Trabajo Práctico 2

Minería de Textos

Irene J. Ventura Farias

Introducción

Cluto es un software de clustering de documentos, el cual que se encarga de aplicar clustering sobre colecciones de datos usando diferentes tipos de algoritmos y combinaciones de parámetros, para posteriormente analizar sus características y comparar la diferencia entre ellos.

Concretamente, para este trabajo práctico hemos utilizado esta versión del software: CLUTO - Software for Clustering High-Dimensional Datasets en su versión *"Stand Alone"* y nos hemos enfocado exclusivamente en los algoritmos de partición y los aglomerativos. Los cuales nos han permitido analizar y comparar sus resultados en varias ejecuciones, aplicando diferentes combinaciones de parámetros sobre un mismo dataset (*re0*): Noticias de la agencia de noticias Reuters.

Colección de Datos

La colección de datos que vamos a utilizar para nuestros clústers es el *re0*, se encuentra en la colección de datos de prueba que ofrece CLUTO: datasets. De forma más detallada, re0 es un subconjunto del dataset de la agencia de noticias Reuters, la cual contiene noticias publicadas por la propia agencia en 1987.

La colección *re0* contiene 1504 documentos, 2886 términos y 13 clases. La primera versión de estos documentos fue recopilados, indexados y clasificados en categorías por el personal de *Reuters Ltd.* y de *Carnegie Group, Inc.* en el curso del desarrollo del sistema de categorización de textos CONSTRUE en 1987 y para esta versión cada documento en la colección pertenece únicamente a una clase.

Luego en 1990, se hicieron públicos y disponibles estos documentos con el propósito de investigación. En esta práctica hemos trabajado concretamente con la versión de *Reuters-21578,* que fue desarrollada por Steve Finch y David D Lewis en 1996. Tras una limpieza y tratamientos de los datos, cada documento puede pertenecer a una o más categorías que se encuentran agrupadas por tipos [EXCHANGES, ORGS, PEOPLE, PLACES o TOPICS]. Pero en el caso de re0 se han seleccionado documentos que pertenecen únicamente a una categoría, por lo que podrían agruparse en diferentes clústeres, siendo cada clúster la representación de una categoría (housing, money, trade, reserves, cpi, interest, gnp, retail, ipi, Jobs, lei, bop y wpi).

Método de Representación de la Colección de Datos

En la colección de datos re0 se realizó previamente un preprocesamiento de los datos para reducir la lista de rasgos en cada documento:

- Se descartaron palabras comunes utilizando una lista de palabras “vacías” (stop-list)
- Se realizo Stemming sobre los rasgos, aplicando el algoritmo de Porter para reducir las palabras a su raíz.

Una vez terminada la limpieza de los documentos, para la representación del dataset se ha aplicado un Modelo Espacio Vectorial (VSM). Cada documento está representado como un vector, en donde cada valor en el vector representa la relevancia de cada término o rasgo presente en el documento. La relevancia se mide con una función de pesado global, que en este caso particular la función de pesado global es Frecuencia del término por Frecuencia inversa del Documento (TF-IDF):

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

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

Esta función nos permite considerar que si un rasgo que se repiten en diferentes documentos su frecuencia es bajo (se acerca a 0), mientras que si un rasgo es exclusivo en un documento su frecuencia es mayor.

Además, todos los valores se han normalizado a 1, esto nos permite evaluar documentos con diferentes longitudes; de manera que los vectores que representan cada documento tendrán una longitud unitaria.

Parámetros utilizados en el *vcluster*

Según el enunciado de la práctica los requisitos para realizar el clustering y analizar los resultados en 13 clústeres diferentes, deben estar definidos en base a diferentes combinaciones de los siguientes parámetros:

- **Algoritmos**
  - *De Partición* (**`direct`**): este método consiste en encontrar simultáneamente todos los clústers y aunque su tiempo de ejecución sea el más lento. En términos de calidad, se recomienda utilizar este método para valores pequeños de k (10-20) y en nuestro caso necesitamos aplicarlo concretamente para 13 clúster.

  - *Aglomerativo* (**`agglo`**): este método usa el paradigma de aglomeración, cuyo objetivo es optimizar una función de criterio. La solución se obtiene deteniendo el proceso aglomeración cuando se llegan al número de clúster especificados.

- **Funciones de Similitud (-`sim`)**: es la métrica que se utiliza para la agrupación.
  - Función Coseno (**`cos`**)
  - Coeficiente de Correlación (**`corr`**)

- **Funciones de Criterio (-`crfun`):** este parámetro permite asignar la función que se usa para encontrar los clústeres. Según el manual recomienda utilizar los siguientes valores:

  - i2: A medida que la esta función de criterio sea mayor, mejores resultados

$$\text{maximize} \sum_{i=1}^{k} \sqrt{\sum_{v,u \in S_i} \text{sim}(v,u)}$$

  - h2

$$\text{maximize} \frac{\mathcal{I}_2}{\mathcal{E}_1}$$

Resultados de las ejecuciones de vcluster

Finalmente he realizado un total del ocho ejecuciones diferentes con los parámetros seleccionados en el apartado anterior:

```
.\vcluster -clmethod=direct -sim=cos -crfun=i2 -rclassfile='re0.mat.rclass'
.\re0.mat 13
```

```
*******************************************************************************
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

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

Options -----------------------------------------------------------------------
  CLMethod=Direct, CRfun=I2, SimFun=Cosine, #Clusters: 13
  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 ----------------------------------------------------------------------

------------------------------------------------------------------------------------------------------------------------
13-way clustering: [I2=6.04e+002] [1504 of 1504], Entropy: 0.382, Purity: 0.659
------------------------------------------------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | hous mone trad rese  cpi inte  gnp reta  ipi jobs  lei  bop  wpi
------------------------------------------------------------------------------------------------------------------------
  0   112 +0.488 +0.114 +0.034 +0.008 0.114 0.938 |    0  105    4    0    0    2    0    0    0    0    0    1    0
  1    70 +0.428 +0.133 +0.034 +0.012 0.312 0.657 |    0   46    1    0    2   21    0    0    0    0    0    0    0
  2    87 +0.204 +0.069 +0.037 +0.016 0.024 0.989 |    0   86    1    0    0    0    0    0    0    0    0    0    0
  3    86 +0.192 +0.064 +0.035 +0.013 0.554 0.651 |    1   56    3    2    2    3    0    3    1    6    2    6    1
  4    65 +0.197 +0.050 +0.040 +0.016 0.408 0.646 |    0   42    1    5    0   14    2    0    0    0    0    1    0
  5    78 +0.179 +0.051 +0.026 +0.011 0.027 0.987 |    0    0    1    0    0   77    0    0    0    0    0    0    0
  6   166 +0.155 +0.047 +0.033 +0.009 0.803 0.295 |   12   12    1    2   49    1    3   14   28   22    8    0   14
  7   156 +0.158 +0.046 +0.044 +0.011 0.570 0.321 |    1   50   46   28    0    2    1    0    0    0    0   28    0
  8   112 +0.123 +0.033 +0.032 +0.010 0.200 0.848 |    0   14   95    2    0    0    1    0    0    0    0    0    0
  9   142 +0.107 +0.037 +0.032 +0.013 0.296 0.775 |    0  110   19    1    1   10    1    0    0    0    0    0    0
 10   129 +0.105 +0.030 +0.039 +0.013 0.653 0.519 |    1   13   14    1    6    5   67    3    7   10    1    1    0
 11   139 +0.089 +0.027 +0.037 +0.014 0.322 0.547 |    0   59    1    1    0   76    2    0    0    0    0    0    0
 12   162 +0.077 +0.023 +0.027 +0.010 0.287 0.815 |    1   15  132    0    0    8    3    0    1    1    0    1    0
------------------------------------------------------------------------------------------------------------------------

Timing Information ------------------------------------------------------------
   I/O:                                 0.029 sec
   Clustering:                          0.168 sec
   Reporting:                           0.008 sec
*******************************************************************************
```

```
.\vcluster -clmethod=direct -sim=corr -crfun=i2 -rclassfile='re0.mat.rclass'
.\re0.mat 13
```

```
*******************************************************************************
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

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

Options -----------------------------------------------------------------------
  CLMethod=Direct, CRfun=I2, 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=I2, NTrials=10, NIter=10

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

------------------------------------------------------------------------------------------------------------------------
13-way clustering: [I2=8.65e+002] [1504 of 1504], Entropy: 0.370, Purity: 0.666
------------------------------------------------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | hous mone trad rese  cpi inte  gnp reta  ipi jobs  lei  bop  wpi
------------------------------------------------------------------------------------------------------------------------
  0   106 +0.650 +0.095 +0.100 +0.028 0.092 0.953 |    0  101    3    0    0    1    0    0    0    0    0    1    0
  1    61 +0.537 +0.124 +0.084 +0.030 0.323 0.639 |    0   39    1    0    2   19    0    0    0    0    0    0    0
  2   150 +0.405 +0.094 +0.104 +0.026 0.768 0.313 |   11   16    1    0   47    1    1   12   21   22    4    0   14
  3    90 +0.348 +0.098 +0.090 +0.034 0.361 0.744 |    0   67   10    3    0    0    1    0    0    4    0    5    0
  4   166 +0.353 +0.081 +0.096 +0.027 0.605 0.325 |    2   39   54   35    0    3    2    0    0    1    0   30    0
  5   164 +0.368 +0.077 +0.115 +0.033 0.217 0.817 |    1   27    0    0    1  134    1    0    0    0    0    0    0
  6    95 +0.331 +0.095 +0.079 +0.034 0.045 0.979 |    0   93    1    0    0    1    0    0    0    0    0    0    0
  7   120 +0.373 +0.076 +0.121 +0.030 0.753 0.300 |    2   29    1    0    6   15   36    7   13    4    6    0    1
  8    94 +0.280 +0.068 +0.067 +0.025 0.146 0.894 |    0    9   84    0    0    0    1    0    0    0    0    0    0
  9   152 +0.203 +0.054 +0.055 +0.027 0.095 0.934 |    0   10  142    0    0    0    0    0    0    0    0    0    0
 10   101 +0.247 +0.060 +0.112 +0.036 0.381 0.594 |    0   60    5    2    0   32    2    0    0    0    0    0    0
 11   122 +0.188 +0.059 +0.068 +0.031 0.186 0.885 |    0  108    8    1    1    4    0    0    0    0    0    0    0
 12    83 +0.217 +0.058 +0.110 +0.041 0.707 0.434 |    0   10    9    1    3    9   36    1    3    8    1    2    0
------------------------------------------------------------------------------------------------------------------------

Timing Information ------------------------------------------------------------
   I/O:                                 0.026 sec
   Clustering:                          7.861 sec
   Reporting:                           0.050 sec
```

```
.\vcluster -clmethod=direct -sim=cos -crfun=h2 -rclassfile='re0.mat.rclass'
.\re0.mat 13
```

```
*******************************************************************************
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

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

Options -----------------------------------------------------------------------
  CLMethod=Direct, CRfun=H2, SimFun=Cosine, #Clusters: 13
  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 ----------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------------------------
13-way clustering: [H2=2.34e-003] [1504 of 1504], Entropy: 0.373, Purity: 0.657
-------------------------------------------------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | hous mone trad rese  cpi inte  gnp reta  ipi jobs  lei  bop  wpi
-------------------------------------------------------------------------------------------------------------------------
  0   124 +0.421 +0.140 +0.034 +0.009 0.235 0.863 |    1  107    5    0    0    7    2    1    0    0    0    1    0
  1    90 +0.295 +0.142 +0.034 +0.012 0.394 0.611 |    0   55    4    1    3   26    1    0    0    0    0    0    0
  2   104 +0.193 +0.061 +0.039 +0.014 0.491 0.702 |    2   73    3    6    1    5    0    2    1    6    1    3    1
  3   104 +0.168 +0.066 +0.035 +0.015 0.072 0.962 |    0  100    1    0    0    3    0    0    0    0    0    0    0
  4   134 +0.167 +0.045 +0.045 +0.012 0.569 0.313 |    0   32   42   25    0    2    1    0    0    0    0   32    0
  5   176 +0.148 +0.046 +0.032 +0.009 0.800 0.284 |   12   17    1    0   50    2    3   15   30   22   10    0   14
  6   106 +0.136 +0.045 +0.028 +0.013 0.042 0.981 |    0    1    1    0    0  104    0    0    0    0    0    0    0
  7    90 +0.142 +0.052 +0.035 +0.014 0.200 0.856 |    0   77    7    0    0    6    0    0    0    0    0    0    0
  8   100 +0.113 +0.032 +0.030 +0.012 0.378 0.500 |    0   50   42    4    0    4    0    0    0    0    0    0    0
  9   108 +0.120 +0.030 +0.039 +0.015 0.297 0.565 |    0   61    0    0    0   45    2    0    0    0    0    0    0
 10   114 +0.118 +0.029 +0.040 +0.012 0.589 0.588 |    1    6   11    1    6    5   67    2    4   10    0    1    0
 11   166 +0.103 +0.029 +0.026 +0.009 0.092 0.952 |    0    5  158    0    0    0    2    0    0    1    0    0    0
 12    88 +0.083 +0.020 +0.032 +0.011 0.520 0.500 |    0   24   44    5    0   10    2    0    2    0    0    1    0
-------------------------------------------------------------------------------------------------------------------------

Timing Information ------------------------------------------------------------
   I/O:                                 0.027 sec
   Clustering:                          0.228 sec
   Reporting:                           0.009 sec
*******************************************************************************
```

```
.\vcluster -clmethod=direct -sim=corr -crfun=h2 -rclassfile='re0.mat.rclass'
.\re0.mat 13
```

```
*******************************************************************************
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

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

Options -----------------------------------------------------------------------
  CLMethod=Direct, 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=1.95e-003] [1504 of 1504], Entropy: 0.371, Purity: 0.666
-------------------------------------------------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | hous mone trad rese  cpi inte  gnp reta  ipi jobs  lei  bop  wpi
-------------------------------------------------------------------------------------------------------------------------
  0   114 +0.594 +0.133 +0.101 +0.031 0.095 0.947 |    0  108    3    0    0    3    0    0    0    0    0    0    0
  1    81 +0.400 +0.142 +0.084 +0.034 0.382 0.593 |    0   48    2    0    3   26    2    0    0    0    0    0    0
  2   205 +0.386 +0.084 +0.097 +0.026 0.820 0.249 |   11   35    1    0   51    3   11   16   32   20   10    0   15
  3    72 +0.355 +0.108 +0.085 +0.033 0.180 0.875 |    0   63    0    0    0    5    0    0    0    4    0    0    0
  4   173 +0.378 +0.068 +0.117 +0.029 0.230 0.792 |    1   33    0    0    1  137    1    0    0    0    0    0    0
  5   169 +0.352 +0.076 +0.099 +0.029 0.620 0.296 |    2   50   45   36    0    5    5    0    0    0    0   26    0
  6    48 +0.336 +0.096 +0.104 +0.029 0.637 0.417 |    2    2   20    3    0    0    1    2    0    7    0   11    0
  7    96 +0.308 +0.096 +0.077 +0.033 0.079 0.958 |    0   92    2    0    0    2    0    0    0    0    0    0    0
  8    90 +0.317 +0.076 +0.124 +0.034 0.568 0.611 |    0    5    7    0    5    4   55    2    3    7    1    1    0
  9   100 +0.260 +0.073 +0.068 +0.026 0.220 0.850 |    0   11   85    0    0    1    1    0    1    1    0    0    0
 10    93 +0.266 +0.051 +0.108 +0.028 0.414 0.602 |    0   56    5    2    0   26    3    0    1    0    0    0    0
 11   156 +0.200 +0.056 +0.054 +0.026 0.129 0.917 |    0   11  143    0    0    1    1    0    0    0    0    0    0
 12   107 +0.188 +0.062 +0.064 +0.028 0.187 0.879 |    0   94    6    1    0    6    0    0    0    0    0    0    0
-------------------------------------------------------------------------------------------------------------------------

Timing Information ------------------------------------------------------------
   I/O:                                 0.028 sec
   Clustering:                          8.968 sec
   Reporting:                           0.052 sec
*******************************************************************************
```

```
.\vcluster -clmethod=agglo -sim=cos -crfun=i2 -rclassfile='re0.mat.rclass'
.\re0.mat 13
```

```
*******************************************************************************
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

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

Options -----------------------------------------------------------------------
  CLMethod=AGGLO, CRfun=I2, SimFun=Cosine, #Clusters: 13
  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 ----------------------------------------------------------------------

------------------------------------------------------------------------------------------------------------------------
13-way clustering: [I2=5.79e+002] [1504 of 1504], Entropy: 0.443, Purity: 0.577
------------------------------------------------------------------------------------------------------------------------
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   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
  2    78 +0.126 +0.042 +0.027 +0.009 0.497 0.500 |    0   39   23    0    3    8    1    0    0    4    0    0    0
  3   250 +0.086 +0.028 +0.038 +0.014 0.870 0.224 |   14   48   18    5    9   15   56   13   24   32    7    9    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   177 +0.066 +0.020 +0.037 +0.014 0.518 0.424 |    0   75   64    3    0   20   11    1    2    0    1    0    0
 10    67 +0.258 +0.065 +0.038 +0.008 0.404 0.657 |    0    1    1    0   44    0    0    3    4    0    0    0   14
 11    49 +0.245 +0.068 +0.027 +0.012 0.039 0.980 |    0    1    0    0    0   48    0    0    0    0    0    0    0
 12   165 +0.081 +0.026 +0.037 +0.014 0.374 0.527 |    0   67    3    1    0   87    6    0    0    0    0    1    0
------------------------------------------------------------------------------------------------------------------------

Timing Information ------------------------------------------------------------
   I/O:                                 0.030 sec
   Clustering:                          0.219 sec
   Reporting:                           0.011 sec
*******************************************************************************
```

```
.\vcluster -clmethod=agglo -sim=corr -crfun=i2 -rclassfile='re0.mat.rclass'
.\re0.mat 13
```

```
*******************************************************************************
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

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

Options -----------------------------------------------------------------------
  CLMethod=AGGLO, CRfun=I2, 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=I2, NTrials=10, NIter=10

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

------------------------------------------------------------------------------------------------------------------------
13-way clustering: [I2=8.36e+002] [1504 of 1504], Entropy: 0.374, Purity: 0.651
------------------------------------------------------------------------------------------------------------------------
cid  Size  ISim  ISdev   ESim  ESdev  Entpy Purty | hous mone trad rese  cpi inte  gnp reta  ipi jobs  lei  bop  wpi
------------------------------------------------------------------------------------------------------------------------
  0   147 +0.264 +0.077 +0.119 +0.030 0.731 0.463 |    3   16    9    0    8   10   68    6    9   11    4    2    1
  1   131 +0.204 +0.060 +0.056 +0.026 0.125 0.939 |    1    3  123    0    0    0    0    0    2    1    0    1    0
  2    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
  3   100 +0.679 +0.076 +0.102 +0.028 0.022 0.990 |    0   99    0    0    0    1    0    0    0    0    0    0    0
  4   119 +0.176 +0.058 +0.100 +0.039 0.340 0.639 |    0   76    3    0    0   35    5    0    0    0    0    0    0
  5   172 +0.345 +0.082 +0.096 +0.028 0.561 0.349 |    1   41   60   33    0    1    1    0    0    0    0   35    0
  6   166 +0.118 +0.038 +0.079 +0.039 0.467 0.542 |    0   90   49    8    4   11    4    0    0    0    0    0    0
  7    60 +0.395 +0.093 +0.084 +0.031 0.128 0.917 |    0   55    0    0    0    1    0    0    0    4    0    0    0
  8   185 +0.375 +0.096 +0.103 +0.029 0.790 0.259 |   11   36    1    0   48    4    1   14   26   23    7    0   14
  9   161 +0.389 +0.069 +0.121 +0.027 0.198 0.814 |    0   29    0    1    0  131    0    0    0    0    0    0    0
 10    71 +0.306 +0.068 +0.067 +0.022 0.079 0.958 |    0    2   68    0    0    0    1    0    0    0    0    0    0
 11    50 +0.526 +0.068 +0.087 +0.033 0.000 1.000 |    0   50    0    0    0    0    0    0    0    0    0    0    0
 12    83 +0.230 +0.064 +0.081 +0.031 0.200 0.855 |    0   71    6    0    0    6    0    0    0    0    0    0    0
------------------------------------------------------------------------------------------------------------------------

Timing Information ------------------------------------------------------------
   I/O:                                 0.032 sec
   Clustering:                          6.578 sec
   Reporting:                           0.051 sec
*******************************************************************************
```

```
.\vcluster -clmethod=agglo -sim=cos -crfun=h2 -rclassfile='re0.mat.rclass'
.\re0.mat 13
*******************************************************************************
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

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

Options -----------------------------------------------------------------------
  CLMethod=AGGLO, CRfun=H2, SimFun=Cosine, #Clusters: 13
  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 ----------------------------------------------------------------------

-------------------------------------------------------------------------------------------------------------------------
13-way clustering: [H2=2.11e-003] [1504 of 1504], Entropy: 0.449, Purity: 0.578
-------------------------------------------------------------------------------------------------------------------------
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   110 +0.193 +0.044 +0.036 +0.016 0.103 0.936 |    0  103    1    0    0    6    0    0    0    0    0    0    0
  2   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
  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   129 +0.139 +0.047 +0.044 +0.016 0.728 0.279 |    1   29   36   19    3    7    3    1    2    0    2   25    1
  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    85 +0.212 +0.078 +0.044 +0.013 0.363 0.765 |    1   65    0    6    0    5    4    0    2    2    0    0    0
 10    83 +0.107 +0.042 +0.032 +0.011 0.386 0.542 |    0   45   24    0    0   14    0    0    0    0    0    0    0
 11    99 +0.111 +0.033 +0.041 +0.013 0.732 0.414 |    0    8   11    1   10    6   41    2    6    9    0    5    0
 12    69 +0.110 +0.033 +0.036 +0.017 0.585 0.377 |    0   26   25    1    3    3    7    0    1    0    0    3    0
-------------------------------------------------------------------------------------------------------------------------

Timing Information ------------------------------------------------------------
   I/O:                                 0.030 sec
   Clustering:                          4.771 sec
   Reporting:                           0.009 sec
*******************************************************************************
```

```
.\vcluster -clmethod=agglo -sim=corr -crfun=h2 -rclassfile='re0.mat.rclass'
.\re0.mat 13
*******************************************************************************
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

Matrix Information ------------------------------------------------------------
  Name: .\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:                                 0.030 sec
   Clustering:                         11.254 sec
   Reporting:                           0.057 sec
*******************************************************************************
```

Análisis de Resultados

Una vez realizadas las 8 ejecuciones posibles con los parámetros seleccionados, podemos comentar las medidas que nos proporcionan las ejecuciones:

Método **F. Similitud** **F. Criterio**	direct cos i2	direct corr i2	direct cos h2	direct corr h2	agglo cos i2	agglo corr i2	agglo cos h2	agglo corr i2
Medidas Internas								
F. Criterio	6,04e+002	8,65e+002	2,34e-003	1,95e-003	5,79e+002	8,36+002	2,11e-003	1,83e-003
Medidas Externas								
Entropía	0,382	0,370	0,373	0,371	0,577	0,374	0,449	0,377
Pureza	0,659	0,666	0,657	0,666	0,443	0,651	0,578	0,662

Según las medidas internas, todas las ejecuciones han asignado todos los documentos a clúster (1504 de 1504), pero evaluando las dos funciones de criterio podemos decir que el algoritmo que ofrece mejores resultados es el de partición `direct` y que concretamente el caso donde la función de similitud es `corr` y la medida de criterio es `i2`, ofrece mejores resultados de agrupamiento, basándonos en las medidas de entropía y pureza.

En líneas generales, las ejecuciones donde aplicamos el algoritmo aglomerativo obtiene peores resultados que los de partición. Esto se debe a que este tipo de métodos suele propagar errores que se pueden producir durante las primeras decisiones de agrupación.