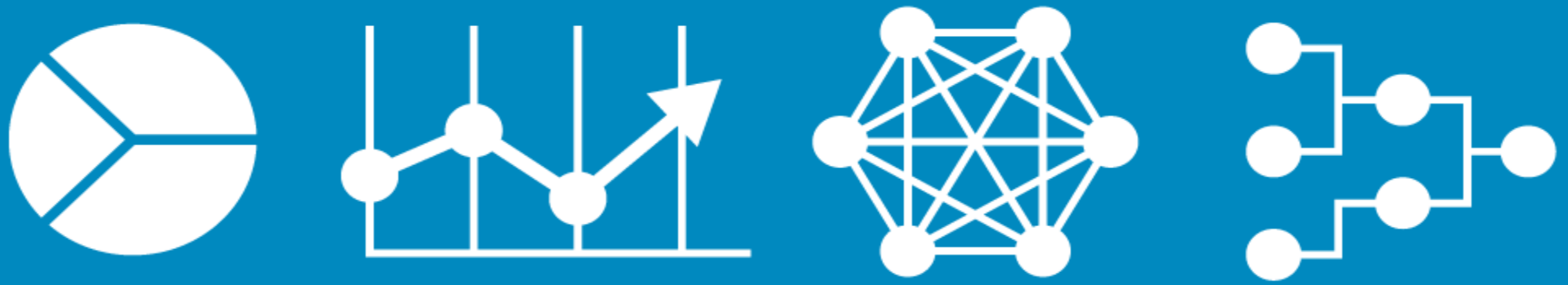
ctrl F Agency



Astuces en text-mining : removeSparseTerms

Text-mining Basics : Tips and Tricks R(S)

Par Thibaut LOMBARD

Décembre 2016

ctrl F Agency

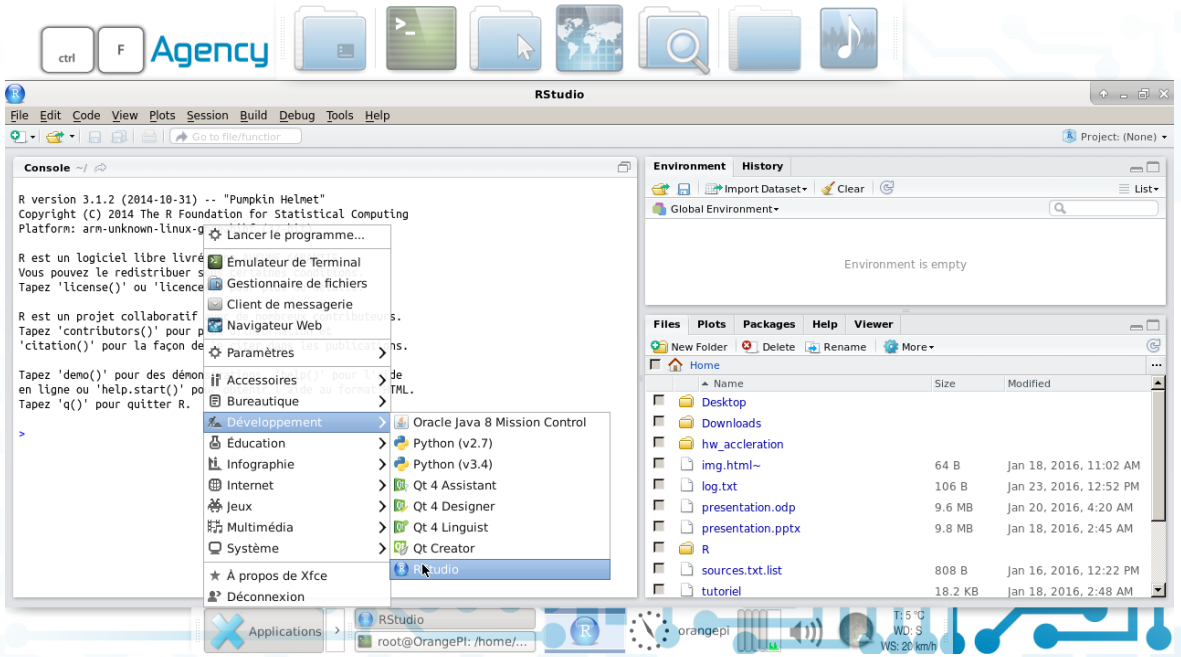
- Founder of ctrl+f agency (and all stuff related)
- Web/mobile Dev,
- Data science,
- Research
- Teaching

Age : 30 Years Old

ctrl F Agency

Pour commencer

- Ctrl+F agency OS : l'ordinateur le moins onéreux au monde
- Installez le thème ctrl+f agency sur OrangePI
- OrangePi et Ubuntu : tutorial
- Build/Installez Rstudio sur OrangePI



ctrl F Agency

Sparse.r

```
packages <- function(paquets)
{
        new.paquets <- paquets[!(paquets %in% installed.packages()[, "Package"])] if
        (length(new.paquets))
        install.packages(new.paquets, dependencies = TRUE, repos='http://cran.rstudio.com/')
        sapply(paquets, require, character.only = TRUE)
}
packages(c("NLP", "tm", "NMF"))
```

ctrl F Agency

Rscript sparse.r

```
monTexte <- c("Bâchez la queue du wagon-taxi avec les pyjamas du fakir.",
              "la matrice du wagon-taxi",
              "le fakir est dans la matrice")
monCorpus <- Corpus(VectorSource(monTexte))
maTdm <- DocumentTermMatrix(monCorpus, control = list(minWordLength = 1))
```

ctrl F Agency

inspect(maTdm)

```
[1] "******************************************************"
[1] "Inspection de la matrice Term document maTdm avec as.matrix()"
[1] "******************************************************"
<<DocumentTermMatrix (documents: 3, terms: 11)>>
Non-/sparse entries: 13/20
Sparsity           : 61%
Maximal term length: 10
Weighting          : term frequency (tf)
    Terms
Docs avec bâchez dans est fakir fakir. les matrice pyjamas queue wagon-taxi
   1    1       1    0   0     0      1   1       0       1     1          1
   2    0       0    0   0     0      0   0       1       0     0          1
   3    0       0    1   1     1      0   0       1       0     0          0
```

ctrl F Agency

Le calcul de sparsity (parcimonie) 50%

```
c <- removeSparseTerms(maTdm, 0.5)

<<DocumentTermMatrix (documents: 3, terms: 2)>>
Non-/sparse entries: 4/2
Sparsity           : 33%
Maximal term length: 10
Weighting          : term frequency (tf)

    Terms
Docs matrice wagon-taxi
   1       0          1
   2       1          1
   3       1          0
```

ctrl F Agency

Le calcul de sparsity (parcimonie) 90%

```
d <- removeSparseTerms(maTdm, 0.9)

<<DocumentTermMatrix (documents: 3, terms: 11)>>
Non-/sparse entries: 13/20
Sparsity           : 61%
Maximal term length: 10
Weighting          : term frequency (tf)

    Terms
Docs avec bâchez dans est fakir fakir. les matrice pyjamas queue wagon-taxi
   1    1      1    0   0     0      1   1       0       1     1          1
   2    0      0    0   0     0      0   0       1       0     0          1
   3    0      0    1   1     1      0   0       1       0     0          0
```

ctrl F Agency

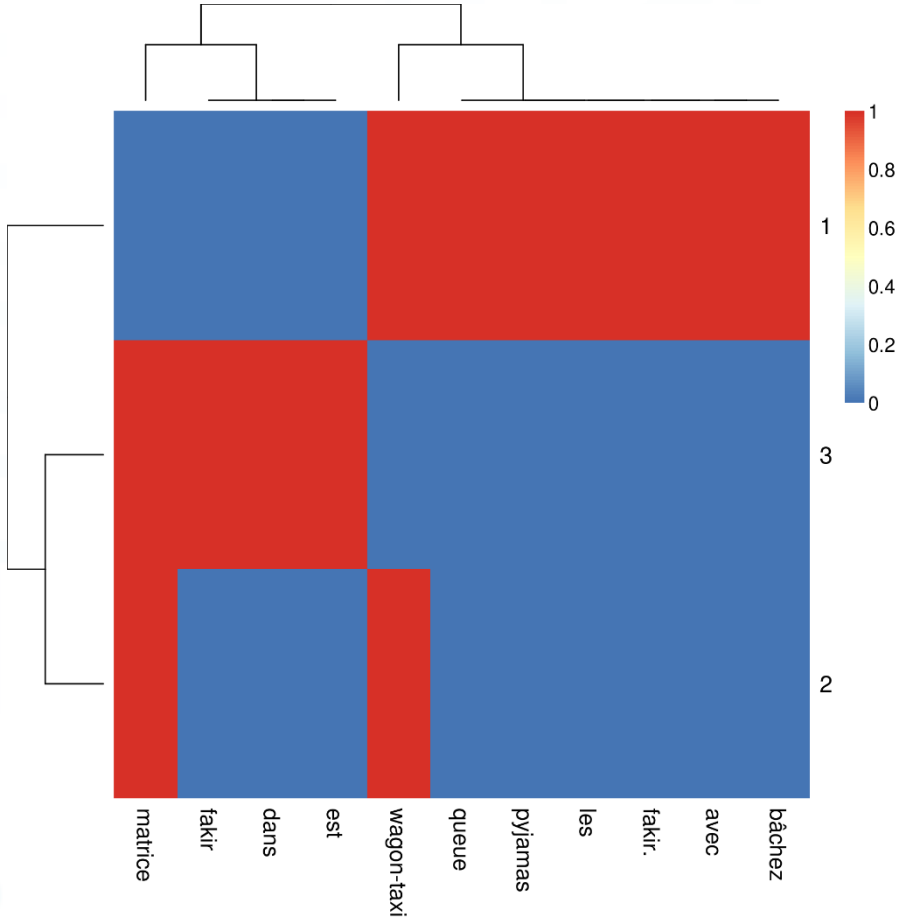
Création des annotated Heatmap

```
b <- as.matrix(maTdm)
ch <- as.matrix(removeSparseTerms(maTdm, 0.5))
dh <- as.matrix(removeSparseTerms(maTdm, 0.9))

aheatmap(b, filename = "sparse-heatmap-matrix.png")
aheatmap(ch, filename = "sparse-heatmap-50percent.png")
aheatmap(dh, filename = "sparse-heatmap-99percent.png")
```

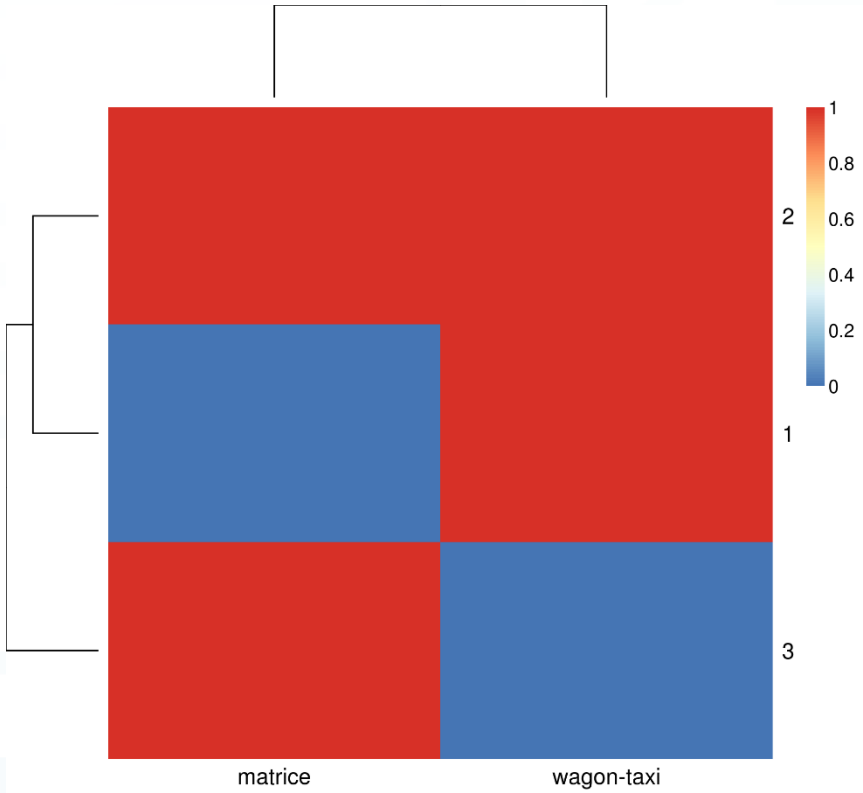
ctrl F Agency

Annotated Heatmap (matrice maTdm)



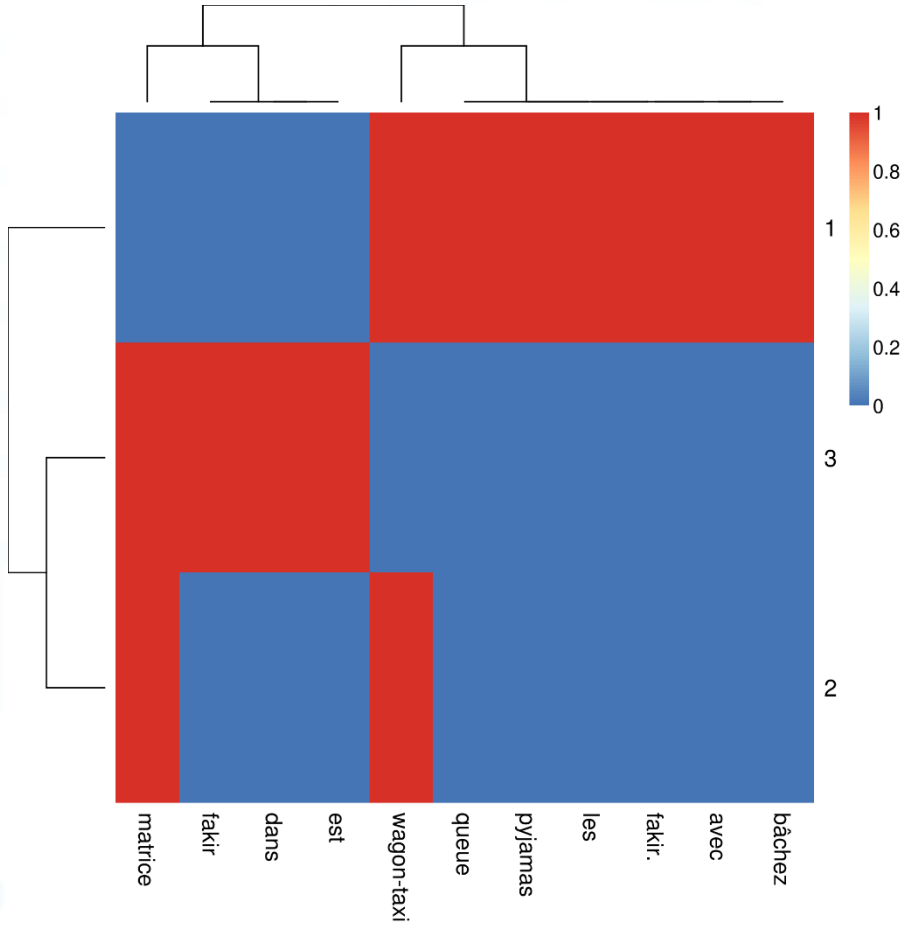
ctrl F Agency

Annotated Heatmap Parcimonie à 50%



ctrl F Agency

Annotated Heatmap Parcimonie à 90%



Création d'une distance matrix

```
distMatrix <- dist(t(d), method="euclidian")
distMatrix
```

```
          avec   bâchez     dans      est    fakir   fakir.      les
bâchez    0.000000
dans      1.414214 1.414214
est       1.414214 1.414214 0.000000
fakir     1.414214 1.414214 0.000000 0.000000
fakir.    0.000000 0.000000 1.414214 1.414214 1.414214
les       0.000000 0.000000 1.414214 1.414214 1.414214 0.000000
matrice   1.732051 1.732051 1.000000 1.000000 1.000000 1.732051 1.732051
pyjamas   0.000000 0.000000 1.414214 1.414214 1.414214 0.000000 0.000000
queue     0.000000 0.000000 1.414214 1.414214 1.414214 0.000000 0.000000
wagon-taxi 1.000000 1.000000 1.732051 1.732051 1.732051 1.000000 1.000000
           matrice  pyjamas    queue
pyjamas   1.732051
queue     1.732051 0.000000
wagon-taxi 1.414214 1.000000 1.000000
```

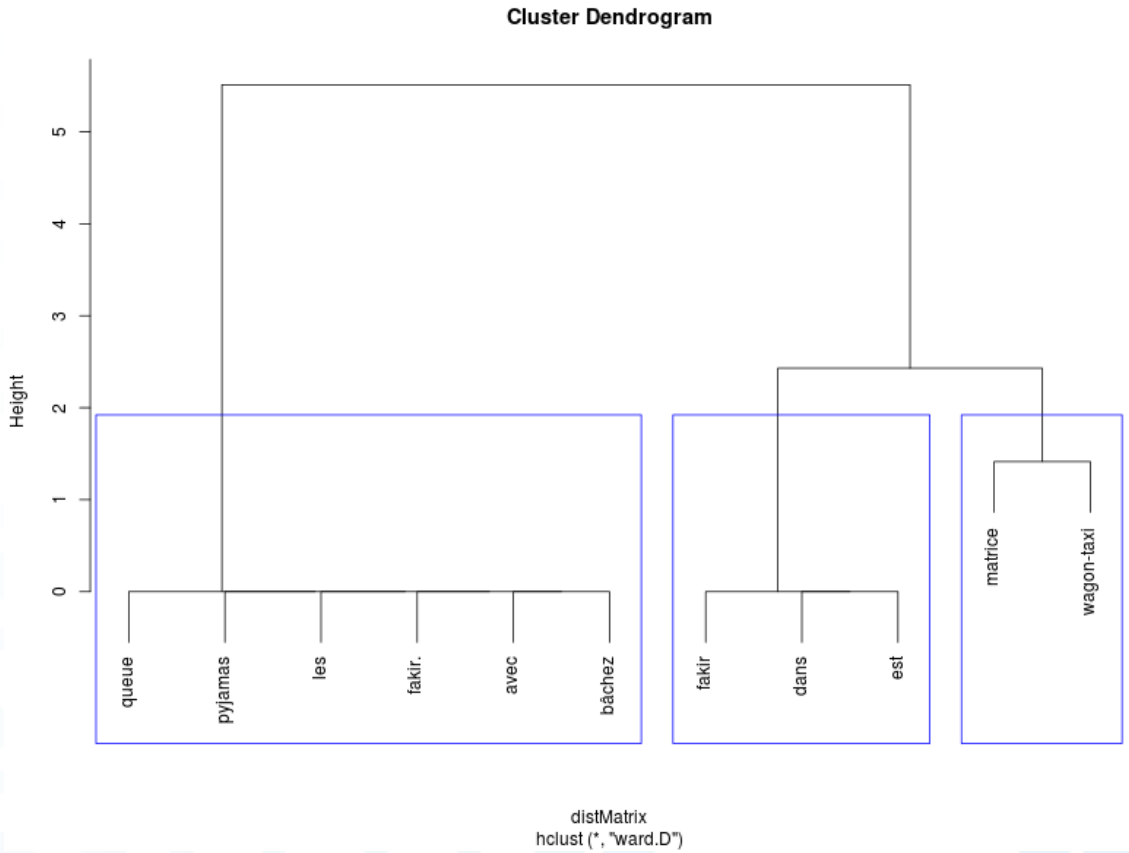
ctrl F Agency

Création d'un graphique de clutering (méthode Ward.D)

```
fit <- hclust(distMatrix, method = "ward.D")
png(filename="sparse-clust.png",width=800,height=600)
plot(fit)
rect.hclust(fit, k = 3,border="blue")
```

ctrl F Agency

Résultat du clustering



ctrl F Agency

Q/A