Xplortext

Wiesbaden R Users Group

5 May 2021

Las Médulas. Spanish Roman over ground gold mine

Authors:

Ramón Alvarez-Esteban (1) (Maintainer)

Mónica Bécue-Bertaut (2)

Josep-Anton Sánchez-Espigares (2)

Belchin Kostov (2)

(1) University of Leon / Spain. ramon.alvarez@unileon.es
(2) UPC Universitat Politècnica de Catalunya / Spain

Xplortext

Multivariate statistical methods for analyze textual data

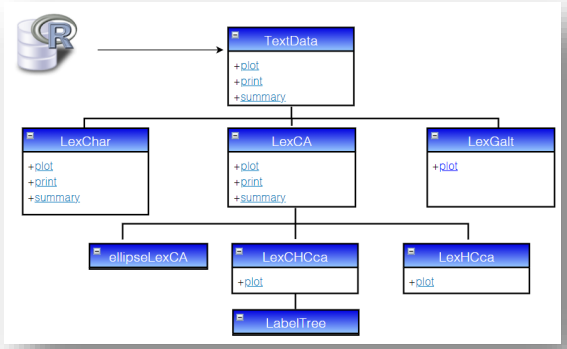
http://www.xplortext.org/

https://xplortext.unileon.es/

This script

Home | Flowchart | Analyse textuelle avec R | Textual Data Science With R | Spanish political speeches

R User Group



- print
- summary
- plot

Packages:

tm(>= 0.7-8)
slam(>= 0.1-48)

ggplot2
ggdendro(>= 0.1.22)
ggforce(>= 0.3.2)
ggrepel(>= 0.9.0),
graphics
gridExtra(>= 2.3),

FactoMineR, MASS, methods, stats, utils, flexclust, flashClust

stringi(>= 1.5.3), stringr(>= 1.4.0)

Open question:

what is the most important thing in life?

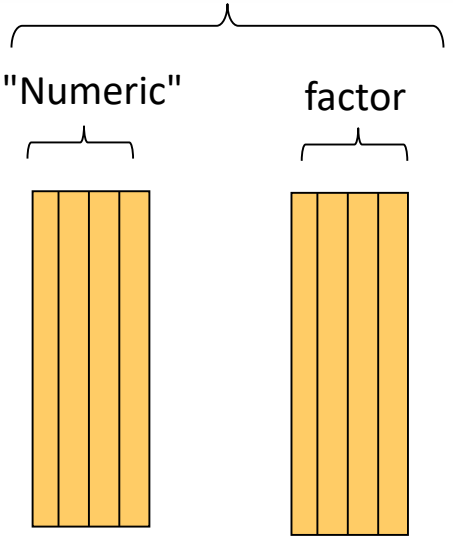
Enriching the processing of texts with contextual variables

UTF-8

Textual variable (Corpus)

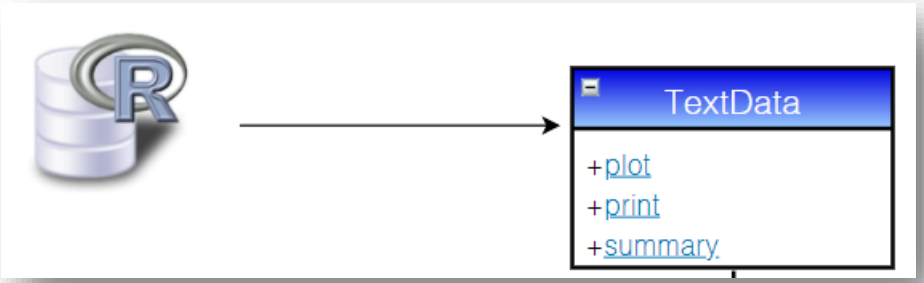
Documents

1	Good health. Happiness
2	Healthy, have enough to eat,…
.	
.	
.	
1043	job, good life, money, health



Working with dataframes (rownames for the documents)
No tibbles

UTF-8



TextData function

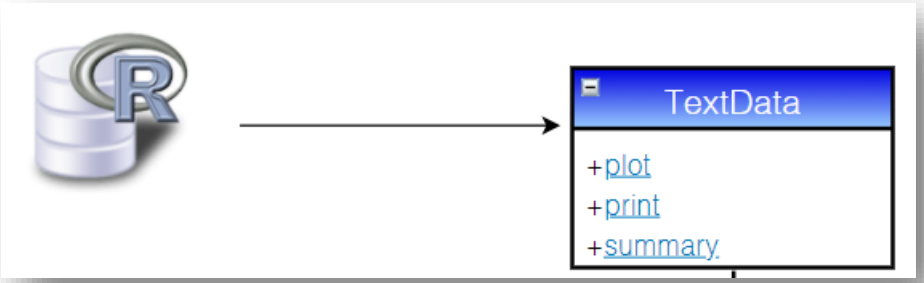
var.agg
context.quali
context.quanti
selDoc

tm options
lower=TRUE, remov.number,
lminword, Fmin, Dmin, Fmax
stop.word.tm=FALSE, idiom="en"
stop.word.user

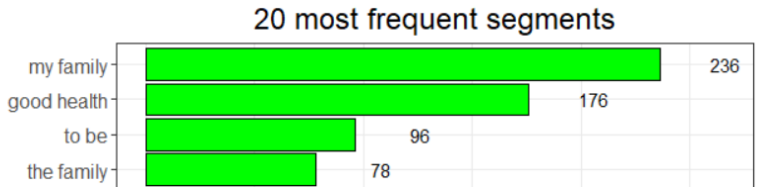
segments

```
res.TD   <-TextData(data, var.text=c(9,10), var.agg="Age12Categ", remov.number=TRUE,
                    Fmin=10, Dmin=10, stop.word.tm=TRUE)
summary(res.TD)
```

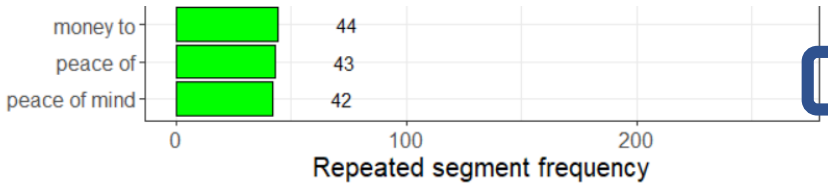
	Before	After
Documents	1043.00	12.00
Occurrences	13917.00	5911.00
Words	1334.00	132.00
Mean-length	13.34	492.58
NonEmpty.Docs	1040.00	12.00
NonEmpty.Mean-length	13.38	492.58



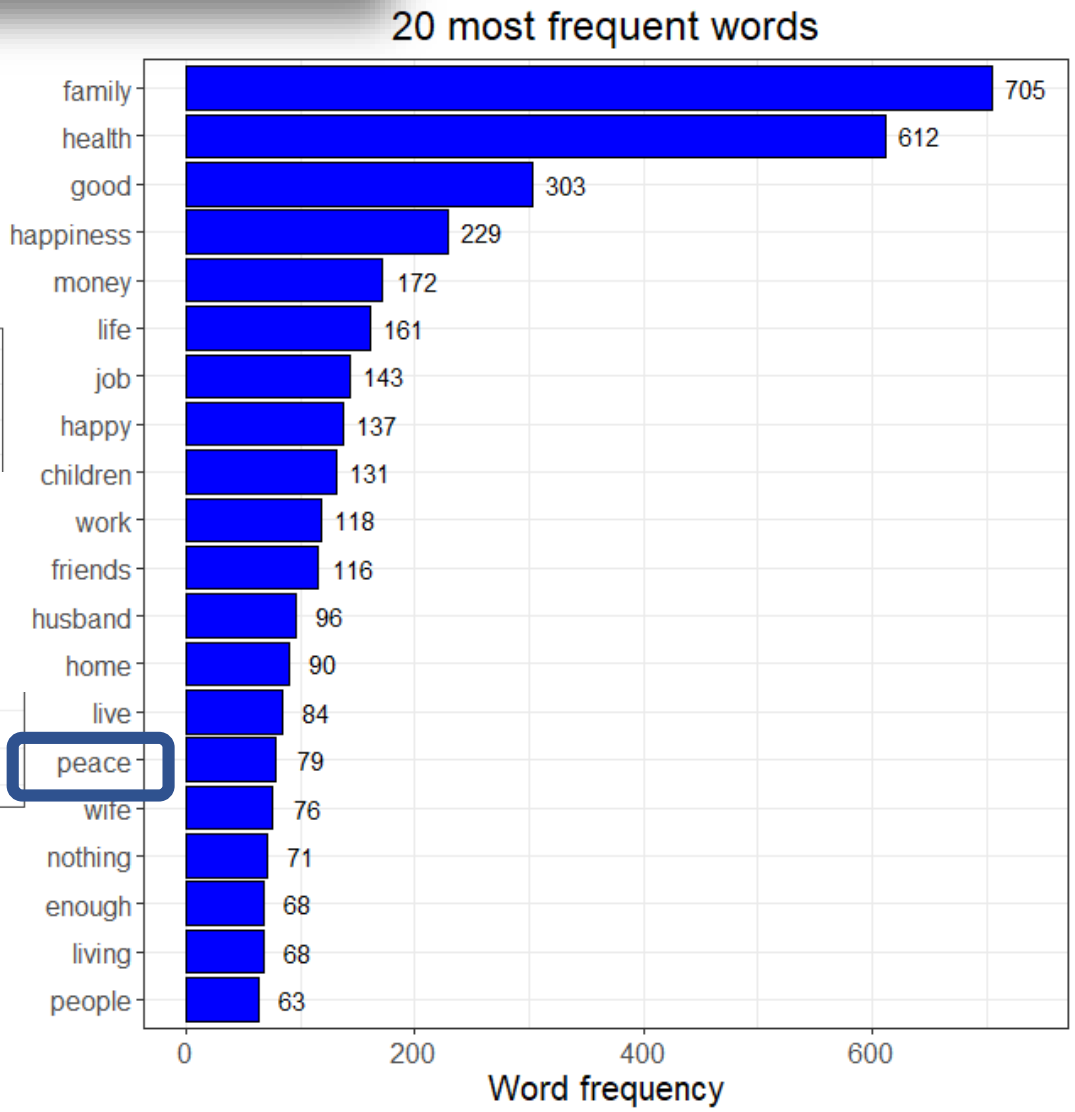
plot(res.TD, col.fill="blue")



peace of mind: 42

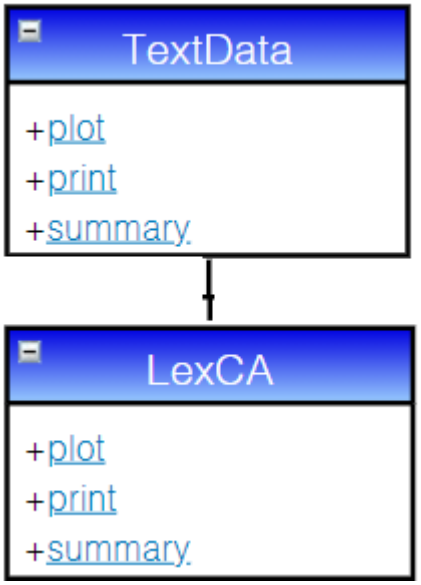


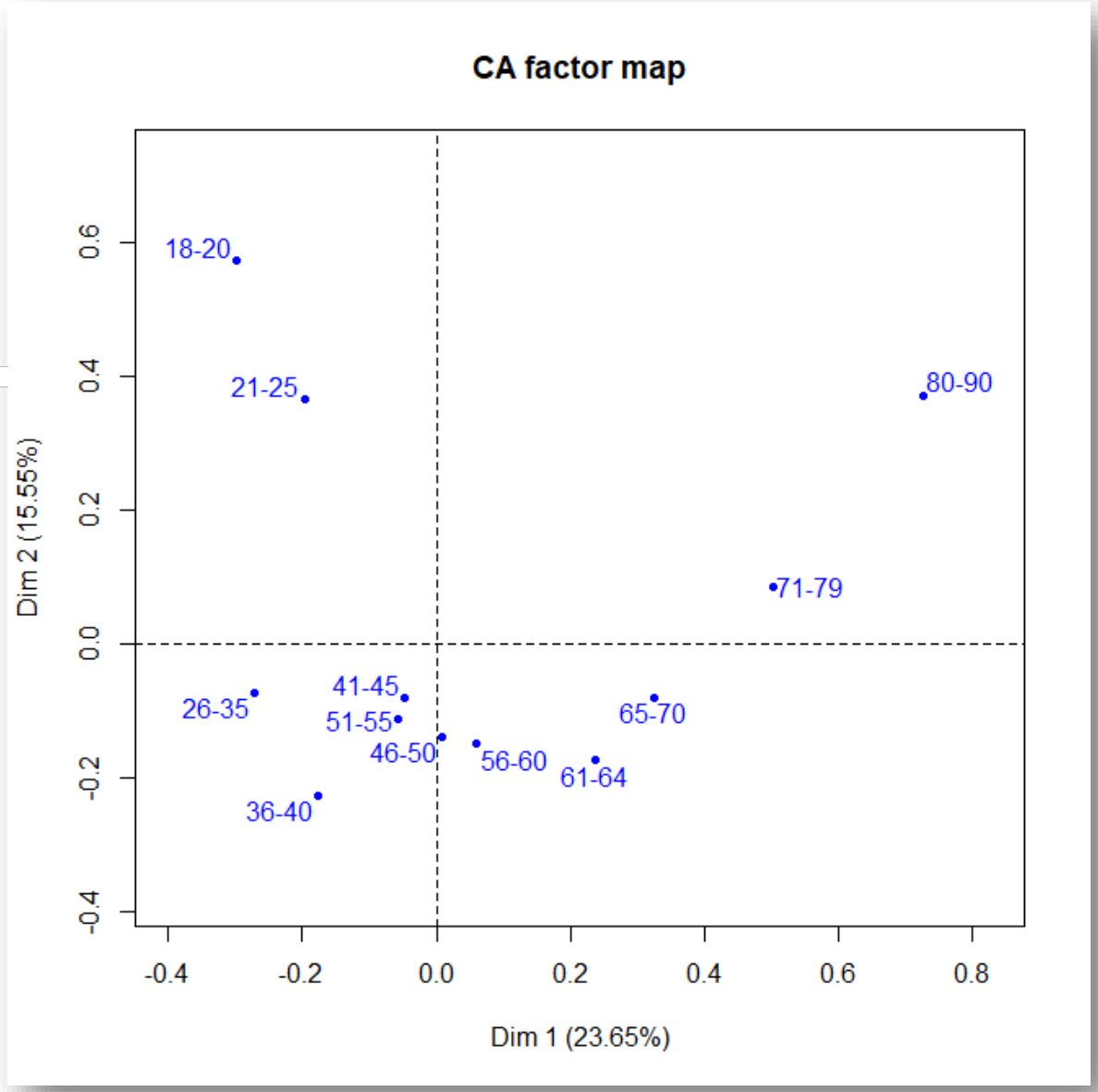
peace in the world: 6



LexCA function (Correspondence Analysis) (CA)

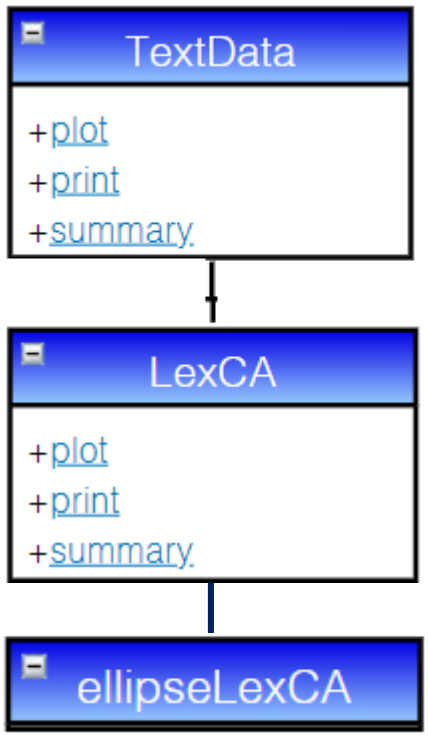
```
res.LexCA.2F <- LexCA(res.TD, ncp=2, graph=FALSE)
plot(res.LexCA.2F, selWord=NULL)
```

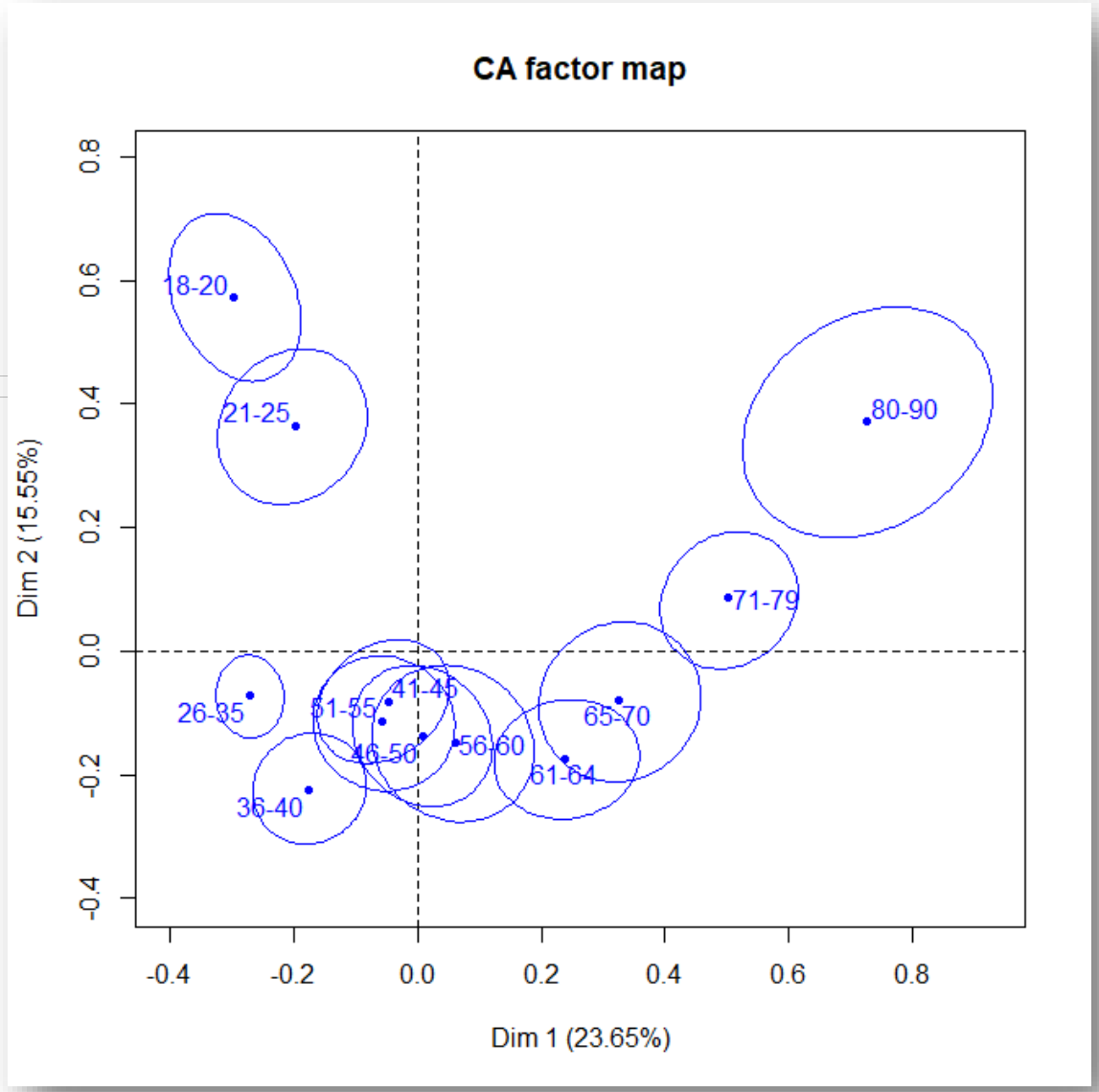




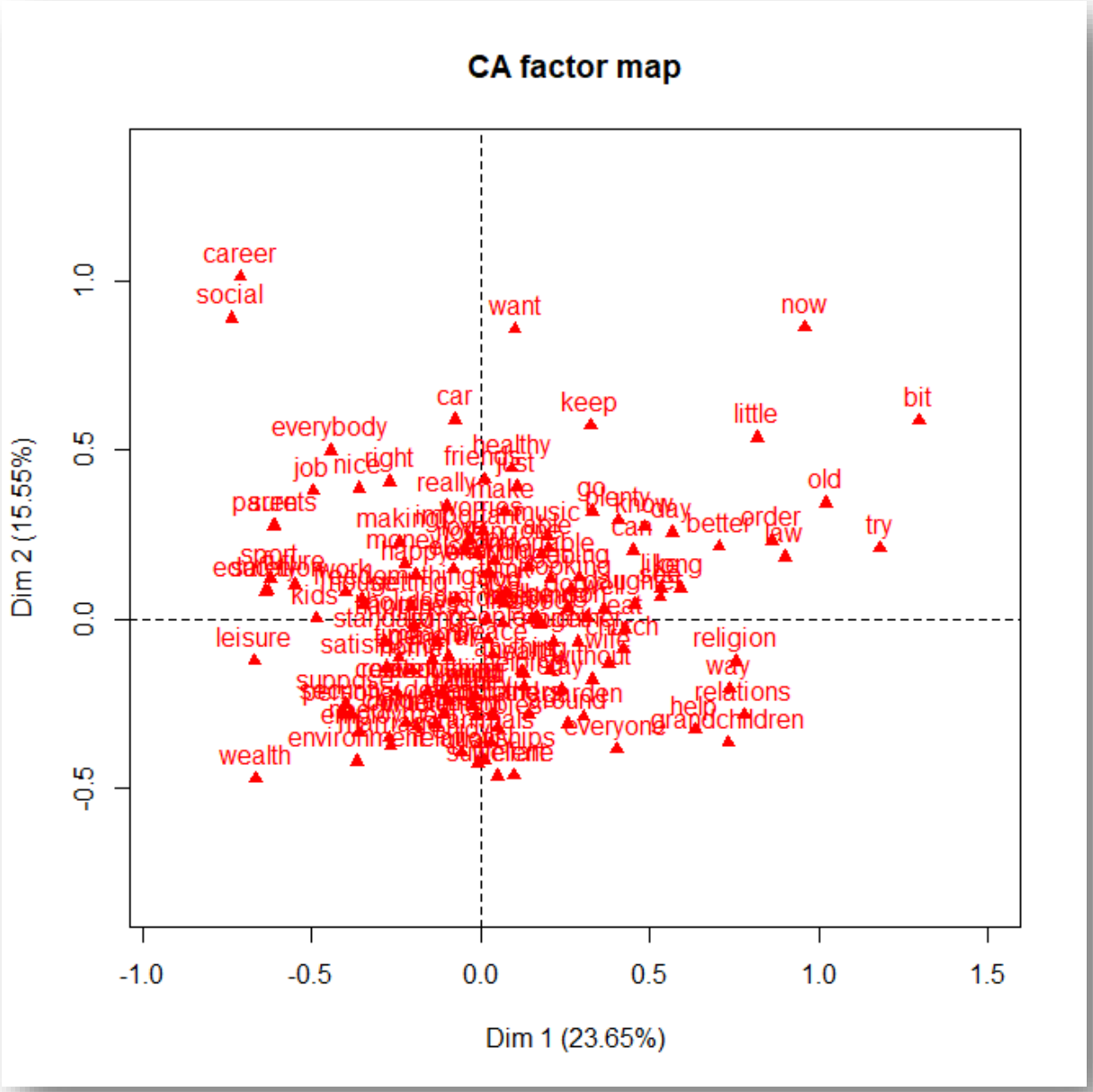
LexCA function (Correspondence Analysis) (CA) / ellipseLexCA

```
res.LexCA.2F <- LexCA(res.TD, ncp=2, graph=FALSE)
ellipseLexCA(res.LexCA.2F, selWord=NULL, selDoc="ALL")
```

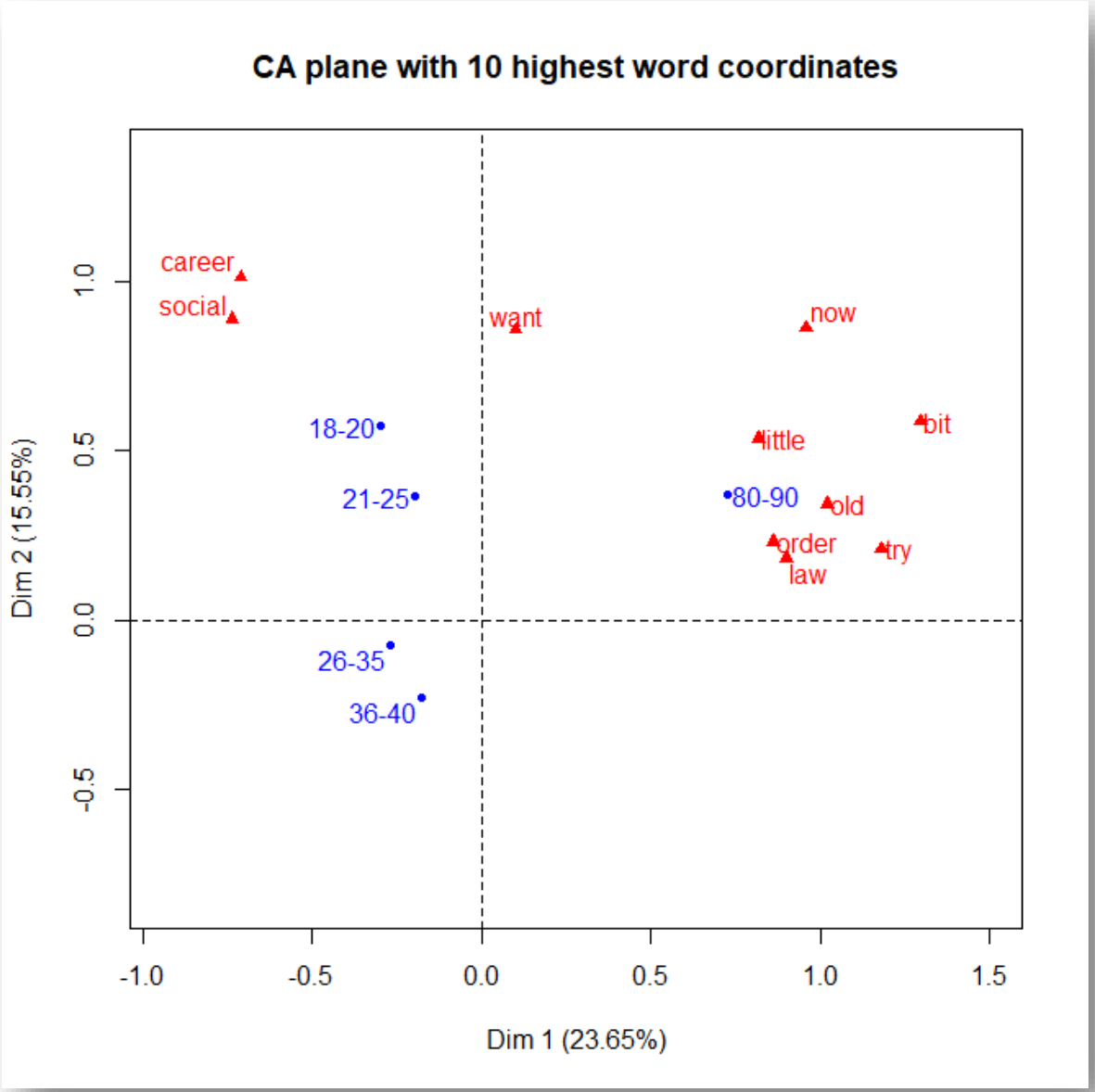




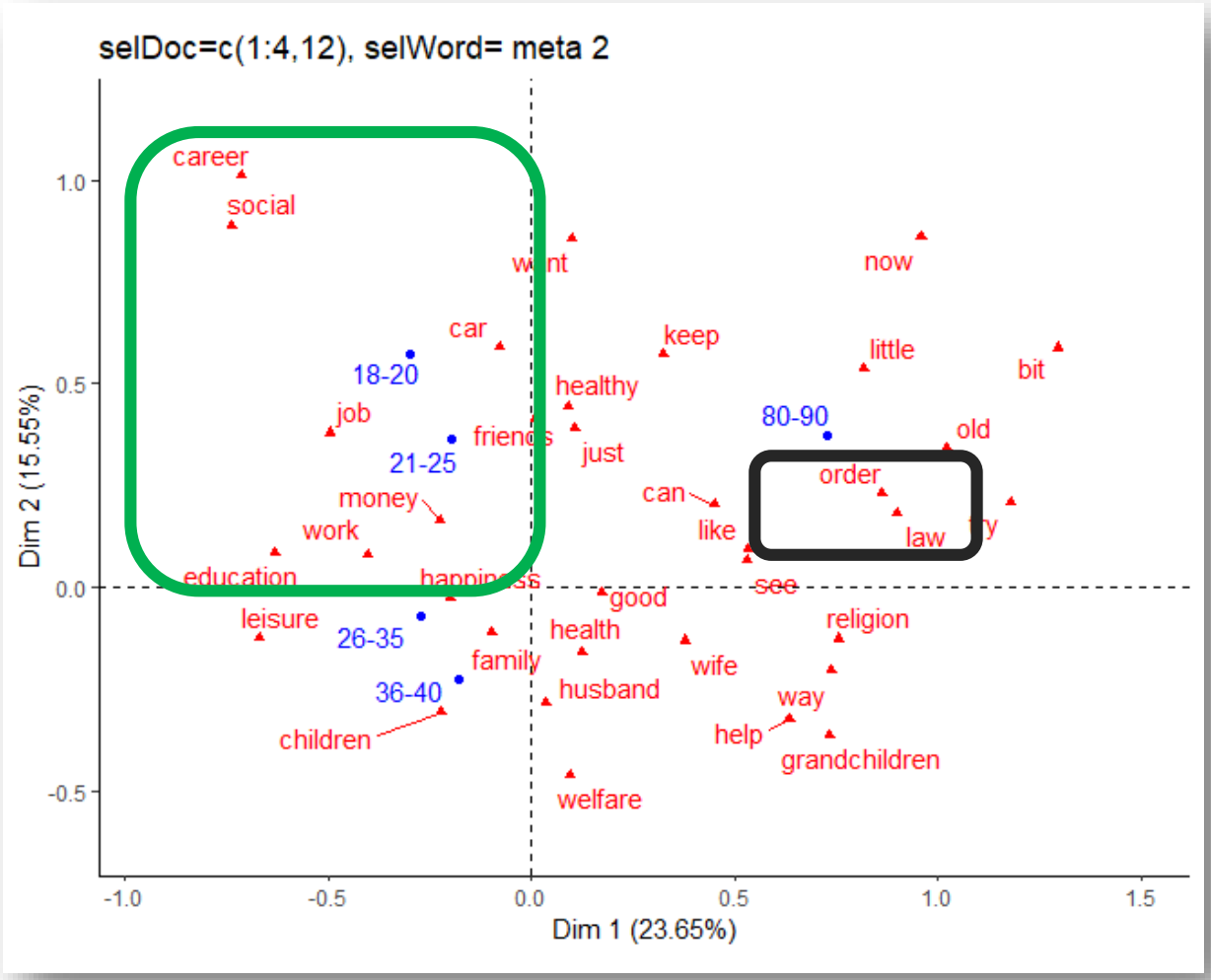
plot(res.LexCA.2F, selDoc=NULL)



plot(res.LexCA.2F, selDoc=c(1,2,3,4, 12), selWord="coord 10")



plot(res.LexCA.2F, selDoc=c(1:4,12), selWord="meta 2")



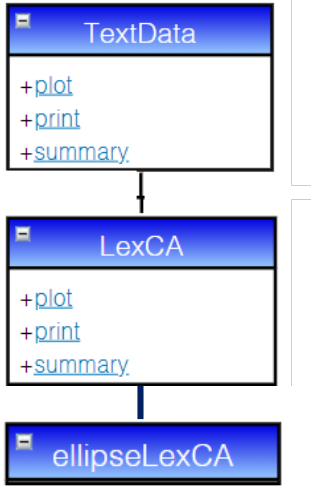
cos2

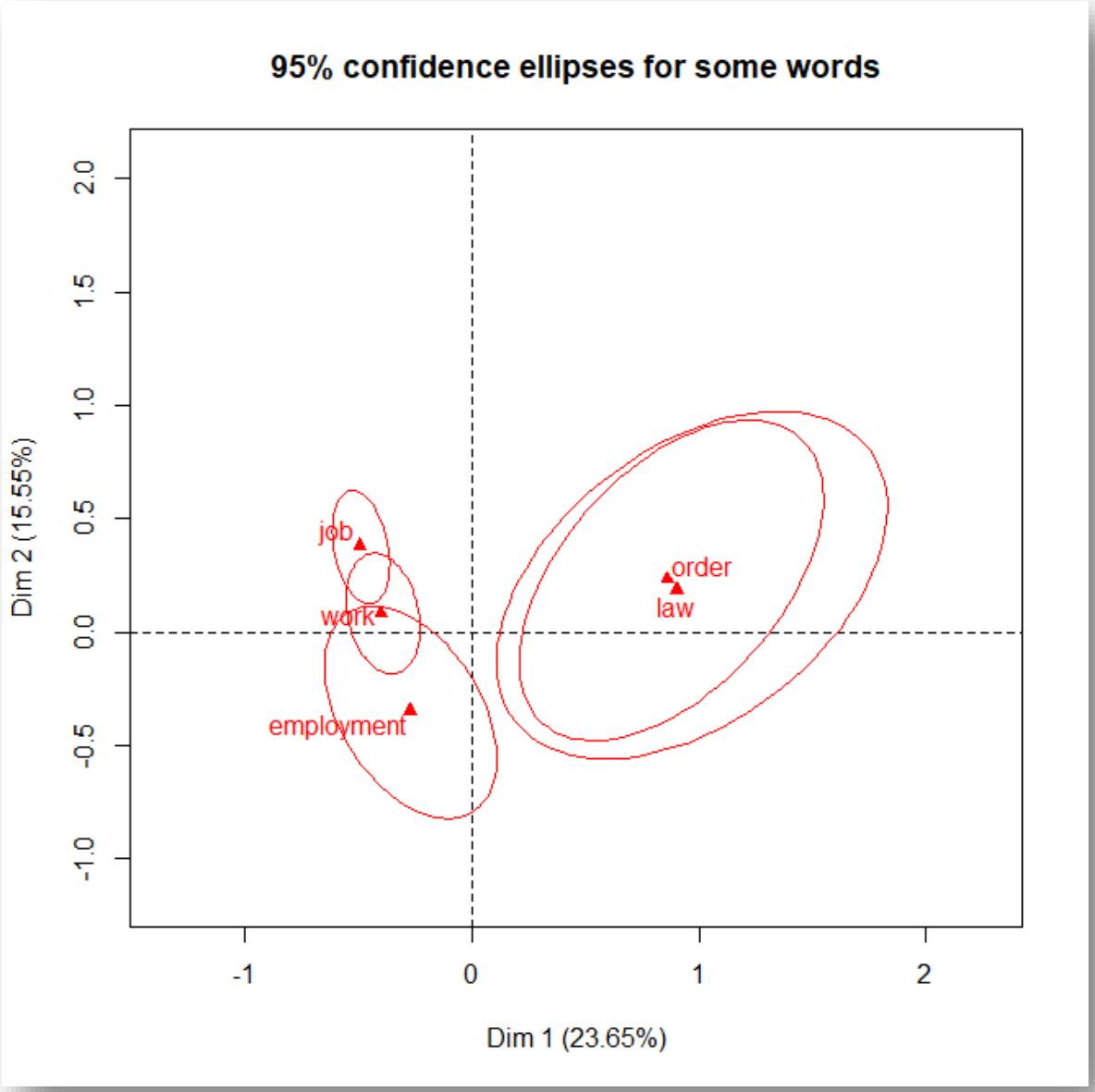
meta 2:

the words with a contribution over 2 times the average word contribution on any of the two axes are plotted.

LexCA function (Correspondence Analysis) (CA) / ellipseLexCA

ellipseLexCA(ellipseLexCA(res.LexCA.2F, selDoc=NULL, selWord=c("car","order","law", "job", "work", "employment", "money"))

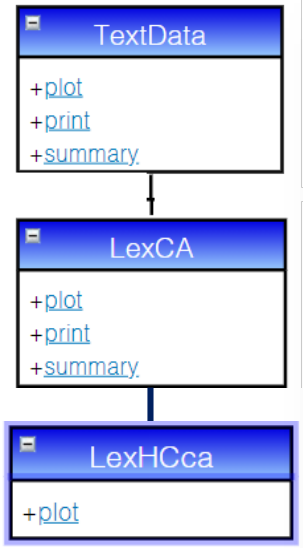


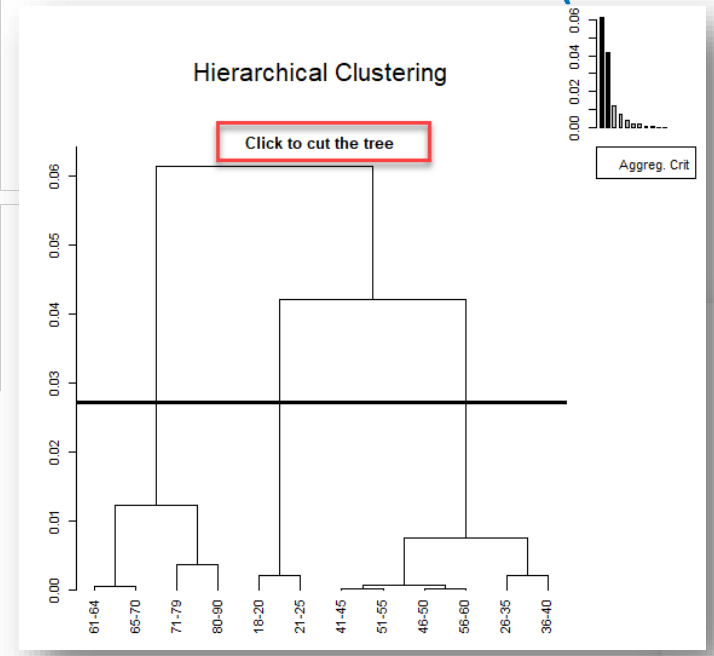


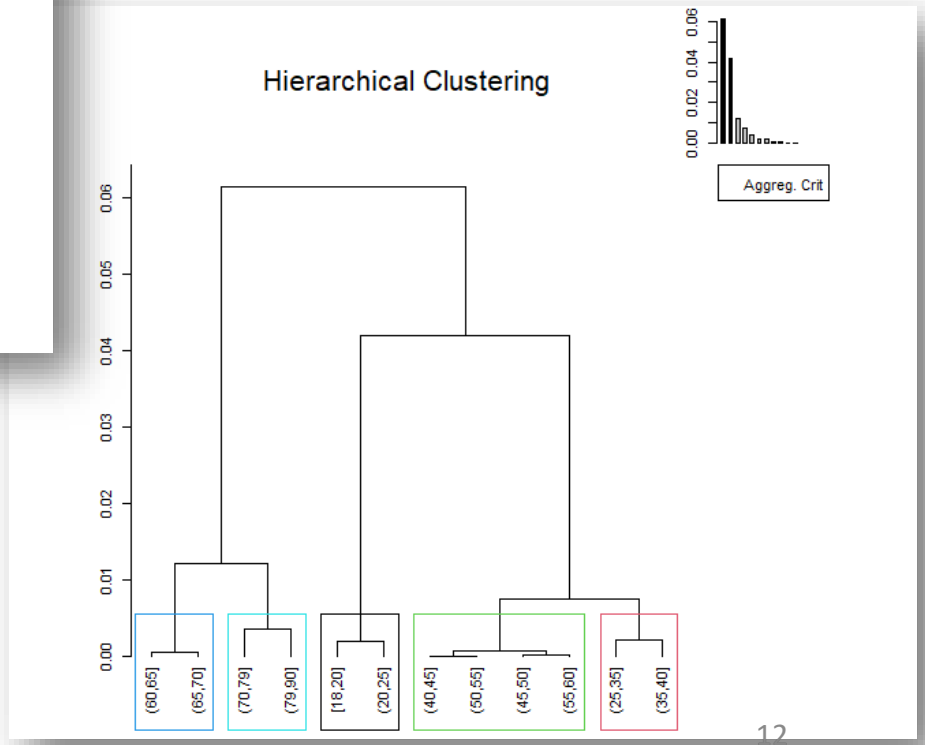
LexHCca function
Hierarchical Clustering of Documents/Words on Textual Correspondence Analysis Coordinates)

```
res.LexCA.2F <- LexCA(res.TD, ncp=2, graph=FALSE)
res.LexHCca.2F.5C <-  LexHCca(res.LexCA.2F, nb.clust="click")
res.LexHCca.2F.5C <-  LexHCca(res.LexCA.2F, nb.clust=5)
```

cluster.CA="docs"
cluster.CA="words"

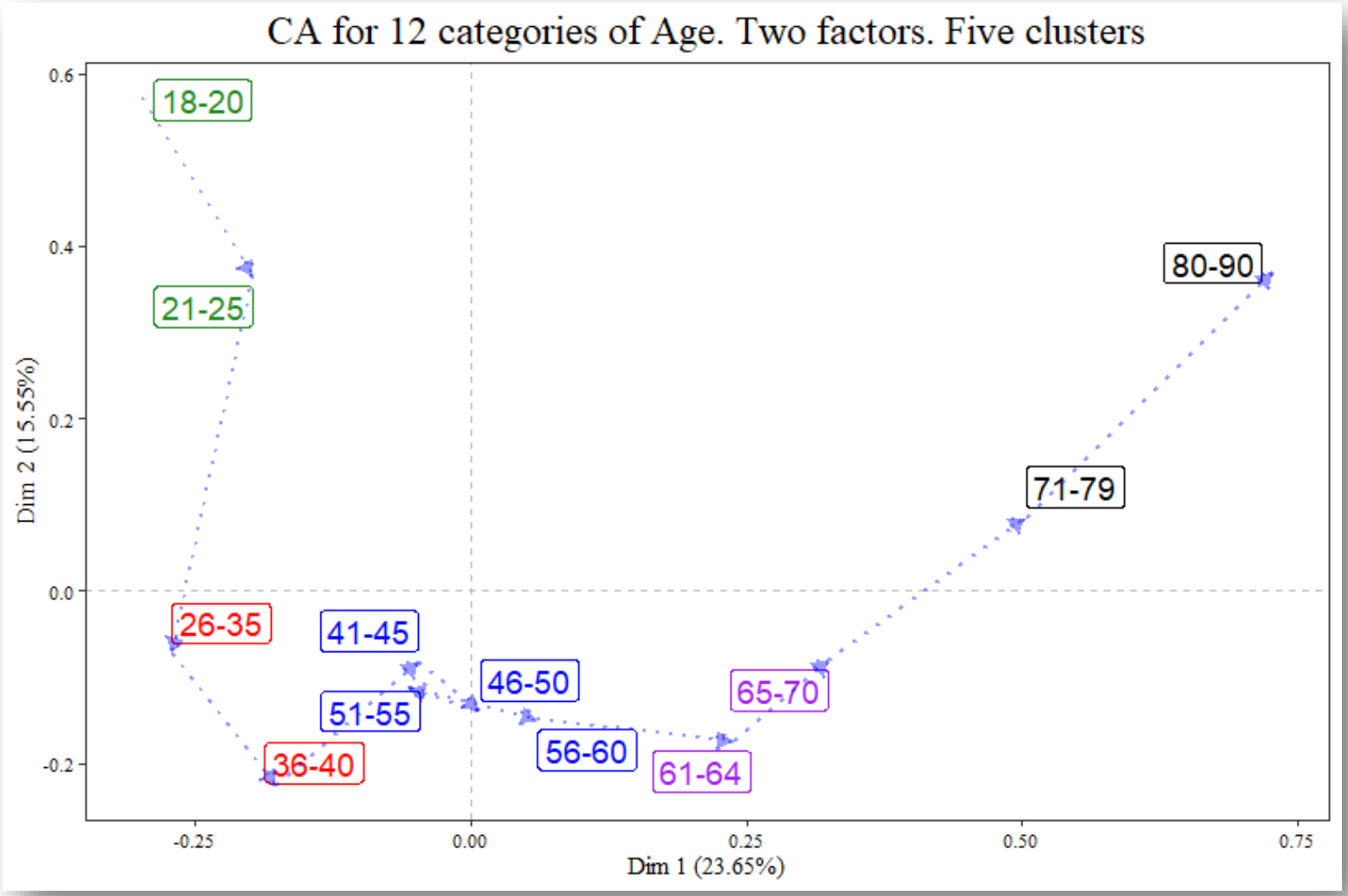


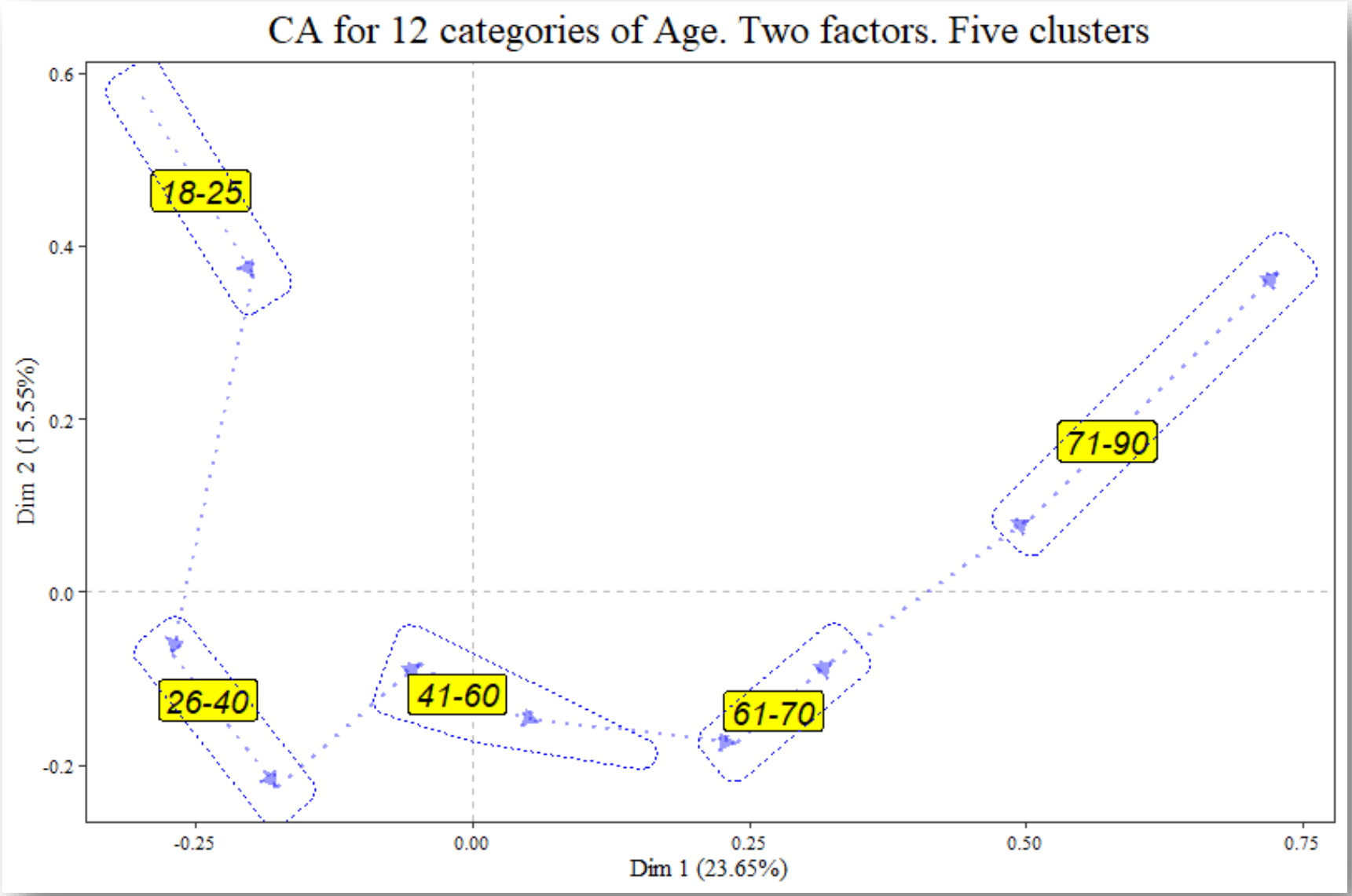




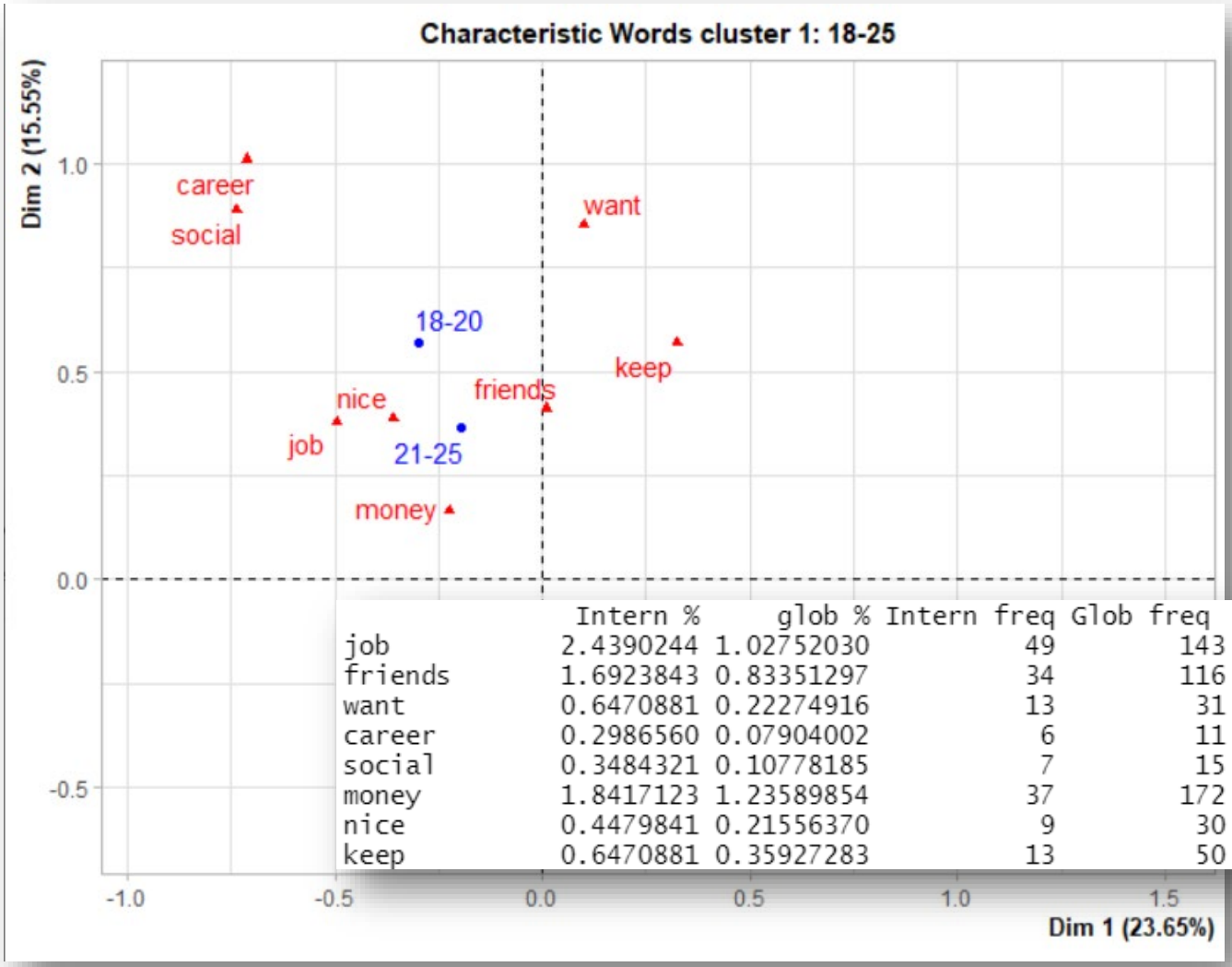

```
class(res.LexHCca.2F.5C$call$t$tree)
hclust
```

```
c.palette <- c("forestgreen", "red", "blue", "purple", "black")
plot(res.LexHCca.2F.5C, plot=c("labels","traject"),
    labels=c(size=5, force=0.2, set.seed=998, rect=TRUE), palette=c.palette,
    traject=c(color="blue", linetype="dotted", alpha.t=.4),
    title=c(text="CA for 12 categories of Age. Two factors. Five clusters"))
```



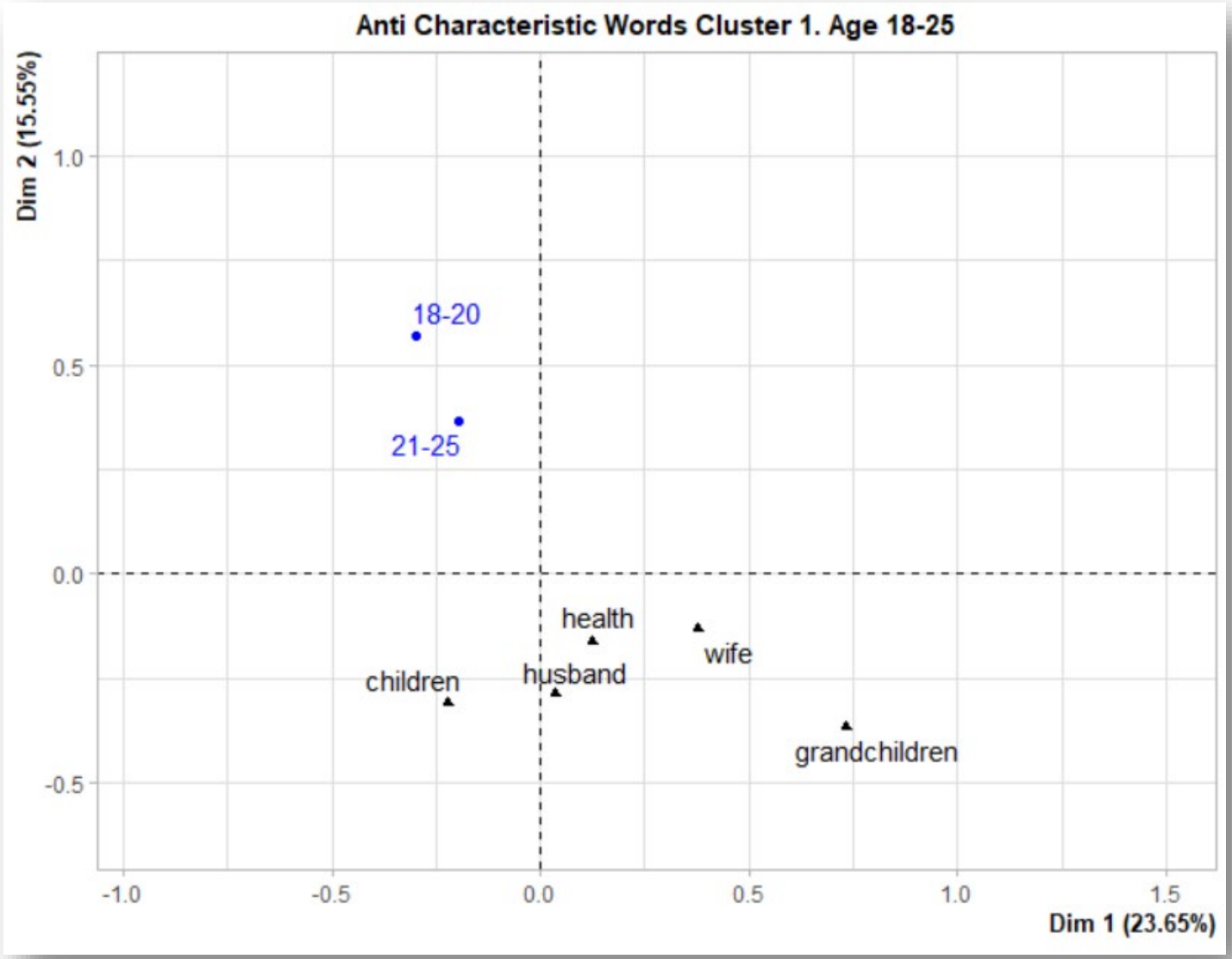

CA for 12 categories of Age. Two factors. Five clusters
18-25
26-40
41-60
61-70
71-90
Dim 2 (15.55%)
Dim 1 (23.65%)
0.6
0.4
0.2
0.0
-0.2
-0.25
0.00
0.25
0.50
0.75

```
df <- as.data.frame(res.LexHCca.2F.5C$description$desc.cluster.doc$words$cluster_1)
words_1 <- rownames(df[df$v.test>0,])
plot(res.LexCA.2F, selDoc=c(1,2), selWord=words_1, graph.type = "ggplot",
  title="Characteristic Words 18-25")
```



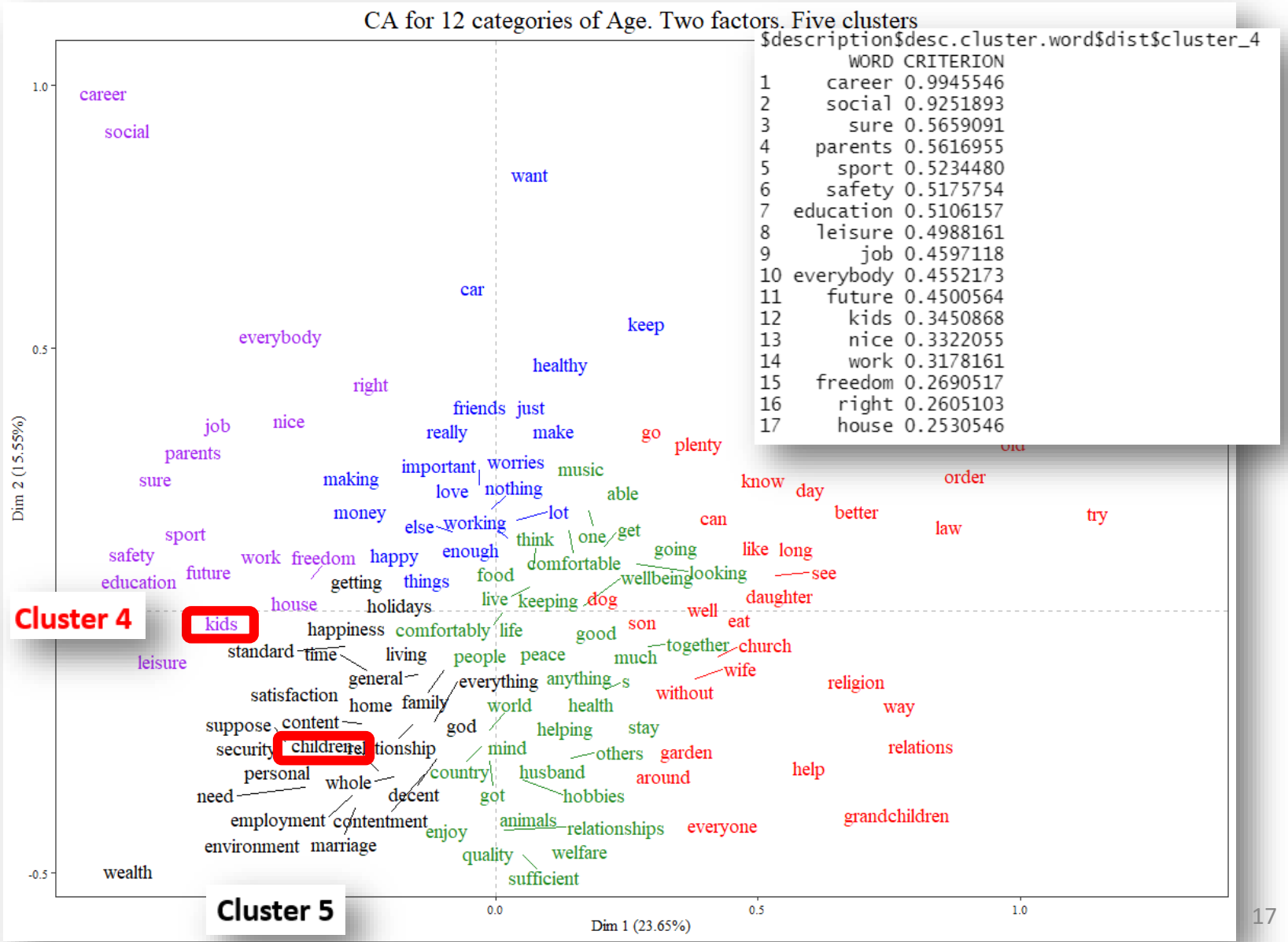
	Intern %	glob %	Intern freq	Glob freq	p.value	v.test
job	2.4390244	1.02752030	49	143	3.766107e-09	5.894152
friends	1.6923843	0.83351297	34	116	5.342068e-05	4.040134
want	0.6470881	0.22274916	13	31	3.659746e-04	3.563477
career	0.2986560	0.07904002	6	11	4.317838e-03	2.853957
social	0.3484321	0.10778185	7	15	5.730287e-03	2.762829
money	1.8417123	1.23589854	37	172	1.491207e-02	2.434508
nice	0.4479841	0.21556370	9	30	4.399428e-02	2.014145
keep	0.6470881	0.35927283	13	50	4.480995e-02	2.006434

```
anti_words_1 <- rownames(df[df$v.test<0,])
plot(res.LexCA.2F, selDoc=c(1,2), selWord=anti_words_1, col.word = "black", graph.type = "ggplot",
    title="Anti Characteristic Words 18-25")
```

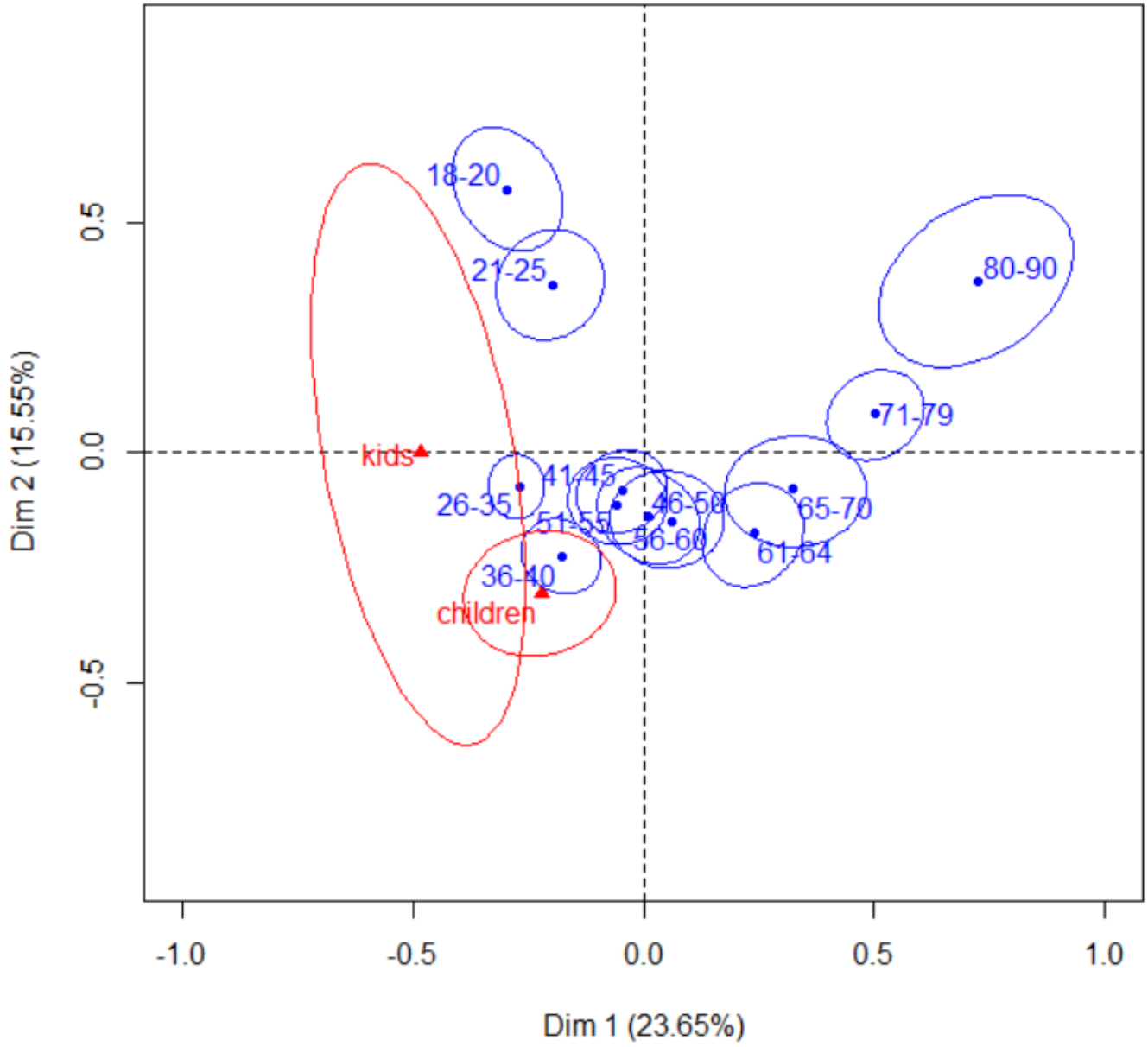


LexHCca function
Hierarchical Clustering of Words

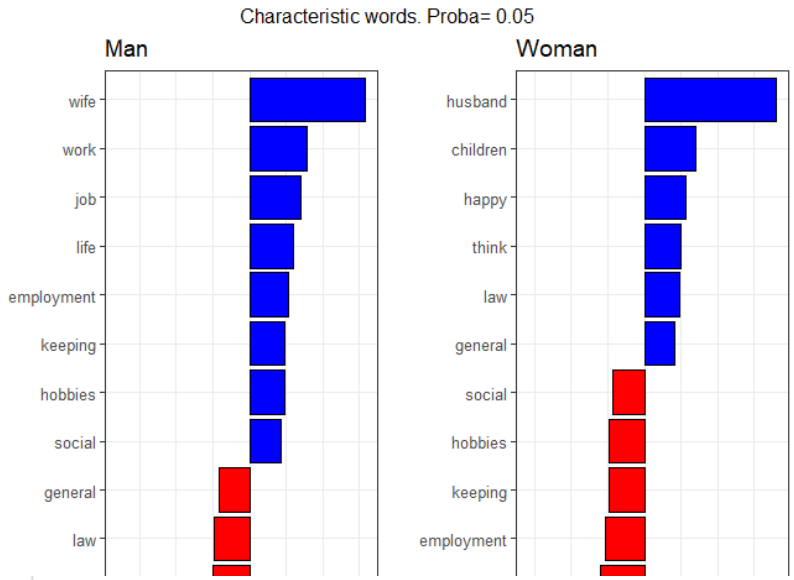
Words.C <- LexHCca(res.LexCA.2F, **cluster.CA="words"**, …

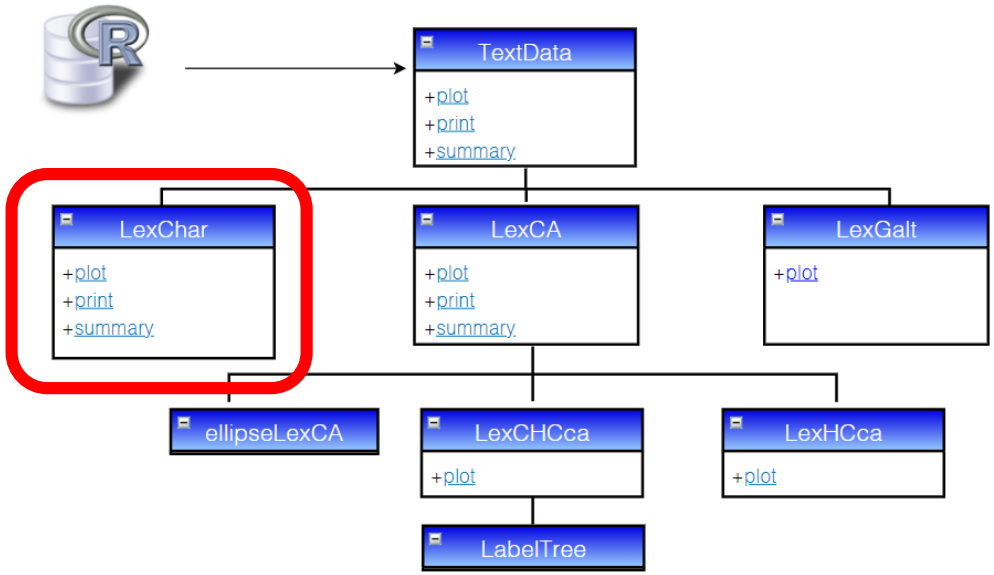



```
$description$desc.cluster.word$dist$cluster_4
        WORD CRITERION
1     career 0.9945546
2     social 0.9251893
3       sure 0.5659091
4    parents 0.5616955
5      sport 0.5234480
6     safety 0.5175754
7  education 0.5106157
8    leisure 0.4988161
9        job 0.4597118
10 everybody 0.4552173
11   future 0.4500564
12     kids 0.3450868
13     nice 0.3322055
14     work 0.3178161
15  freedom 0.2690517
16    right 0.2605103
17    house 0.2530546
```


18-20
21-25
80-90
71-79
kids
41-45
26-35
46-50
51-55
65-70
56-60
61-64
36-40
children
0.5
0.0
-0.5
Dim 2 (15.55%)
-1.0
-0.5
0.0
0.5
1.0
Dim 1 (23.65%)

Another analysis not presented here

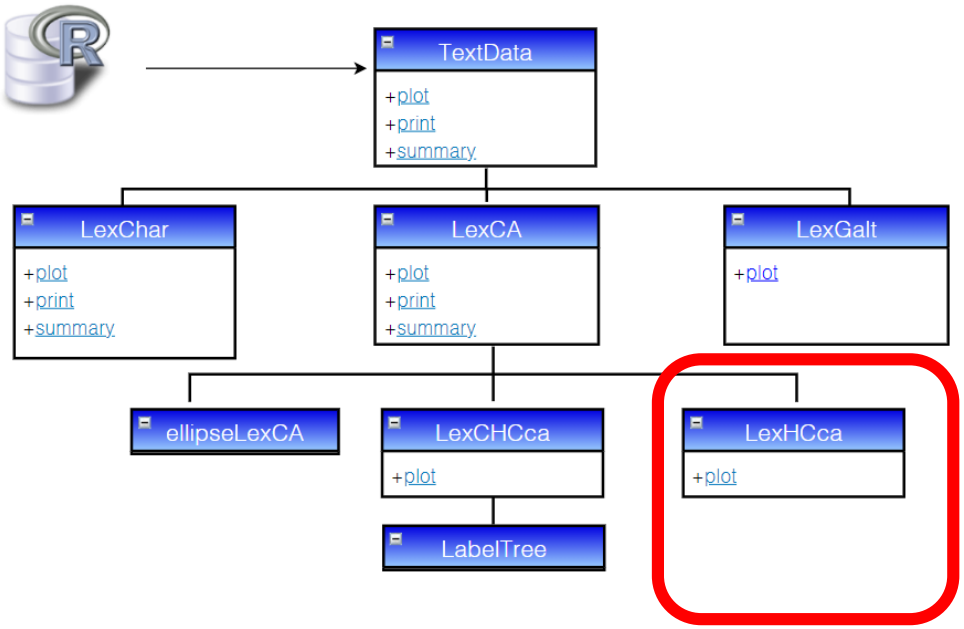




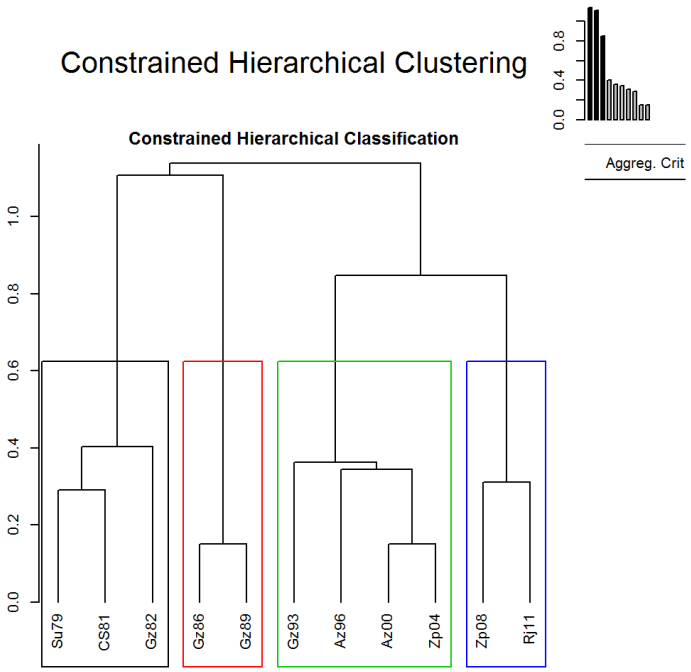
$Woman

	DOCUMENT	CRITERION	TEXT
1	4	0	health. happiness, money, family
2	5	0	to be happy. healthy, have enough to eat, enough money to live on
3	7	0	health. happiness
4	8	0	to be healthy.just to live long enough to see the children grow up, I do not think there is a lot t
5	9	0	health,. keeping going, family, going out, shopping, visiting
6	10	0	husband. new baby grand daughter, life in general,
7	12	0	good health. happiness, togetherness,
8	13	0	family. friends, pets,
9	15	0	my family really. health, walking
10	16	0	my children.my husband, my family and relations, health and wellbeing of family

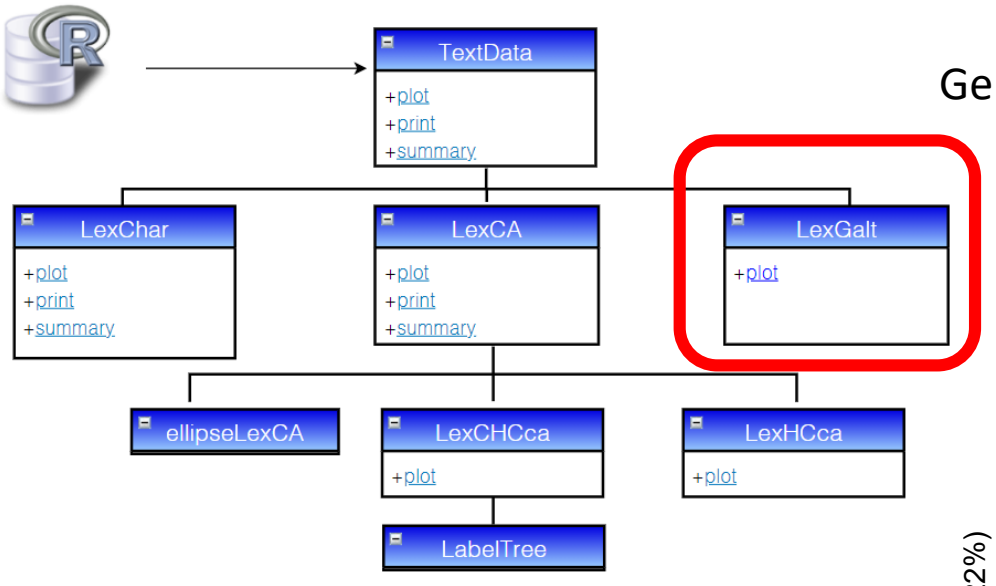
Another analysis not presented here



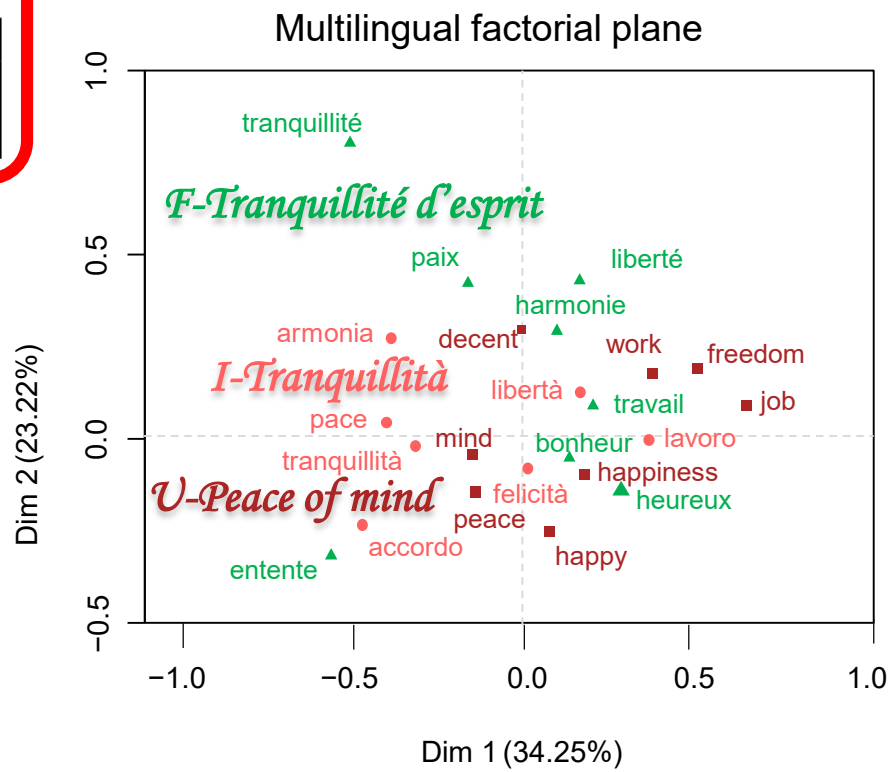
Chronogically Constrained Cluster



Another analysis not presented here



Generalised Aggregate Lexical Table (LexGalt)



5 May 2021

Las Médulas. Spanish Roman over ground gold mine

Thank you very much

Authors:

Ramón Alvarez-Esteban (2) (Maintainer)

Mónica Bécue-Bertaut (1)

Josep-Anton Sánchez-Espigares (1)

Belchin Kostov (1)

(1) UPC Universitat Politècnica de Catalunya / Spain
(2) University of Leon / Spain. ramon.alvarez@unileon.es

ramon.alvarez@unileon.es