Memoria Trabajo Práctico 2 (TP2)
Clustering

***Alumno**: Francisco Javier Piqueras Martínez*
***Asignatura**: Minería de Textos*
***Fecha de entrega**: 2 de febrero de 2021*

Índice

1. Descripción del documento

Este documento consiste en la memoria de la entrega TP2 (Trabajo Práctico 2) de la asignatura de Minería de Textos del Máster en Ingeniería y Ciencia de Datos de la UNED.

Este Trabajo Práctico pertenece los temas 3, 4 y 5 de la asignatura. Además, está compuesto por dos ejercicios, de los cuales la realización del primero es obligatoria y la realización del segundo es opcional. En este caso, se ha realizado solamente el primer ejercicio.

El contenido de esta memoria es el siguiente: En primer lugar, se va a describir en qué consiste la tarea a realizar, en segundo lugar se implementará la memoria con los pasos indicados en el enunciado del ejercicio.

2. Descripción de la tarea

La tarea consiste en hacer uso de un software de clustering, en concreto "*CLUTO – Software for Clustering High Dimensional Datasets"* en su forma "Stand Alone". Guía 3 del manual de uso.

Para la realización de esta tarea es necesario leerse el Manual de uso.

Como datos de prueba se usarán los accesibles en la página oficial de CLUTO: datasets.tar.gz

La realización de la práctica consiste en lo siguiente:

- Descarga el software necesario, manual de uso, conjunto de datos con los que se va a experimentar, descripción y toda la información adicional que se considere necesaria.
- Seleccionar la colección "`re0`", que consiste en noticias de la agencia de noticias Reuters. Es un subconjunto de la colección "Reuters-21578", tiene 1504 documentos, 2886 términos o rasgos y 13 clases o clústeres.
- Utilizar el programa `vcluster` que usa como representación de los objetos el modelo espacio-vectorial. La entrada de este programa es una matriz que corresponde al fichero con extensión "`.mat`" de los que aparecen en datasets.tar.gz, en concreto en este caso "`re0.mat`". Mediante parámetros se le puede indicar que haga el clustering con diferentes algoritmos, funciones, criterio, etc.
- Realizar el clustering y hacer el análisis de los resultados del clustering de la citada colección en 13 clústeres con:
    - Un algoritmo de partición y un algoritmo aglomerativo.
    - Dos funciones de similitud.
    - Dos funciones de criterio.
    - Habrá que combinar estos 6 elementos entre ellos en la medida de lo posible.

Tras la realización de la tarea, preparar una memoria (este documento) en la que:

- Se describa la colección seleccionada.
- Se indique el método de representación de la colección en cuestión.
- Se indiquen los parámetros de `vcluster` que se han utilizado indicando el tipo de algoritmo, funciones de similitud y criterio, así como cualquier otra variación de los parámetros por defecto que se haya utilizado.
- Se añada la información de la salida de las ejecuciones `vcluster`. La calidad del clustering se mide con las medidas internas y externas. Para obtener información de la calidad estadística del clustering comparando la solución del algoritmo con la solución manual (medida externa) se debe utilizar el parámetro `-rclassfile` en la ejecución de `vcluster` junto con el fichero correspondiente a la colección con extensión "`.rclass`" de datasets.tar.gz.
- Se analicen los resultados en términos de calidad de las medidas internas y externas con los diferentes algoritmos, funciones de similitud y criterio, así como el tiempo de ejecución.

3. Memoria

3.1. Descripción de la colección seleccionada

La colección seleccionada recibe el nombre de "re0". Esta colección es un subconjunto de la colección Reuters-21578, la cual tiene 1504 documentos, 2886 términos o rasgos y 13 clases o clústeres.

Pertenece a la colección de conjuntos de datos de prueba del software de CLUTO, que está compuesta por un total de 15 conjuntos de datos:

Data	Source	# of documents	# of terms	# of classes
classic	CACM/CISI/CRANFIELD/MEDLINE	7089	12009	4
fbis	FBIS (TREC)	2463	12674	17
hitech	San Jose Mercury (TREC)	2301	13170	6
reviews	San Jose Mercury (TREC)	4069	23220	5
sports	San Jose Mercury (TREC)	8580	18324	7
la12	LA Times (TREC)	6279	21604	6
new3	TREC	9558	36306	44
tr31	TREC	927	10128	7
tr41	TREC	878	7454	10
ohscal	OHSUMED-233445	11162	11465	10
re0	Reuters-21578	1504	2886	13
re1	Reuters-21578	1657	3758	25
k1a	WebACE	2340	13879	20
k1b	WebACE	2340	13879	6
wap	WebACE	1560	8460	20

Para todos y cada uno de ellos, se ha utilizado una "*stop-list*" para eliminar los términos comunes haciendo uso del algoritmo de "*Porter suffix-stripping*". Además, se han eliminado los términos que aparecen en menos de dos documentos.

3.2. Método de Representación

Para la representación de la colección se ha usado el modelo espacio vectorial (VSM: Vector Space Model), en la que cada documento queda representado como una combinación lineal de vectores, donde cada coeficiente representará la relevancia de cada rasgo en el contenido del documento calculada con una función de pesado.

$$\vec{d}_j = (t_{1j}, t_{2j}, \dots, t_{nj})$$

Así, definir una representación dentro del VSM se reduce a encontrar una función de asignación de pesos: $F: f(t_i) = t_{ij}$ para calcular el valor de cada componente del vector d_j.

La función de pesado usada en esta colección de documentos es la función global TF-IDF (Frecuencia del término por Frecuencia inversa del Documento). Un valor TF-IDF alto se alcanza con una elevada frecuencia del término en el documento dado y una pequeña frecuencia de ocurrencia del término en la colección completa de documentos. Su fórmula es:

$$F: TF - IDF(\vec{t}_i, \vec{d}_j) = f_{ij} \cdot \log(\frac{N}{df(\vec{t}_i)})$$

Siendo f_{ij}, la frecuencia del rasgo "i-ésimo" en el documento "j-ésimo".

En conclusión, cuando un rasgo aparece en muchos documentos el valor TF-IDF se acerca acero, cuando el término es exclusivo, este crece.

Además, para esta representación, todos los valores se han normalizado, por lo que estos tendrán una longitud unitaria.

3.3. Parámetros `vcluster`

Los parámetros que se han utilizado han sido los siguientes:

- Algoritmo (**`-clmethod`**):
  - **`rb`**
    - Este es el algoritmo de partición clásico, que parte del conjunto entero y va haciendo cada vez la partición más perfecta posible, dividiendo cada conjunto en dos, hasta que finalmente obtiene tantos clústeres como instancias (rasgos). En este caso, el algoritmo dejará de particionar cuando haya llegado al número de clústeres que se le ha pasado como parámetro.
  - **`agglo`**
    - La idea de este algoritmo es completamente la contraria al del algoritmo anterior (`rb`). En este se parte desde todas las instancias (rasgos) que se van uniendo, formando clústeres más

grandes en cuanto a su similitud hasta llegar a un único clúster. En este caso, el algoritmo dejará de unir clústeres cuando haya llegado al número indicado en el parámetro de ejecución.

- Funciones de similitud (`-sim`):
  - `cos`
    - La similitud entre objetos se obtiene mediante la función coseno.
  - `corr`
    - La similitud entre objetos se calcula usando el coeficiente de correlación.
- Funciones de criterio (`-rfun`):
  - `i2`
  - `h2`

Criterion Function	Optimazition Function
$\mathcal{I}_1$	$\text{maximize} \sum_{i=1}^{k} \frac{1}{n_i} \left(\sum_{v,u \in S_i} \text{sim}(v,u) \right)$ (1)
$\mathcal{I}_2$	$\text{maximize} \sum_{i=1}^{k} \sqrt{\sum_{v,u \in S_i} \text{sim}(v,u)}$ (2)
$\mathcal{E}_1$	$\text{minimize} \sum_{i=1}^{k} n_i \frac{\sum_{v \in S_i, u \in S} \text{sim}(v,u)}{\sqrt{\sum_{v,u \in S_i} \text{sim}(v,u)}}$ (3)
$\mathcal{G}_1$	$\text{minimize} \sum_{i=1}^{k} \frac{\sum_{v \in S_i, u \in S} \text{sim}(v,u)}{\sum_{v,u \in S_i} \text{sim}(v,u)}$ (4)
$\mathcal{G}'_1$	$\text{minimize} \sum_{i=1}^{k} n_i^2 \frac{\sum_{v \in S_i, u \in S} \text{sim}(v,u)}{\sum_{v,u \in S_i} \text{sim}(v,u)}$ (5)
$\mathcal{H}_1$	$\text{maximize} \frac{\mathcal{I}_1}{\mathcal{E}_1}$ (6)
$\mathcal{H}_2$	$\text{maximize} \frac{\mathcal{I}_2}{\mathcal{E}_1}$ (7)

Se han probado las seis combinaciones posibles.

3.4. Ejecución y salida

Se van a probar las 6 combinaciones posibles. En las ejecuciones, se añadirá el parámetro `–rclassfile` con la ruta del archivo "`.rclass`" para obtener la medida externa de ejecución del algoritmo de agrupamiento.

Se ha optado por elegir el número de clústeres a 10 ya que es un valor estándar que además no coincide con el real y nos va a permitir poder sacar valor de las diferentes medidas que nos ofrece la salida de `vcluster`.

Ejecución 1:

- `-clmethod=rb`
- `-sim=cos`

- `-rfun=i2`

La salida es la siguiente:

```
C:\Users\fjpiqueras\Desktop\CLUTO\cluto-2.1.2\MSWIN-x86>vcluster.exe -
clmethod=rb -sim=cos -crfun=i2 -rclassfile=../../datasets/re0.mat.rclass
../../datasets/re0.mat 10
```

```
*******************************************************************************
vcluster (CLUTO 2.1.2) Copyright 2001-06, Regents of the University of Minnesota

Matrix Information ------------------------------------------------------------
  Name: ../../datasets/re0.mat, #Rows: 1504, #Columns: 2886, #NonZeros: 77808

Options -----------------------------------------------------------------------
  CLMethod=RB, CRfun=I2, SimFun=Cosine, #Clusters: 10
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CSType=Best, AggloFrom=0, AggloCRFun=I2, NTrials=10, NIter=10

Solution ----------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------------------------------
10-way clustering: [I2=5.68e+002] [1504 of 1504], Entropy: 0.410, Purity: 0.622
-------------------------------------------------------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | hous mone trad rese  cpi inte  gnp reta  ipi jobs  lei  bop  wpi
-------------------------------------------------------------------------------------------------------------------------------
  0    66 +0.457 +0.132 +0.034 +0.008 0.248 0.667 |    0   44    0    0    0   22    0    0    0    0    0    0    0
  1   126 +0.410 +0.141 +0.033 +0.008 0.222 0.857 |    1  108    6    0    0    9    0    0    0    0    0    2    0
  2   118 +0.174 +0.052 +0.045 +0.013 0.564 0.398 |    1   15   47   18    0    0    3    0    1    0    0   33    0
  3    65 +0.157 +0.047 +0.028 +0.011 0.368 0.554 |    0   36   24    0    2    3    0    0    0    0    0    0    0
  4   134 +0.160 +0.052 +0.041 +0.013 0.330 0.724 |    0   97    0   12    0   21    0    0    0    4    0    0    0
  5   231 +0.125 +0.040 +0.032 +0.011 0.865 0.216 |   13   35    1    2   50    9   23   18   29   26   10    0   15
  6   173 +0.120 +0.048 +0.031 +0.013 0.118 0.931 |    0  161    2    1    0    9    0    0    0    0    0    0    0
  7   190 +0.093 +0.030 +0.032 +0.016 0.277 0.705 |    0   52    1    1    1  134    1    0    0    0    0    0    0
  8   234 +0.083 +0.024 +0.027 +0.010 0.186 0.876 |    0   22  205    4    0    0    0    0    1    1    0    1    0
  9   167 +0.079 +0.022 +0.038 +0.012 0.727 0.317 |    1   38   33    4    7   12   53    2    6    8    1    2    0
-------------------------------------------------------------------------------------------------------------------------------

Timing Information ------------------------------------------------------------
   I/O:                          0.053 sec
   Clustering:                   0.138 sec
   Reporting:                    0.083 sec
Memory Usage Information ------------------------------------------------------
   Maximum memory used:       2293760 bytes
   Current memory used:        846720 bytes
*******************************************************************************
```

Ejecución 2:

- `-clmethod=rb`
- `-sim=cos`
- `-rfun=h2`

La salida es la siguiente:

```
C:\Users\fjpiqueras\Desktop\CLUTO\cluto-2.1.2\MSWIN-x86>vcluster.exe -
clmethod=rb -sim=cos -crfun=h2 -rclassfile=../../datasets/re0.mat.rclass
../../datasets/re0.mat 10
```

```
*******************************************************************************
vcluster (CLUTO 2.1.2) Copyright 2001-06, Regents of the University of Minnesota

Matrix Information ------------------------------------------------------------
  Name: ../../datasets/re0.mat, #Rows: 1504, #Columns: 2886, #NonZeros: 77808

Options -----------------------------------------------------------------------
  CLMethod=RB, CRfun=H2, SimFun=Cosine, #Clusters: 10
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CSType=Best, AggloFrom=0, AggloCRFun=H2, NTrials=10, NIter=10

Solution ----------------------------------------------------------------------

------------------------------------------------------------------------------------------------------------------------------------------------
10-way clustering: [H2=2.02e-003] [1504 of 1504], Entropy: 0.413, Purity: 0.610
------------------------------------------------------------------------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | hous mone trad rese  cpi inte  gnp reta  ipi jobs  lei  bop  wpi
------------------------------------------------------------------------------------------------------------------------------------------------
  0   125 +0.410 +0.144 +0.033 +0.007 0.256 0.832 |    1  104    6    0    0   11    1    0    0    0    0    2    0
  1    78 +0.379 +0.134 +0.037 +0.013 0.269 0.667 |    0   52    0    1    0   25    0    0    0    0    0    0    0
  2    85 +0.160 +0.051 +0.025 +0.010 0.000 1.000 |    0    0    0    0    0   85    0    0    0    0    0    0    0
  3   116 +0.170 +0.057 +0.041 +0.011 0.277 0.767 |    0   89    1    7    0   19    0    0    0    0    0    0    0
  4   143 +0.148 +0.050 +0.043 +0.014 0.629 0.343 |    1   21   49   29    0    3    5    1    1    0    0   33    0
  5   244 +0.122 +0.038 +0.031 +0.011 0.845 0.209 |   14   36    1    0   51    6   32   18   31   30   10    0   15
  6   199 +0.108 +0.044 +0.032 +0.014 0.216 0.879 |    0  175    6    0    1    9    3    0    1    4    0    0    0
  7   219 +0.087 +0.025 +0.027 +0.010 0.161 0.895 |    0   18  196    3    0    0    0    0    0    1    0    1    0
  8   121 +0.093 +0.030 +0.040 +0.015 0.410 0.471 |    0   57    0    0    3   48   13    0    0    0    0    0    0
  9   174 +0.073 +0.024 +0.033 +0.013 0.642 0.345 |    0   56   60    2    5   13   26    1    4    4    1    2    0
------------------------------------------------------------------------------------------------------------------------------------------------

Timing Information ------------------------------------------------------------
   I/O:                                 0.062 sec
   Clustering:                          0.152 sec
   Reporting:                           0.017 sec
Memory Usage Information ------------------------------------------------------
   Maximum memory used:              2162688 bytes
   Current memory used:               846720 bytes
*******************************************************************************
```

Ejecución 3:

- `-clmethod=rb`
- `-sim=corr`
- `-rfun=h2`

La salida es la siguiente:

```
C:\Users\fjpiqueras\Desktop\CLUTO\cluto-2.1.2\MSWIN-x86>vcluster.exe -
clmethod=rb -sim=corr -crfun=h2 -rclassfile=../../datasets/re0.mat.rclass
../../datasets/re0.mat 10
```

```
*******************************************************************************
vcluster (CLUTO 2.1.2) Copyright 2001-06, Regents of the University of Minnesota

Matrix Information ------------------------------------------------------------
  Name: ../../datasets/re0.mat, #Rows: 1504, #Columns: 2886, #NonZeros: 77808

Options -----------------------------------------------------------------------
  CLMethod=RB, CRfun=H2, SimFun=CorrCoef, #Clusters: 10
  RowModel=None, ColModel=None, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CSType=Best, AggloFrom=0, AggloCRFun=H2, NTrials=10, NIter=10

Solution ----------------------------------------------------------------------

------------------------------------------------------------------------------------------------------------------------------------------------
10-way clustering: [H2=1.76e-003] [1504 of 1504], Entropy: 0.427, Purity: 0.613
------------------------------------------------------------------------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | hous mone trad rese  cpi inte  gnp reta  ipi jobs  lei  bop  wpi
------------------------------------------------------------------------------------------------------------------------------------------------
  0   154 +0.428 +0.153 +0.094 +0.029 0.292 0.805 |    0  124   15    0    0    4    2    0    0    2    0    7    0
  1    85 +0.368 +0.093 +0.081 +0.033 0.254 0.800 |    0   68    1    0    0   12    0    0    0    4    0    0    0
  2   181 +0.352 +0.077 +0.096 +0.026 0.599 0.298 |    2   54   53   35    0    7    1    0    0    0    0   29    0
  3   322 +0.315 +0.087 +0.087 +0.029 0.870 0.199 |   12   51    9    2   56   16   64   20   35   30   11    1   15
  4    94 +0.279 +0.069 +0.068 +0.025 0.115 0.926 |    0    6   87    0    0    0    0    0    0    1    0    0    0
  5   215 +0.320 +0.078 +0.111 +0.032 0.320 0.693 |    2   56    2    0    1  149    3    0    0    2    0    0    0
  6   123 +0.282 +0.093 +0.081 +0.035 0.177 0.886 |    0  109    5    1    0    8    0    0    0    0    0    0    0
  7   156 +0.195 +0.054 +0.056 +0.028 0.177 0.891 |    0   12  139    0    1    1    2    0    1    0    0    0    0
  8   113 +0.185 +0.060 +0.072 +0.031 0.213 0.876 |    0   99    4    1    2    6    1    0    0    0    0    0    0
  9    61 +0.190 +0.106 +0.090 +0.035 0.551 0.475 |    0   29    4    3    0   16    7    0    1    0    0    1    0
------------------------------------------------------------------------------------------------------------------------------------------------

Timing Information ------------------------------------------------------------
   I/O:                                 0.068 sec
   Clustering:                          5.541 sec
   Reporting:                           0.109 sec
Memory Usage Information ------------------------------------------------------
   Maximum memory used:              56688640 bytes
   Current memory used:                846680 bytes
*******************************************************************************
```

Ejecución 4:

- `-clmethod=agglo`
- `-sim=cos`
- `-rfun=i2`

La salida es la siguiente:

```
C:\Users\fjpiqueras\Desktop\CLUTO\cluto-2.1.2\MSWIN-x86>vcluster.exe -
clmethod=agglo -sim=cos -crfun=i2 -rclassfile=../../datasets/re0.mat.rclass
../../datasets/re0.mat 10
```

```
*******************************************************************************
vcluster (CLUTO 2.1.2) Copyright 2001-06, Regents of the University of Minnesota

Matrix Information ------------------------------------------------------------
  Name: ../../datasets/re0.mat, #Rows: 1504, #Columns: 2886, #NonZeros: 77808

Options -----------------------------------------------------------------------
  CLMethod=AGGLO, CRfun=I2, SimFun=Cosine, #Clusters: 10
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CSType=Best, AggloFrom=0, AggloCRFun=I2, NTrials=10, NIter=10

Solution ----------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------------------------
10-way clustering: [I2=5.47e+002] [1504 of 1504], Entropy: 0.481, Purity: 0.548
-------------------------------------------------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | hous mone trad rese  cpi inte  gnp reta  ipi jobs  lei  bop  wpi
-------------------------------------------------------------------------------------------------------------------------
  0   104 +0.512 +0.115 +0.034 +0.005 0.037 0.981 |    0  102    0    0    0    2    0    0    0    0    0    0    0
  1   255 +0.060 +0.021 +0.033 +0.013 0.532 0.447 |    0  114   87    3    3   28   12    1    2    4    1    0    0
  2   317 +0.093 +0.034 +0.034 +0.013 0.908 0.177 |   14   49   19    5   53   15   56   16   28   32    7    9   14
  3   214 +0.081 +0.027 +0.033 +0.015 0.337 0.631 |    0   68    3    1    0  135    6    0    0    0    0    1    0
  4    57 +0.548 +0.095 +0.036 +0.007 0.248 0.667 |    0   38    0    0    0   19    0    0    0    0    0    0    0
  5    96 +0.199 +0.045 +0.034 +0.015 0.091 0.938 |    0   90    0    0    0    6    0    0    0    0    0    0    0
  6    82 +0.194 +0.058 +0.038 +0.014 0.567 0.646 |    1   53    4    3    3    2    0    3    2    2    2    6    1
  7   176 +0.091 +0.029 +0.027 +0.010 0.221 0.858 |    0   17  151    0    1    0    2    0    3    1    1    0    0
  8    74 +0.168 +0.036 +0.043 +0.013 0.568 0.500 |    0   37    7   11    0   10    3    0    0    0    0    6    0
  9   129 +0.136 +0.052 +0.045 +0.015 0.576 0.372 |    1   40   48   19    0    2    1    0    2    0    0   16    0
-------------------------------------------------------------------------------------------------------------------------

Timing Information ------------------------------------------------------------
   I/O:                                0.062 sec
   Clustering:                         0.357 sec
   Reporting:                          0.052 sec
Memory Usage Information ------------------------------------------------------
   Maximum memory used:               23855104 bytes
   Current memory used:                 858760 bytes
*******************************************************************************
```

Ejecución 5:

- `-clmethod=agglo`
- `-sim=cos`
- `-rfun=h2`

La salida es la siguiente:

```
C:\Users\fjpiqueras\Desktop\CLUTO\cluto-2.1.2\MSWIN-x86>vcluster.exe -
clmethod=agglo -sim=cos -crfun=h2 -rclassfile=../../datasets/re0.mat.rclass
../../datasets/re0.mat 10
```

```
*******************************************************************************
vcluster (CLUTO 2.1.2) Copyright 2001-06, Regents of the University of Minnesota

Matrix Information ------------------------------------------------------------
  Name: ../../datasets/re0.mat, #Rows: 1504, #Columns: 2886, #NonZeros: 77808

Options -----------------------------------------------------------------------
  CLMethod=AGGLO, CRfun=H2, SimFun=Cosine, #Clusters: 10
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CSType=Best, AggloFrom=0, AggloCRFun=H2, NTrials=10, NIter=10

Solution ----------------------------------------------------------------------

------------------------------------------------------------------------------------------------------------------------------------------
10-way clustering: [H2=1.86e-003] [1504 of 1504], Entropy: 0.478, Purity: 0.561
------------------------------------------------------------------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | hous mone trad rese  cpi inte  gnp reta  ipi jobs  lei  bop  wpi
------------------------------------------------------------------------------------------------------------------------------------------
  0   192 +0.089 +0.028 +0.027 +0.010 0.233 0.839 |    0   23  161    0    0    1    3    1    0    1    2    0    0
  1   168 +0.077 +0.022 +0.039 +0.015 0.743 0.286 |    0   34   36    2   13    9   48    2    7    9    0    8    0
  2   214 +0.122 +0.043 +0.041 +0.014 0.683 0.439 |    2   94   36   25    3   12    7    1    4    2    2   25    1
  3   100 +0.536 +0.100 +0.035 +0.005 0.038 0.980 |    0   98    0    0    0    2    0    0    0    0    0    0    0
  4   193 +0.108 +0.041 +0.033 +0.014 0.274 0.767 |    0  148   25    0    0   20    0    0    0    0    0    0    0
  5   115 +0.186 +0.120 +0.034 +0.011 0.401 0.583 |    0   67    6    1    1   37    2    0    1    0    0    0    0
  6   181 +0.123 +0.049 +0.034 +0.010 0.834 0.227 |   14   31    4    0   41    2    6   14   22   26    7    0   14
  7   118 +0.079 +0.028 +0.035 +0.012 0.560 0.398 |    0   47   46    8    2    4    5    0    3    1    0    2    0
  8   108 +0.132 +0.040 +0.031 +0.015 0.218 0.824 |    0   16    0    0    0   89    2    0    0    0    0    1    0
  9   115 +0.123 +0.036 +0.041 +0.014 0.519 0.435 |    0   50    5    6    0   43    7    2    0    0    0    2    0
------------------------------------------------------------------------------------------------------------------------------------------

Timing Information ------------------------------------------------------------
   I/O:                                0.058 sec
   Clustering:                        15.363 sec
   Reporting:                          0.096 sec
Memory Usage Information ------------------------------------------------------
   Maximum memory used:            11468800 bytes
   Current memory used:              858720 bytes
*******************************************************************************
```

Ejecución 6:

- `-clmethod=agglo`
- `-sim=corr`
- `-rfun=h2`

La salida es la siguiente:

```
C:\Users\fjpiqueras\Desktop\CLUTO\cluto-2.1.2\MSWIN-x86>vcluster.exe -
clmethod=agglo -sim=corr -crfun=h2 -rclassfile=../../datasets/re0.mat.rclass
../../datasets/re0.mat 10
```

```
*******************************************************************************
vcluster (CLUTO 2.1.2) Copyright 2001-06, Regents of the University of Minnesota

Matrix Information ------------------------------------------------------------
  Name: ../../datasets/re0.mat, #Rows: 1504, #Columns: 2886, #NonZeros: 77808

Options -----------------------------------------------------------------------
  CLMethod=AGGLO, CRfun=H2, SimFun=CorrCoef, #Clusters: 10
  RowModel=None, ColModel=None, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CSType=Best, AggloFrom=0, AggloCRFun=H2, NTrials=10, NIter=10

Solution ----------------------------------------------------------------------

------------------------------------------------------------------------------------------------------------------------------------------
10-way clustering: [H2=1.72e-003] [1504 of 1504], Entropy: 0.407, Purity: 0.636
------------------------------------------------------------------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | hous mone trad rese  cpi inte  gnp reta  ipi jobs  lei  bop  wpi
------------------------------------------------------------------------------------------------------------------------------------------
  0   295 +0.324 +0.087 +0.090 +0.028 0.840 0.217 |   13   40    7    0   55    3   64   20   34   32   11    1   15
  1   121 +0.311 +0.100 +0.090 +0.033 0.431 0.702 |    1   85   13    6    0    4    1    0    0    4    0    7    0
  2   226 +0.182 +0.058 +0.053 +0.029 0.147 0.912 |    1   15  206    0    2    0    0    0    0    1    0    1    0
  3   105 +0.653 +0.094 +0.103 +0.030 0.021 0.990 |    0  104    0    0    0    1    0    0    0    0    0    0    0
  4   184 +0.362 +0.076 +0.117 +0.028 0.234 0.783 |    1   37    0    1    0  144    1    0    0    0    0    0    0
  5   186 +0.267 +0.087 +0.100 +0.029 0.655 0.339 |    0   36   63   35    0   11    9    0    1    2    0   29    0
  6   167 +0.140 +0.054 +0.065 +0.029 0.348 0.743 |    0  124   19    0    3   15    5    0    1    0    0    0    0
  7    59 +0.567 +0.101 +0.089 +0.025 0.245 0.678 |    0   40    0    0    0   19    0    0    0    0    0    0    0
  8    81 +0.306 +0.119 +0.087 +0.035 0.173 0.889 |    0   72    2    0    0    6    0    0    1    0    0    0    0
  9    80 +0.243 +0.050 +0.114 +0.027 0.322 0.688 |    0   55    9    0    0   16    0    0    0    0    0    0    0
------------------------------------------------------------------------------------------------------------------------------------------

Timing Information ------------------------------------------------------------
   I/O:                                0.056 sec
   Clustering:                        22.872 sec
   Reporting:                          0.101 sec
Memory Usage Information ------------------------------------------------------
   Maximum memory used:            79560704 bytes
   Current memory used:              858768 bytes
*******************************************************************************
```

3.5. Análisis de resultados

Dentro del aprendizaje automático, el **clustering** es una técnica de aprendizaje no supervisado que se utiliza para numerosas tareas que pretendan extraer conocimiento del contenido actual.

En ella, se trata de realizar algún tipo de organización lógica de la información para facilitar posteriores análisis. Además, no se conoce de antemano las clases (clústeres, grupos).

La entrada en las ejecuciones de los algoritmos de agrupamiento es:

- Los n objetos que se quieren agrupar (estos están en el archivo "`.mat`").
- El número de grupos (clústeres).

En cuanto a la entrada, la representación de la colección de documentos tiene el siguiente formato:

- Matriz rasgo-documento(término-documento).

La salida:

- Los k clústeres de objetos

En estos casos hemos hecho agrupamiento de tipo ***hard***. Es decir, hemos asignado cada objeto a un único clúster. Además, estos han sido de tipo jerárquico ya que, en los algoritmos utilizados, se ha construido la estructura tanto de arriba abajo (partición) como de abajo a arriba (aglomerativo).

Basándonos en los resultados obtenidos en las ejecuciones del apartado anterior, cabe notar que los resultados obtenidos no son especialmente buenos. El principal motivo se debe a que estamos intentando clasificar en 10 grupos un conjunto de documentos que está separado en 13 grupos. Si intentáramos realizar el agrupamiento en 13 grupos, las medidas de evaluación mejorarían independientemente de los parámetros que utilicemos. Por ejemplo:

```
vcluster.exe -clmethod=agglo -sim=corr -crfun=h2 -
rclassfile=../../datasets/re0.mat.rclass ../../datasets/re0.mat 13
```

```
*******************************************************************************
vcluster (CLUTO 2.1.2) Copyright 2001-06, Regents of the University of Minnesota

Matrix Information ------------------------------------------------------------
  Name: ../../datasets/re0.mat, #Rows: 1504, #Columns: 2886, #NonZeros: 77808

Options -----------------------------------------------------------------------
  CLMethod=AGGLO, CRfun=H2, SimFun=CorrCoef, #Clusters: 13
  RowModel=None, ColModel=None, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CSType=Best, AggloFrom=0, AggloCRFun=H2, NTrials=10, NIter=10

Solution ----------------------------------------------------------------------

------------------------------------------------------------------------------------------------------------------------------------------
13-way clustering: [H2=1.83e-003] [1504 of 1504], Entropy: 0.377, Purity: 0.662
------------------------------------------------------------------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | hous mone trad rese  cpi inte  gnp reta  ipi jobs  lei  bop  wpi
------------------------------------------------------------------------------------------------------------------------------------------
  0    69 +0.207 +0.050 +0.111 +0.033 0.650 0.304 |    0   21   18    7    0   11    9    0    1    2    0    0    0
  1   121 +0.311 +0.100 +0.090 +0.033 0.431 0.702 |    1   85   13    6    0    4    1    0    0    4    0    7    0
  2    80 +0.243 +0.050 +0.114 +0.027 0.322 0.688 |    0   55    9    0    0   16    0    0    0    0    0    0    0
  3   105 +0.653 +0.094 +0.103 +0.030 0.021 0.990 |    0  104    0    0    0    1    0    0    0    0    0    0    0
  4   184 +0.362 +0.076 +0.117 +0.028 0.234 0.783 |    1   37    0    1    0  144    1    0    0    0    0    0    0
  5   166 +0.177 +0.059 +0.056 +0.030 0.174 0.892 |    1   13  148    0    2    0    0    0    0    1    0    1    0
  6   167 +0.140 +0.054 +0.065 +0.029 0.348 0.743 |    0  124   19    0    3   15    5    0    1    0    0    0    0
  7    59 +0.567 +0.101 +0.089 +0.025 0.245 0.678 |    0   40    0    0    0   19    0    0    0    0    0    0    0
  8    81 +0.306 +0.119 +0.087 +0.035 0.173 0.889 |    0   72    2    0    0    6    0    0    1    0    0    0    0
  9    60 +0.345 +0.063 +0.070 +0.021 0.057 0.967 |    0    2   58    0    0    0    0    0    0    0    0    0    0
 10   117 +0.391 +0.079 +0.101 +0.027 0.514 0.385 |    0   15   45   28    0    0    0    0    0    0    0   29    0
 11    74 +0.351 +0.079 +0.130 +0.033 0.519 0.662 |    0    2    6    0    4    2   49    2    4    3    1    1    0
 12   221 +0.356 +0.090 +0.096 +0.026 0.820 0.231 |   13   38    1    0   51    1   15   18   30   29   10    0   15
------------------------------------------------------------------------------------------------------------------------------------------

Timing Information ------------------------------------------------------------
   I/O:                                0.093 sec
   Clustering:                        24.782 sec
   Reporting:                          0.101 sec
Memory Usage Information ------------------------------------------------------
   Maximum memory used:            79560704 bytes
   Current memory used:              858760 bytes
*******************************************************************************
```

A pesar de que la pureza se ha mantenido, la entropía se ha reducido, lo que es un buen indicador.

3.5.1. Medidas internas

Como se puede observar en las diferentes ejecuciones, la medida `ISim` debería ser lo más alta posible mientras que la medida `ESim` debería ser lo más baja. Como se puede comprobar en las distintas ejecuciones, estos indicadores son buenos especialmente en los primeros grupos, pero la calidad del indicador va descendiendo hasta que llegamos a los grupos del final, en los que, por lo general, tienen una similitud bastante baja que se acerca más a la medida de similitud externa (con el resto de objetos). A priori, estos indicadores no son del todo buenos, aunque se podría deber a que ciertos grupos sean menos densos que otros.

Esta información va a ser proporcionada por las medidas externas.

3.5.2. Medidas externas

Las medidas externas nos dan la información comparando la solución del algoritmo con la solución manual.

A primera vista se puede observar claramente que el mayor error de clasificación independientemente de los parámetros se da con el grupo “`mone`”, seguido del grupo “`trad`” en todos los algoritmos. El motivo puede ser que este tema sea más genérico que el resto, y lleve al algoritmo a clasificar de forma errónea este tema en otros más específicos.

Además, los grupos "gnp", "reta", "ipi", "jobs", "lei" se están agrupando todos en el mismo grupo. Podríamos intuir que estos están muy relacionados entre sí o que tienen rasgos muy similares y compartidos.

Por sacar una lista de parámetros que sean los que mejor hayan funcionado en el agrupamiento de esta colección, comparando las ejecuciones que hemos hecho, destaca la Ejecución 6, cuyos parámetros son:

- **-clmethod=agglo**
- **-sim=corr**
- **-rfun=h2**

Sin embargo, como se puede apreciar en los tiempos de ejecución, el agrupamiento con estos parámetros, especialmente gracias al parámetro –rfun=h2 hace que sea mucho más costoso que con –rfun=i2. Aquí puede ser inapreciable, pero con conjuntos de datos mucho más grandes, la diferencia es muy notable, por lo que habría que pensar en si realmente merece la pena la medida de criterio h2 y su mejora en resultados en contra del coste de agrupamiento del algoritmo.

3.6. Prueba adicional

Por probar algún parámetro más disponible en el manual de uso de los que hacen referencia a la visualización del resultado, se ha probado a añadir la opción –showtree, que muestra los pasos que ha dado el algoritmo aglomerativo en su ejecución:

```
vcluster.exe -clmethod=graph -sim=corr -crfun=h2 -showtree -
rclassfile=../../datasets/re0.mat.rclass ../../datasets/re0.mat 13
```

```
*******************************************************************************
vcluster (CLUTO 2.1.2) Copyright 2001-06, Regents of the University of Minnesota

Matrix Information ------------------------------------------------------------
  Name: ../../datasets/re0.mat, #Rows: 1504, #Columns: 2886, #NonZeros: 77808

Options -----------------------------------------------------------------------
  CLMethod=GRAPH, CRfun=Cut, SimFun=CorrCoef, #Clusters: 13
  RowModel=None, ColModel=None, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CSType=Best, AggloFrom=0, AggloCRFun=SLINK_W, NTrials=10, NIter=10

Solution ----------------------------------------------------------------------

------------------------------------------------------------------------------------------------------------------------------------------
13-way clustering: [Cut=8.63e+003] [1490 of 1504], Entropy: 0.374, Purity: 0.654
------------------------------------------------------------------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | hous mone trad rese  cpi inte  gnp reta  ipi jobs  lei  bop  wpi
------------------------------------------------------------------------------------------------------------------------------------------
  0    20 +0.809 +0.107 +0.000 +0.000 0.000 1.000 |    0   20    0    0    0    0    0    0    0    0    0    0    0
  1    54 +0.536 +0.187 +0.000 +0.001 0.000 1.000 |    0   54    0    0    0    0    0    0    0    0    0    0    0
  2    25 +0.367 +0.159 +0.001 +0.001 0.000 1.000 |    0   25    0    0    0    0    0    0    0    0    0    0    0
  3    67 +0.179 +0.099 +0.001 +0.002 0.118 0.925 |    0   62    0    0    0    1    0    0    0    4    0    0    0
  4   104 +0.142 +0.106 +0.001 +0.001 0.369 0.673 |    0   70    7    4    0   22    0    0    0    0    0    1    0
  5    94 +0.089 +0.057 +0.001 +0.001 0.113 0.915 |    0    8   86    0    0    0    0    0    0    0    0    0    0
  6   177 +0.073 +0.044 +0.001 +0.001 0.606 0.322 |    1   43   57   32    0    7    1    0    1    0    0   35    0
  7    82 +0.066 +0.057 +0.001 +0.002 0.683 0.500 |    0    4    6    2    5    2   41    5    4    8    5    0    0
  8   193 +0.063 +0.043 +0.001 +0.001 0.804 0.254 |   14   33    1    0   49    1    8   14   29   23    6    0   15
  9   190 +0.061 +0.041 +0.000 +0.001 0.232 0.779 |    0   39    0    1    0  148    2    0    0    0    0    0    0
 10   121 +0.057 +0.042 +0.001 +0.001 0.093 0.959 |    1    1  116    0    0    0    1    0    1    1    0    0    0
 11   255 +0.035 +0.028 +0.000 +0.001 0.206 0.859 |    0  219   15    0    0   19    2    0    0    0    0    0    0
 12   108 +0.024 +0.014 +0.001 +0.001 0.690 0.250 |    0   26   27    3    5   16   25    0    1    3    0    2    0
------------------------------------------------------------------------------------------------------------------------------------------

------------------------------------------------------------------------------
Hierarchical Tree that optimizes the SLINK_W criterion function...
------------------------------------------------------------------------------
                            hous mone trad rese  cpi inte  gnp reta  ipi jobs  lei  bop  wpi
------------------------------------------------------------------------------------------------------------------------------------------
24
|---22
|    |-------18
|    |       |-----15
|    |       |      |-------6        1   43   57   32    0    7    1    0    1    0    0   35    0
|    |       |      |-------3        0   62    0    0    0    1    0    0    0    4    0    0    0
|    |       |-------------4        0   70    7    4    0   22    0    0    0    0    0    1    0
|    |---20
|        |-19
|        | |---17
|        | |   |-----------11       0  219   15    0    0   19    2    0    0    0    0    0    0
|        | |   |-16
|        | |     |----------12       0   26   27    3    5   16   25    0    1    3    0    2    0
|        | |     |-----13
|        | |           |---10        1    1  116    0    0    0    1    0    1    1    0    0    0
|        | |           |---5         0    8   86    0    0    0    0    0    0    0    0    0    0
|        | |---------14
|        |           |-----8        14   33    1    0   49    1    8   14   29   23    6    0   15
|        |           |-----7         0    4    6    2    5    2   41    5    4    8    5    0    0
|        |-----------------9         0   39    0    1    0  148    2    0    0    0    0    0    0
|-23
   |---21
   |    |-------------------2        0   25    0    0    0    0    0    0    0    0    0    0    0
   |    |-------------------1        0   54    0    0    0    0    0    0    0    0    0    0    0
   |------------------------0        0   20    0    0    0    0    0    0    0    0    0    0    0
------------------------------------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------

Timing Information ------------------------------------------------------------
   I/O:                           0.058 sec
   Clustering:                    8.133 sec
   Reporting:                     0.262 sec
```

3.7. Conclusión

Los resultados no han sido prometedores, y han sido bastante similares independientemente de los parámetros utilizados, destacando una combinación levemente por encima de las otras. La mayor conclusión puede extraerse sobre los datos y sus grupos, cabe destacar que hay un grupo que puede sospechar que es genérico ya que se han agrupado por error sus documentos por el resto de los grupos. Adicionalmente, puede que otros grupos más específicos sean muy similares puesto que todos los algoritmos han agrupado sus documentos en el mismo clúster. Puede que la densidad de estos sea mucho mayor que la densidad del resto de grupos y a su vez estos estén bastante cerca (sean similares entre ellos) a diferencia del resto de grupos, lo que hace que estos se hayan agrupado en un mismo grupo en lugar de en varios.

Además, también se ha podido observar que las funciones de criterio influyen de forma significativa en el tiempo que el algoritmo tarda en agrupar los objetos.