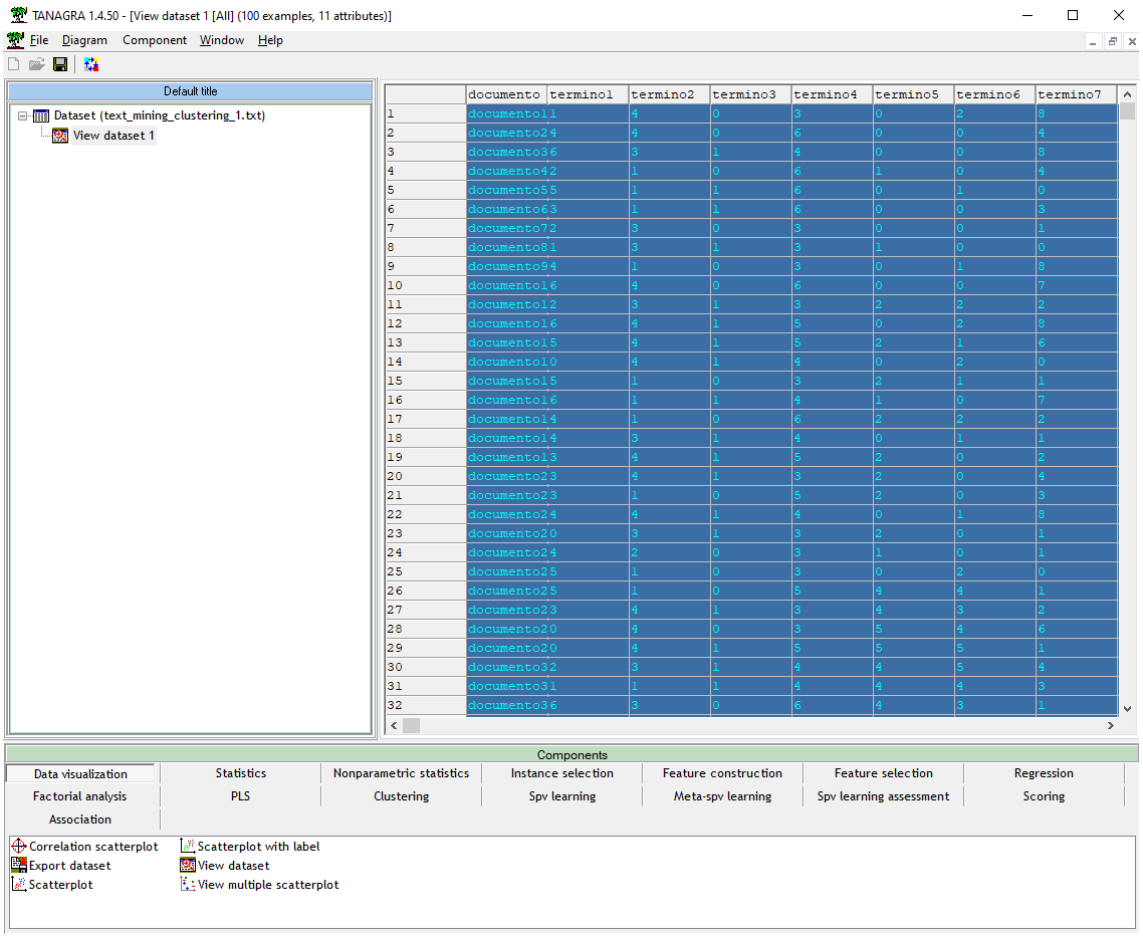
16 Junio 2020 Text Mining Clustering. Tanagra. Parte 3

Datos:text_mining_clustering_1.txt

Presentar un tutorial en que se explique el desarrollo de la tarea

1. **Utilizando los atributos término1 y término3 aplicar clustering jerárquico con las opciones por defecto. Presentar los resultados obtenidos. ¿Qué representa en dendrograma? ¿Cuántos clusters se presentan y porqué? ¿Cuántos elementos tiene cada cluster?**
2. **Visualizar los clusters en un diagrama de dispersión.**
3. **Ejecutar el paso 2 marcando en los parámetros Tree structure. Explicar lo que se presenta.**
4. **Generar clusters utilizando las variables término1 .... término10. Presentar los resultados obtenidos.**

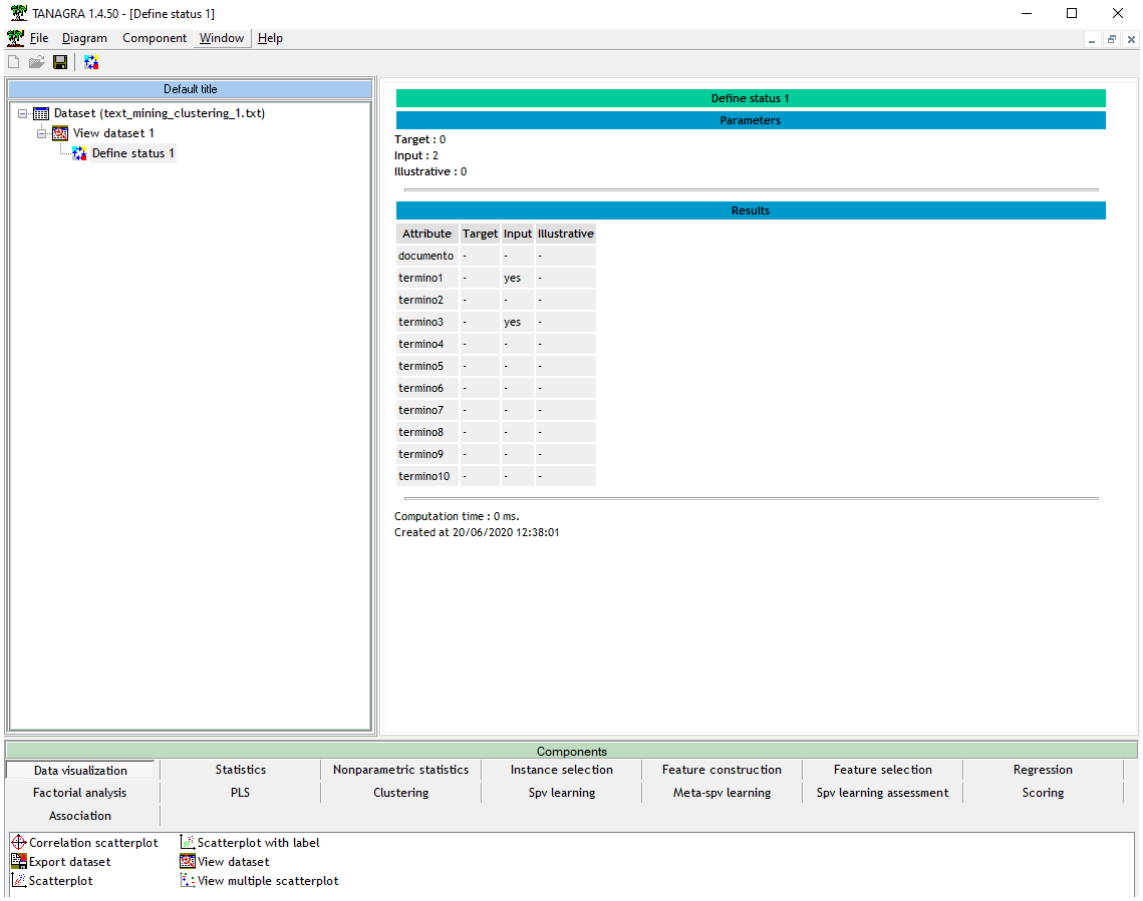
Abrimos la herramienta Tanagra. Seleccionamos File→New.. y en la ventana que nos aparece buscamos el archivo text_mining_clustering_1. Lo seleccionamos y abrimos, y vemos los datos con los que vamos a trabajar, arrastrando View Dataset desde la pestaña inferior Data Visualization, hasta colgarlo debajo de nuestro Dataset:



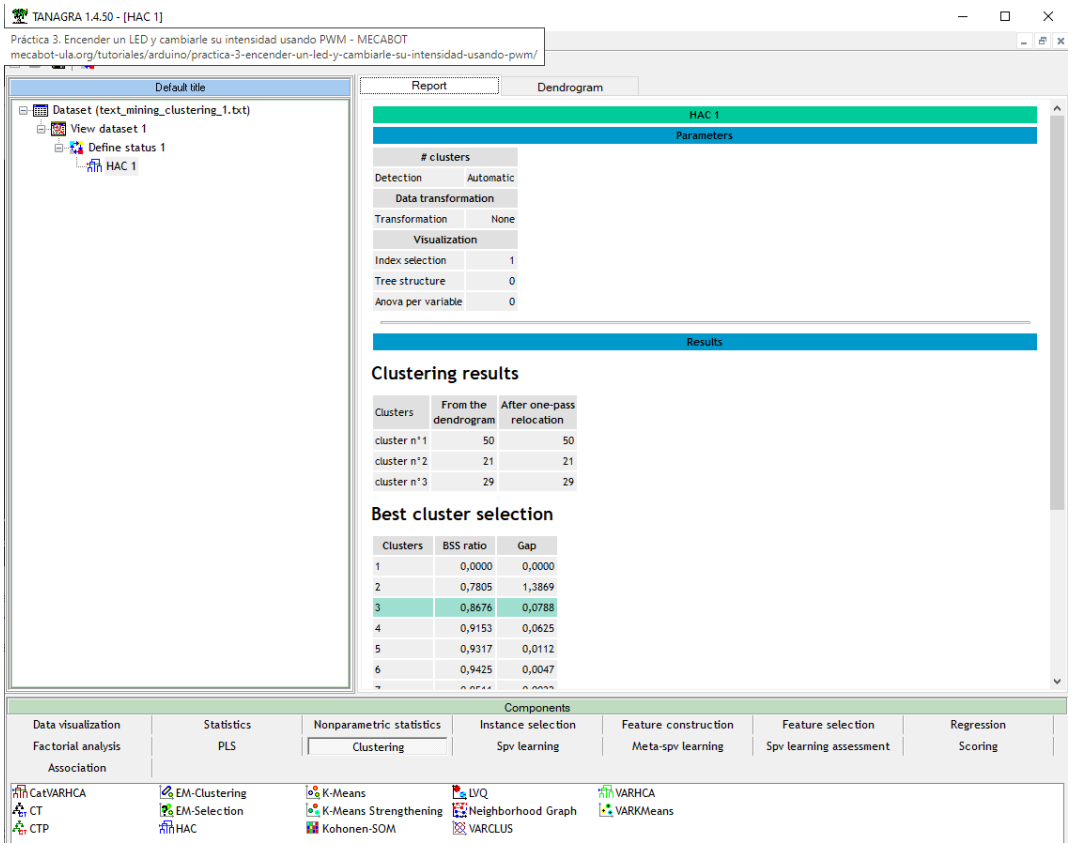
	documento	termino1	termino2	termino3	termino4	termino5	termino6	termino7
1	documento11	4	0	3	0	2	8	
2	documento24	4	0	6	0	0	4	
3	documento36	3	1	4	0	0	8	
4	documento42	1	0	6	1	0	4	
5	documento55	1	1	6	0	1	0	
6	documento63	1	1	6	0	0	3	
7	documento72	3	0	3	0	0	1	
8	documento81	3	1	3	1	0	0	
9	documento94	1	0	3	0	1	8	
10	documento16	4	0	6	0	0	7	
11	documento12	3	1	3	2	2	2	
12	documento16	4	1	5	0	2	8	
13	documento15	4	1	5	2	1	6	
14	documento10	4	1	4	0	2	0	
15	documento15	1	0	3	2	1	1	
16	documento16	1	1	4	1	0	7	
17	documento14	1	0	6	2	2	2	
18	documento14	3	1	4	0	1	1	
19	documento13	4	1	5	2	0	2	
20	documento23	4	1	3	2	0	4	
21	documento23	1	0	5	2	0	3	
22	documento24	4	1	4	0	1	8	
23	documento20	3	1	3	2	0	1	
24	documento24	2	0	3	1	0	1	
25	documento25	1	0	3	0	2	0	
26	documento25	1	0	5	4	4	1	
27	documento23	4	1	3	4	3	2	
28	documento20	4	0	3	5	4	6	
29	documento20	4	1	5	5	5	1	
30	documento32	3	1	4	4	5	4	
31	documento31	1	1	4	4	4	3	
32	documento36	3	0	6	4	3	1	

Queremos aplicar clustering con las variables de entrada término1 y término3. Como ya sabemos, clustering es una técnica de aprendizaje no supervisada, por lo que no tendremos variable objetivo (target).

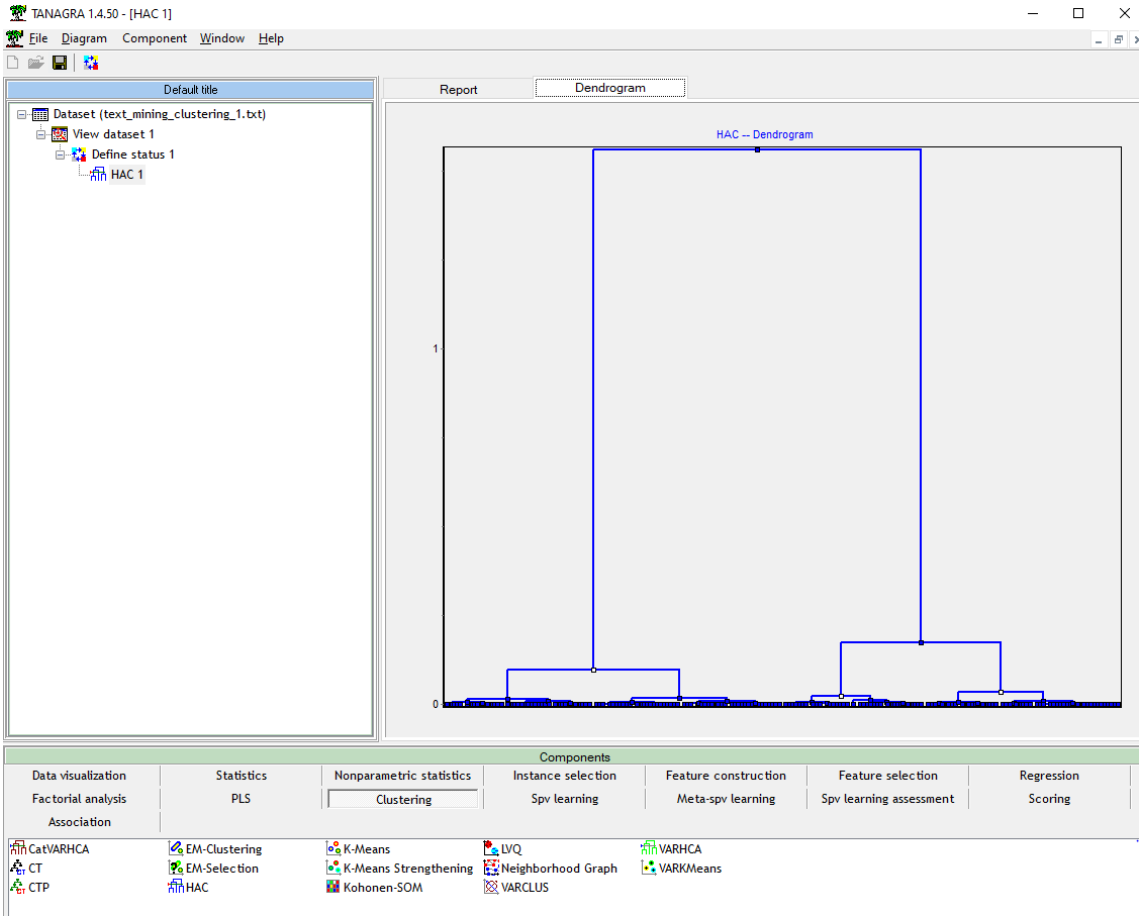
Por lo tanto, creamos un Define Status con termino 1 y término3 como inputs.



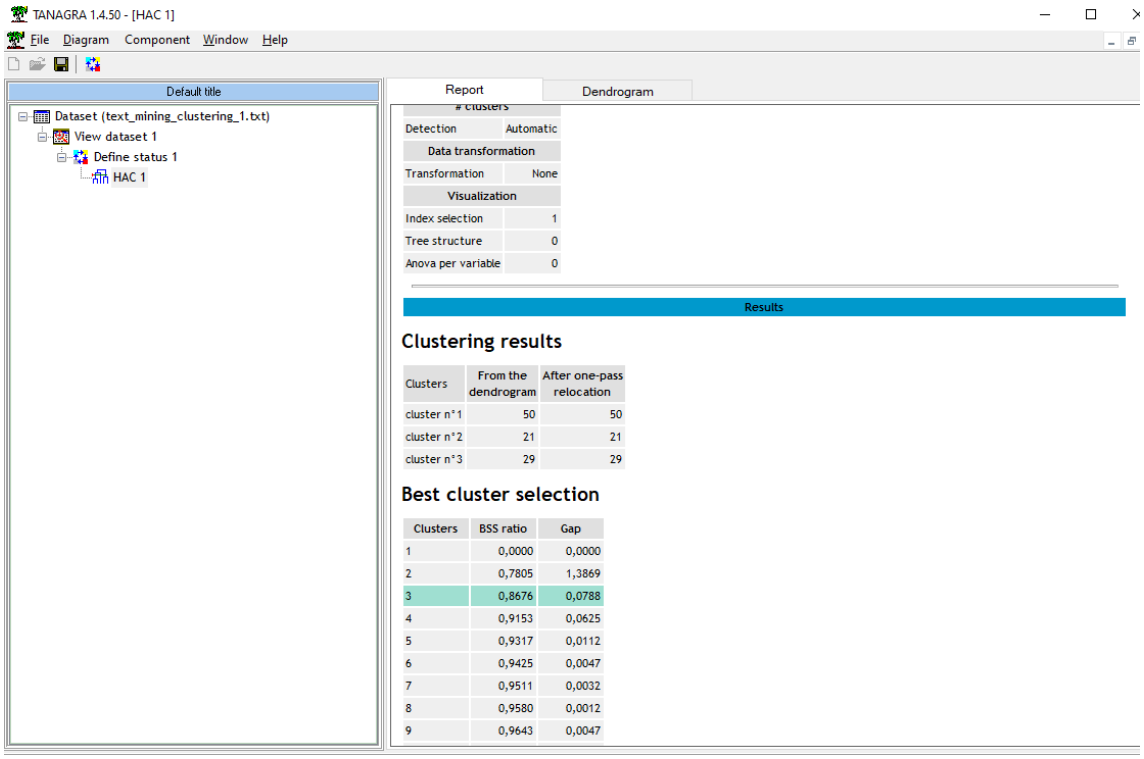
En la pestaña Clustering, seleccionaremos HAC y lo arrastraremos hasta colgarlo debajo de nuestro Define Status. Lo ejecutaremos con las opciones por defecto.



Si hacemos click en la pestaña Dendogram, se nos presentará el dendograma. En el dendograma podemos observar que los datos están unidos, a unas distancias muy pequeñas. Es una representación de como se van agrupando los datos según la semejanza que tengan entre ellos.

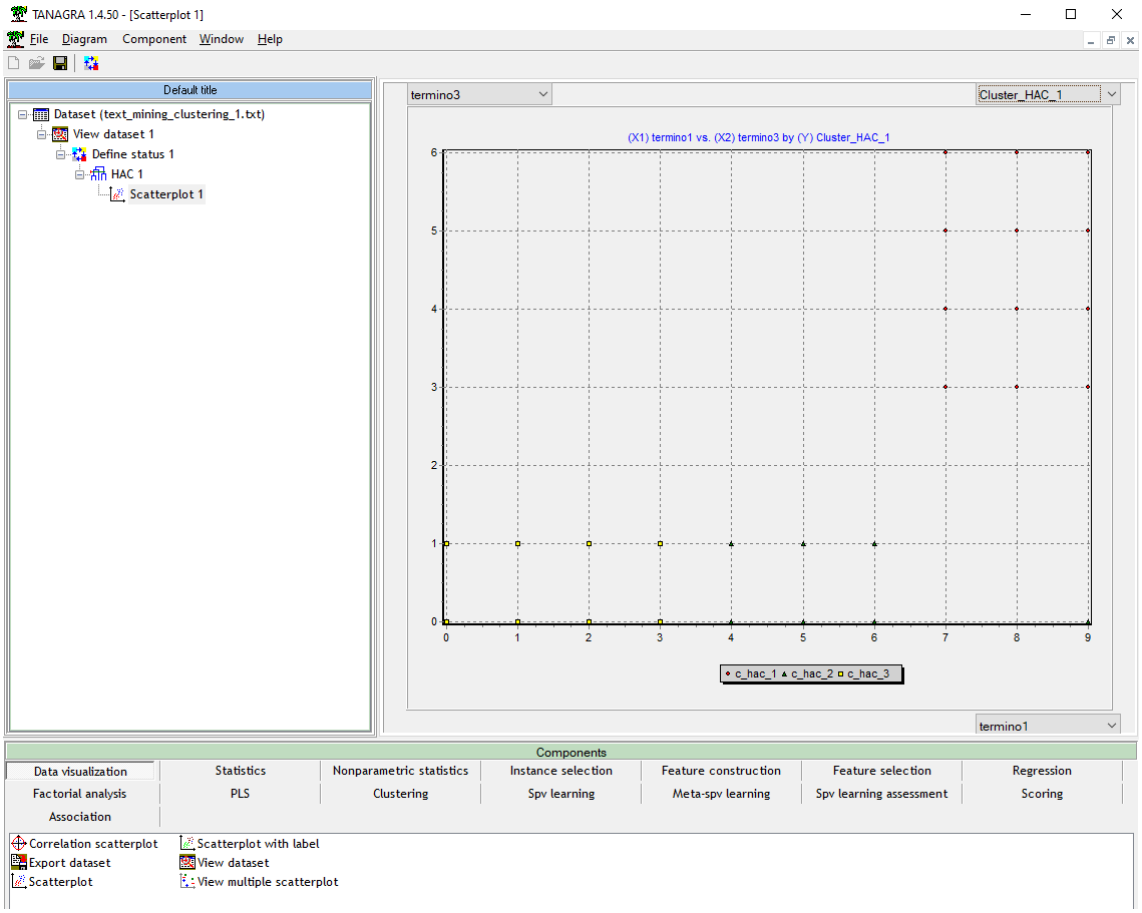


Si queremos saber la cantidad de clusters que se han formado, nos vamos a la pestaña Report de nuevo. Aquí observamos que se han formado 3 clusters, y que los números de los clusters 1, 2 y 3 son 50, 21 y 29 respectivamente.



Procedemos ahora a buscar en la pestaña Data Visualization una herramienta para visualizar un diagrama de dispersión de los clusters. Usaremos Scatterplot. La seleccionamos y arrastramos hasta colgarla debajo de HAC.

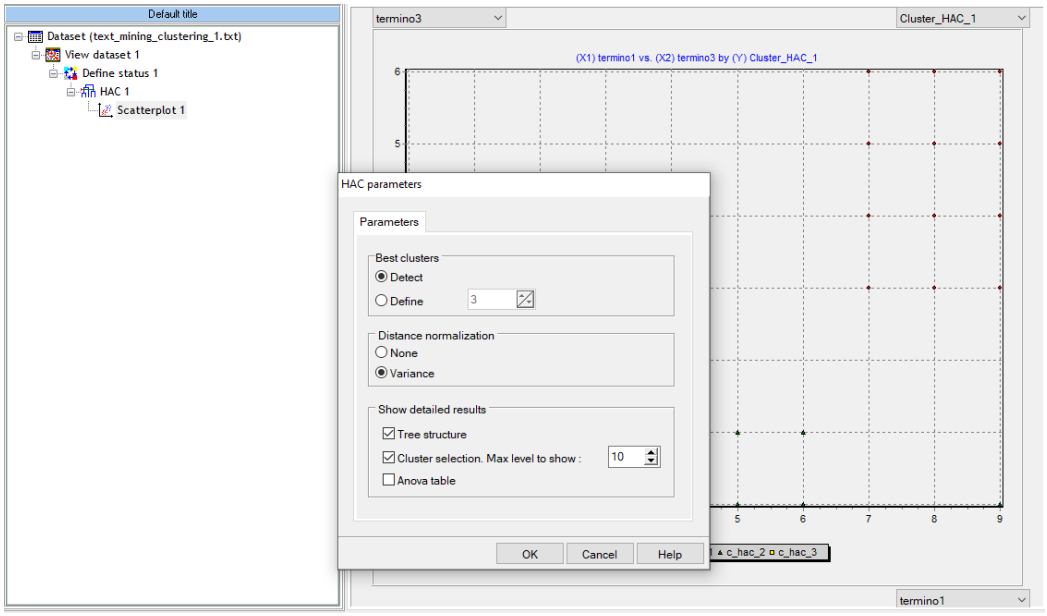
En el eje de las abcisas pondremos término1, en el de las ordenadas término3.



Tenemos también que seleccionar en la pestaña de arriba a la derecha Cluster_HAC_1, que nos permitirá ver como se identifican los datos.

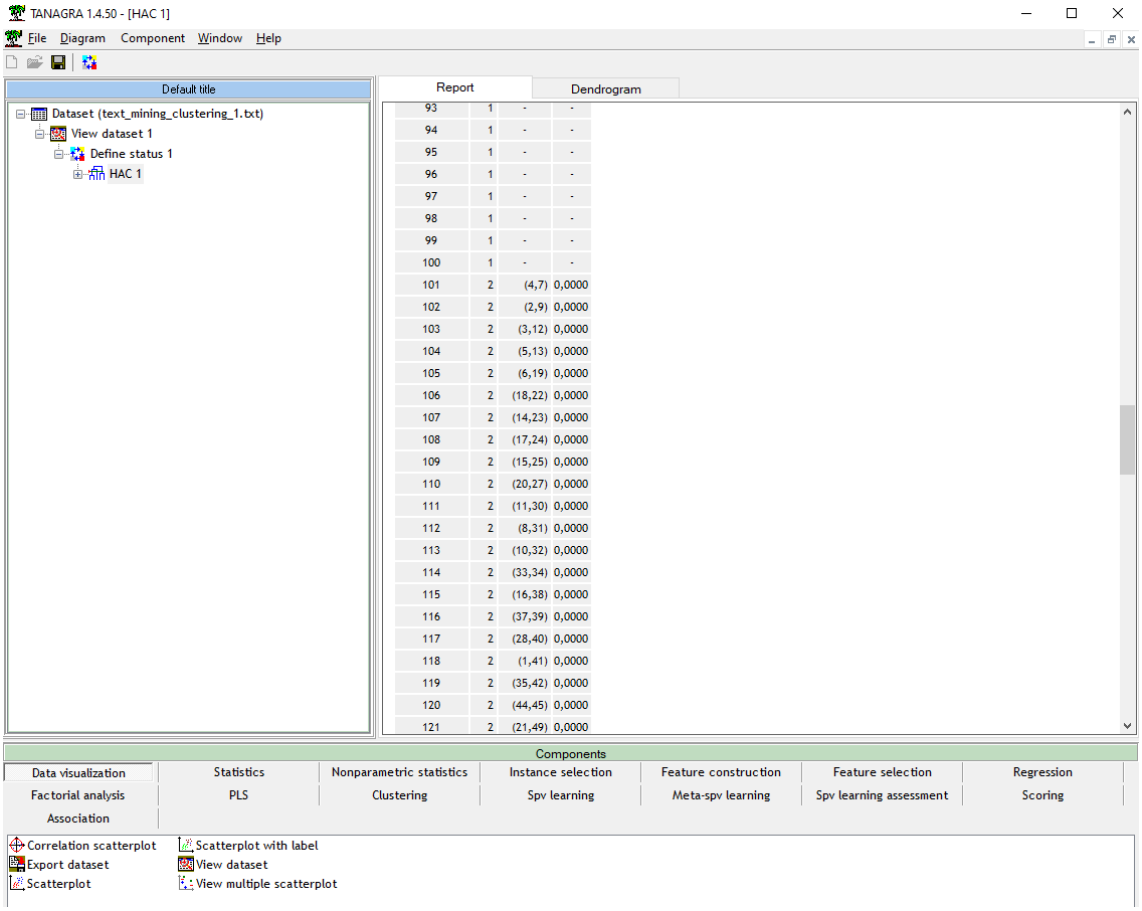
Los clusters se han organizado de la siguiente manera en el diagrama de dispersión: cluster 1, 2 y 3 se representan con un cículo, un triángulo y un cuadrado respectivamente.

Ahora vamos a cambiar la configuración del HAC 1. Para ello hacemos click derecho sobre él y seleccionamos Parameters. Marcamos la casilla que reza Tree Structure y pulsamos Ok:



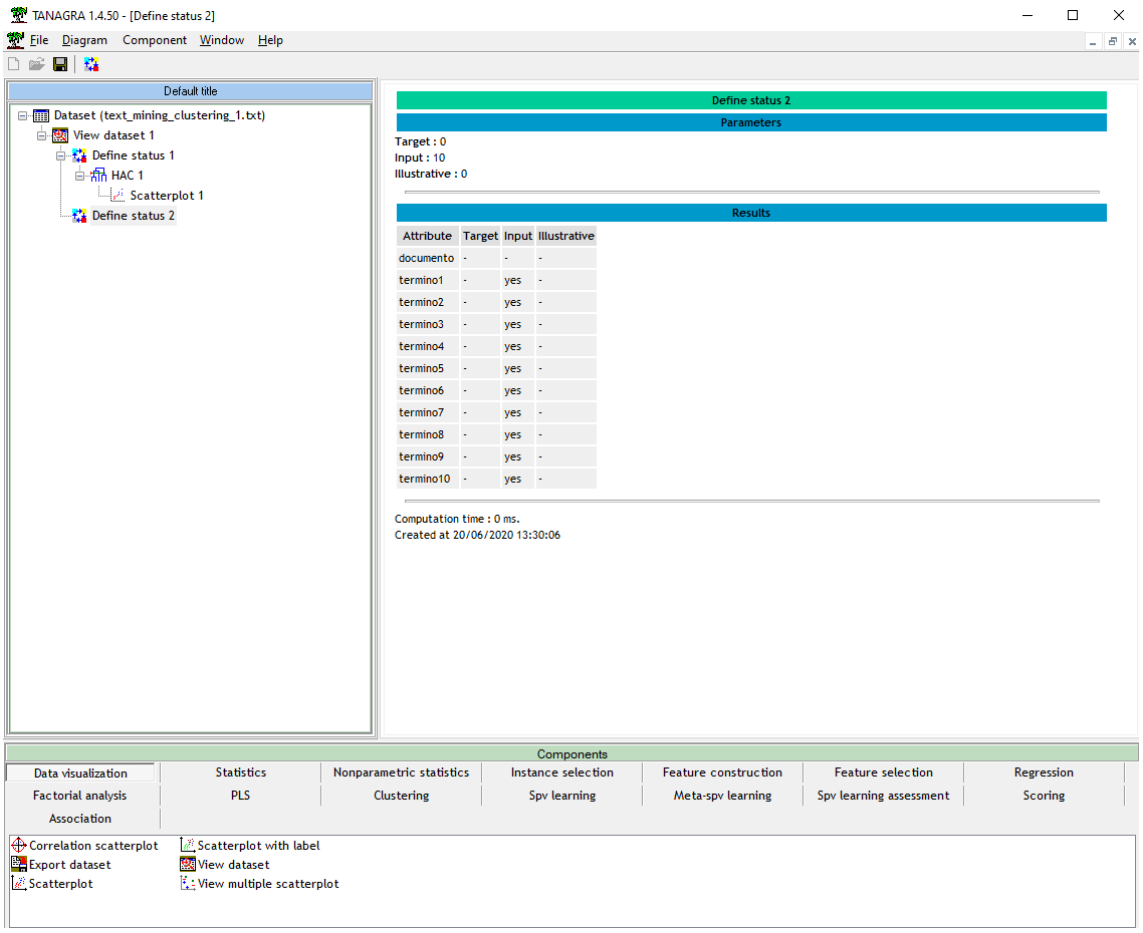
Ejecutamos HAC y observamos que se crea una estructura de árbol de muchas nodo. Los primero 100 nodos significa que se han separado cada uno de los datos que teníamos y están todos separados los unos de los otros (estamos haciendo clustering con un total de 100 documentos).

A partir del 100, se empiezan a agrupar de a dos. Vemos por ejemplo que el nodo 101 se ha creado con los datos 4 y 7, ya que son similares entre ellos, como podemos observar en la distancia que los sepas (0,00).



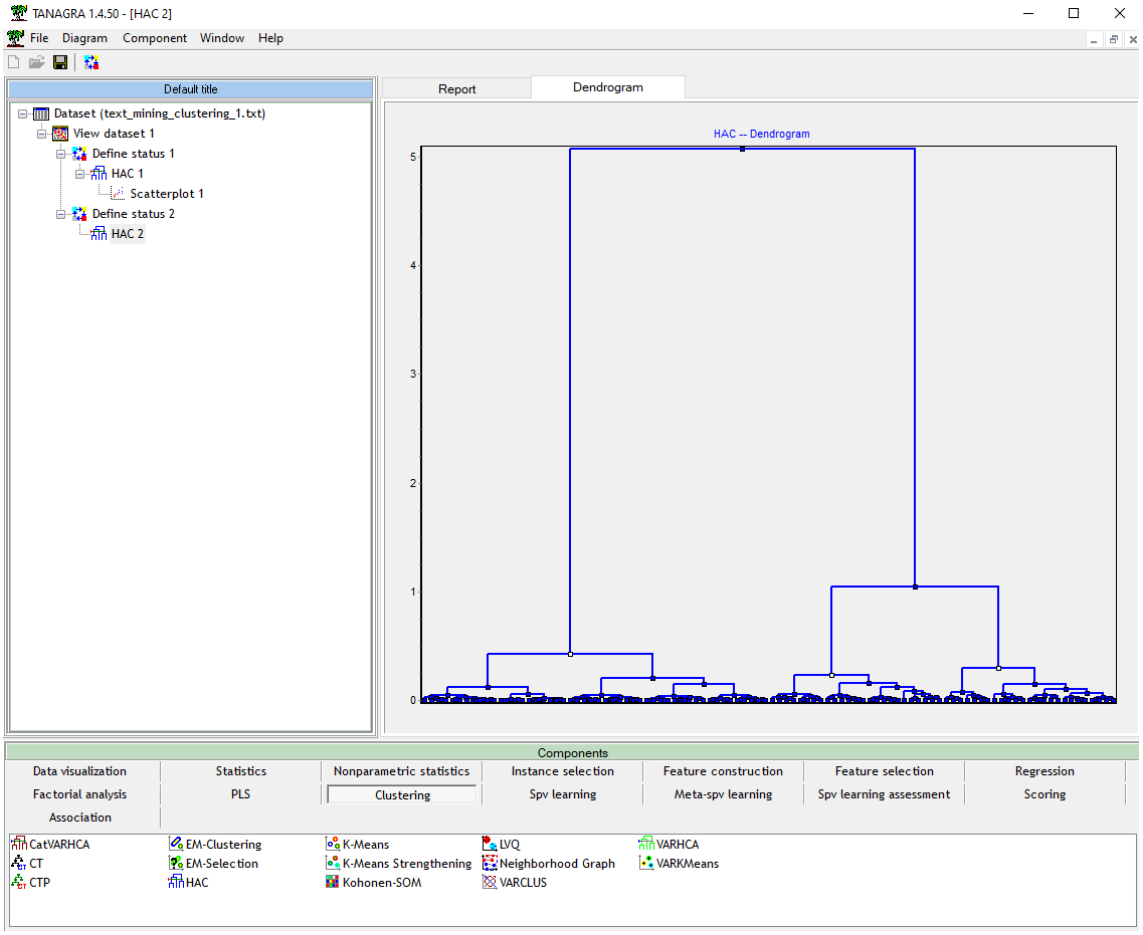
A partir del nodo 174 vemos que ya los datos que se han agrupado, se han agrupado porque se parecen bastante, pero no son exactamente iguales como en los casos anteriores. Esto se puede ver en que la distancia entre ellos es de 0,0015. También podíamos observar esto en el dendograma.

Vamos a crear ahora un nuevo Define Status el que vamos a usar todos los términos como entradas para generar clusters.



Seleccionamos la herramienta HAC de clustering y la colgamos debajo de nuestro nuevo Define Status.

Con los parámetros por defecto ejecutamos el algoritmo. Haciendo click en la pestaña Dendogram, se nos abrirá el dendograma, en el que podremos ver como se han ido organizando los datos para formar clusters.



Para ver cuantos clusters se han generado y el número de elementos que hay en ellos, seleccionamos la pestaña Report. Aquí observamos que se han creado 3 clusters, y que en cluster 1, 2 y 3 hay 50, 25 y 25 elementos respectivamente.

HAC 2

Parameters

# clusters	
Detection	Automatic
Data transformation	
Transformation	None
Visualization	
Index selection	1
Tree structure	0
Anova per variable	0

Results

Clustering results

Clusters	From the dendrogram	After one-pass relocation
cluster n°1	50	50
cluster n°2	25	25
cluster n°3	25	25

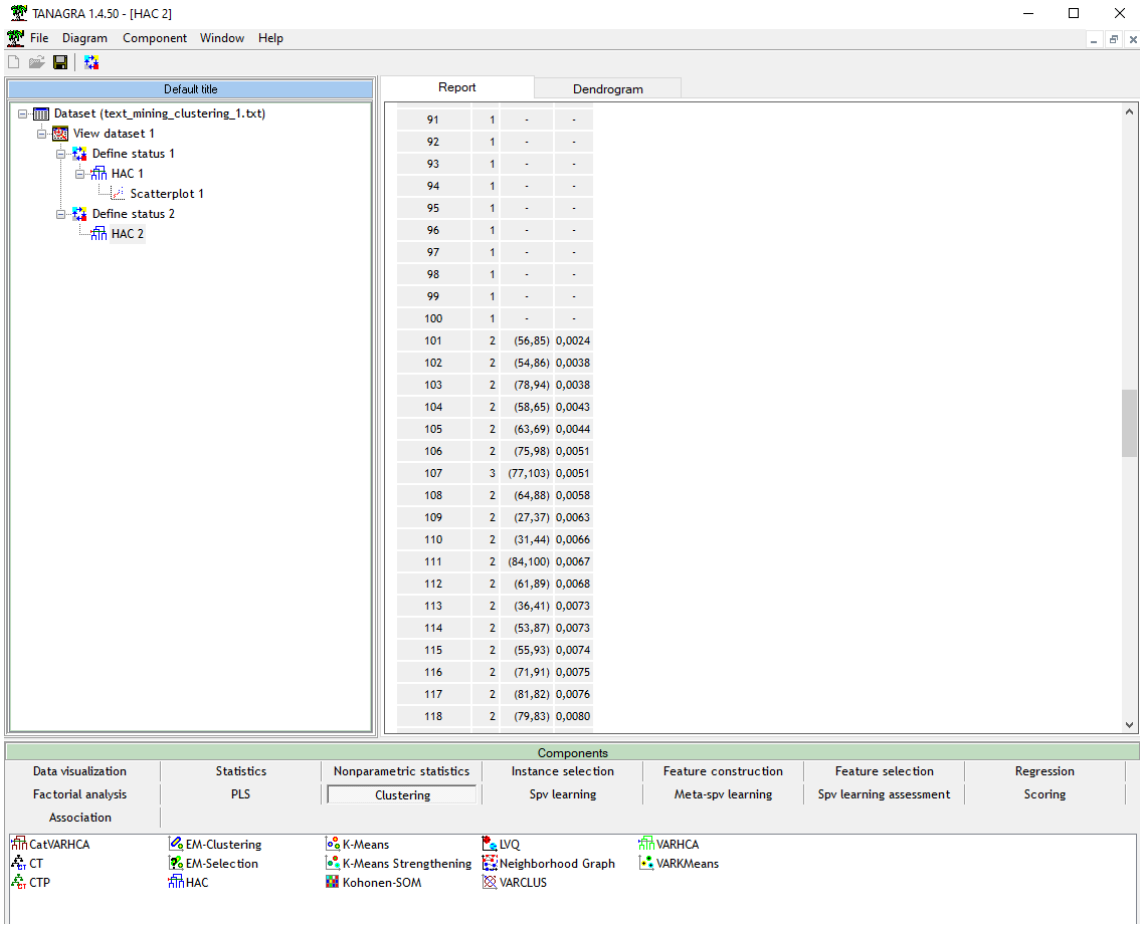
Best cluster selection

Clusters	BSS ratio	Gap
1	0,0000	0,0000
2	0,5075	4,0339
3	0,6117	0,6197
4	0,6538	0,1301
5	0,6830	0,0609
6	0,7061	0,0258

Debido a que tenemos un total de 10 términos, no es posible representar con un diagrama de dispersión los resultados, ya que necesitaríamos 10 dimensiones para poder hacerlo.

Vamos a hacer click con el botón derecho en el algoritmo HAC y vamos a seleccionar Parameters. Aquí seleccionaremos Tree Structure y volveremos a ejecutar el algoritmo.

Como en el caso anterior, los 100 primeros nodos del árbol se han creado con los elementos por separado y solos. A partir del nodo 101 ya empieza el algoritmo a relacionar los datos entre sí. Observamos por ejemplo que el nodo 101 se ha creado con los datos 56 y 85, y que hay una distancia entre ellos de 0,0024.


TANAGRA 1.4.50 - [HAC 2]
File Diagram Component Window Help
Default title
Dataset (text_mining_clustering_1.txt)
View dataset 1
Define status 1
HAC 1
Scatterplot 1
Define status 2
HAC 2
Report
Dendrogram
91 1 - -
92 1 - -
93 1 - -
94 1 - -
95 1 - -
96 1 - -
97 1 - -
98 1 - -
99 1 - -
100 1 - -
101 2 (56,85) 0,0024
102 2 (54,86) 0,0038
103 2 (78,94) 0,0038
104 2 (58,65) 0,0043
105 2 (63,69) 0,0044
106 2 (75,98) 0,0051
107 3 (77,103) 0,0051
108 2 (64,88) 0,0058
109 2 (27,37) 0,0063
110 2 (31,44) 0,0066
111 2 (84,100) 0,0067
112 2 (61,89) 0,0068
113 2 (36,41) 0,0073
114 2 (53,87) 0,0073
115 2 (55,93) 0,0074
116 2 (71,91) 0,0075
117 2 (81,82) 0,0076
118 2 (79,83) 0,0080
Components
Data visualization
Statistics
Nonparametric statistics
Instance selection
Feature construction
Feature selection
Regression
Factorial analysis
PLS
Clustering
Spv learning
Meta-spv learning
Spv learning assessment
Scoring
Association
CatVARHCA
EM-Clustering
K-Means
LVQ
VARHCA
CT
EM-Selection
K-Means Strengthening
Neighborhood Graph
VARKMeans
CTP
HAC
Kohonen-SOM
VARCLUS