Clustering with `textmin`

Ingo Feinerer

This document shows typical text clustering examples that can be performed with `textmin`.

Initialization

```
> library(textmin)

Loading required package: XML

> data(ReutNews)
> tdm <- termdocmatrix(ReutNews)
```

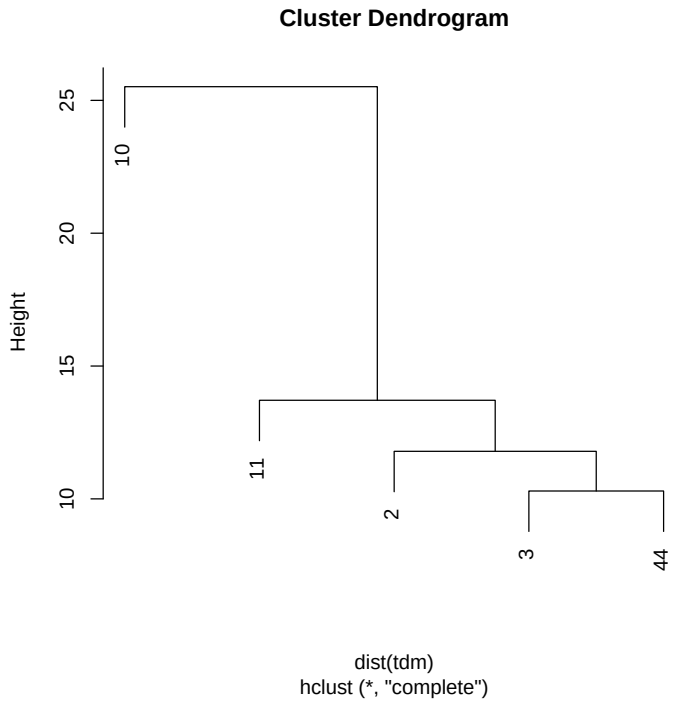
creates a text document collection and the term-document matrix for it.

Hierarchical Clustering

```
> hcl <- hclust(dist(tdm))
```

performs a hierarchical clustering on the distance matrix of the original term-document matrix. The visualized result looks like:

```
> plot(hcl)
```



k-means Clustering

```
> kmeans(tdm, 2)

K-means clustering with 2 clusters of sizes 4, 1

Cluster means:
  <bp> activities also america and borrowing both british called committee
1 0.25       0.25 0.25     0.5 1.5      0.25 0.25    0.25   0.25      0.25
2 0.00       0.00 2.00     0.0 6.0      0.00 0.00    0.00   0.00      0.00
  companies financial form  inc interest investment joint manage management
1      0.25      0.25 0.25 1.25     0.25       0.25  0.25   0.25       0.25
2      0.00      0.00 0.00 1.00     0.00       0.00  0.00   0.00       0.00
  market money north  oil operated oversight owns  pct petroleum plan  plc
1   0.25  0.25   0.5 0.75     0.25      0.25 0.25 0.25      0.25 0.25 0.25
2   1.00  0.00   0.0 0.00     0.00      0.00 0.00 2.00      0.00 1.00 0.00
  reuter said standard subsidiary   the they trading under venture which will
1      1 1.25        1       0.75  3.25 0.25    0.25  0.25     0.5  0.25  0.5
2      1 7.00        0       0.00 15.00 0.00    0.00  1.00     0.0  0.00  0.0
  application assets bancshares bank banking banks billion commerce comptroller
1        0.25   0.75       0.25  0.5    0.25  0.25     0.5      0.5        0.25
2        0.00   0.00       0.00  0.0    0.00  0.00     0.0      0.0        0.00
  county create currency deposits dlrs effort filed harris having houston
1   0.25   0.25     0.25      0.5    1   0.25  0.25   0.25   0.25    0.25
2   0.00   0.00     0.00      0.0    4   0.00  0.00   0.00   0.00    1.00
```

```
  largest link network texas with would  000 125 200 <sedio <woodco acquire
1    0.25 0.25     0.5   0.5 0.25   0.5 0.75   0   0      0       0       0
2    0.00 0.00     0.0   0.0 0.00   3.0 3.00   1   2      1       1       1
  additional any  are but buy certain change circumstances common company
1          0   0 0.25   0   0    0.25      0             0      0       0
2          2   1 1.00   1   1    1.00      1             1      3       4
  completed computer conditions continue control costs current delivery dot
1      0.25        0       0.25        0       0     0       0        0   0
2      1.00        6       1.00        1       1     1       1        1   1
  ensure equal exceed exclusive exercisable five  for forms future generated
1      0     0      0         0           0 0.25 0.25     0      0         0
2      1     1      1         1           2 1.00 4.00     1      1         1
   has help holdings impact improvements inc> including increase involving  its
1 0.25    0        0      0            0    0      0.25        0         0 0.25
2 2.00    1        1      1            1    1      1.00        1         1 5.00
  labels licensee lugano makes matrix mln moves  not occur one operation
1      0        0      0     0      0 1.5     0 0.25     0   0         0
2      1        1      1     1      1 1.0     1 1.00     1   1         1
  outstanding part pay  per price printers product purchase reorganization
1           0    0   0 0.25     0        0       0        0              0
2           1    1   1 2.00     3        1       1        1              1
  right rights sale sedio share shares sold stock switzerland systems tags
1     0      0    0     0     0      0    0     0           0       0    0
2     1      1    1     1     2      3    1     3           1       1    1
  technolgy technology terminal terminals tex ticket time total warrants were
1         0          0        0         0   0      0    0     0        0    0
2         1          2        5         1   1      1    1     1        3    1
  woodco worldwide years  132 1985  299  327  440  472  479  4th  510  807  858
1      0         0     0 0.25 0.25 0.25 0.25 0.25 0.25 0.25 0.25 0.25 0.25 0.25
2      1         1     1 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
  available carry cts extraordinary forward from gain includes loans  net note
1      0.25  0.25 0.5          0.25    0.25 0.25 0.25     0.25  0.25 0.25 0.25
2      0.00  0.00 0.0          0.00    0.00 0.00 0.00     0.00  0.00 0.00 0.00
   qtr shr  tax year <crowley agreed american approval arranging bankruptcy
1 0.25 0.5 0.25 0.25     0.25   0.25      0.5     0.25      0.25       0.25
2 0.00 0.0 0.00 0.00     0.00   0.00      0.0     0.00      0.00       0.00
  bodies charters contract corp> court expected industries lines mariotime
1   0.25     0.25      0.5  0.25  0.25     0.25       0.25  0.75      0.25
2   0.00     0.00      0.0  0.00  0.00     0.00       0.00  0.00      0.00
  mclean negotiations next principle regulatory service south states subject
1   0.25         0.25 0.25      0.25       0.25    0.25  0.25   0.25    0.25
2   0.00         0.00 0.00      0.00       0.00    0.00  0.00   0.00    0.00
  terms transfer transport united various week within
1  0.25      0.5      0.25   0.25    0.25 0.25   0.25
2  0.00      0.0      0.00   0.00    0.00 0.00   0.00

Clustering vector:
 2  3 10 11 44
 1  1  2  1  1
```

```
Within cluster sum of squares by cluster:
[1] 220.25   0.00

Available components:
[1] "cluster"  "centers"  "withinss" "size"
```

performs a k-means clustering with 2 clusters on the term-document matrix.

Spectral clustering

The following example shows a spectral clustering of the text document collection with String Kernels form the `kernlab` package.

```
> library(kernlab)
> stringkern <- stringdot(type = "string")
> specc(ReutNews, 2, kernel = stringkern)

Spectral Clustering object of class "specc"

 Cluster memberships:

2 2 2 1 2

String kernel function.  Type =  string
 Hyperparameters : sub-sequence/string length =  4  lambda =  0.5
 Normalized

Centers:
     [,1]
[1,]   NA

Cluster size:
[1] 1 4

Within-cluster sum of squares:
logical(0)
```