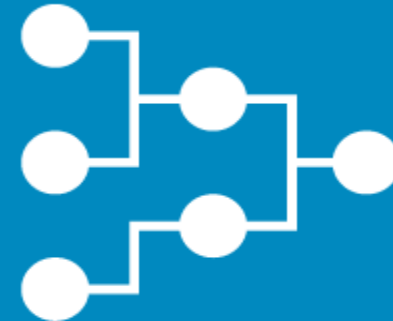




Agency



Astuces en text-mining : Organiser ses barcharts (diagrammes en barres)

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

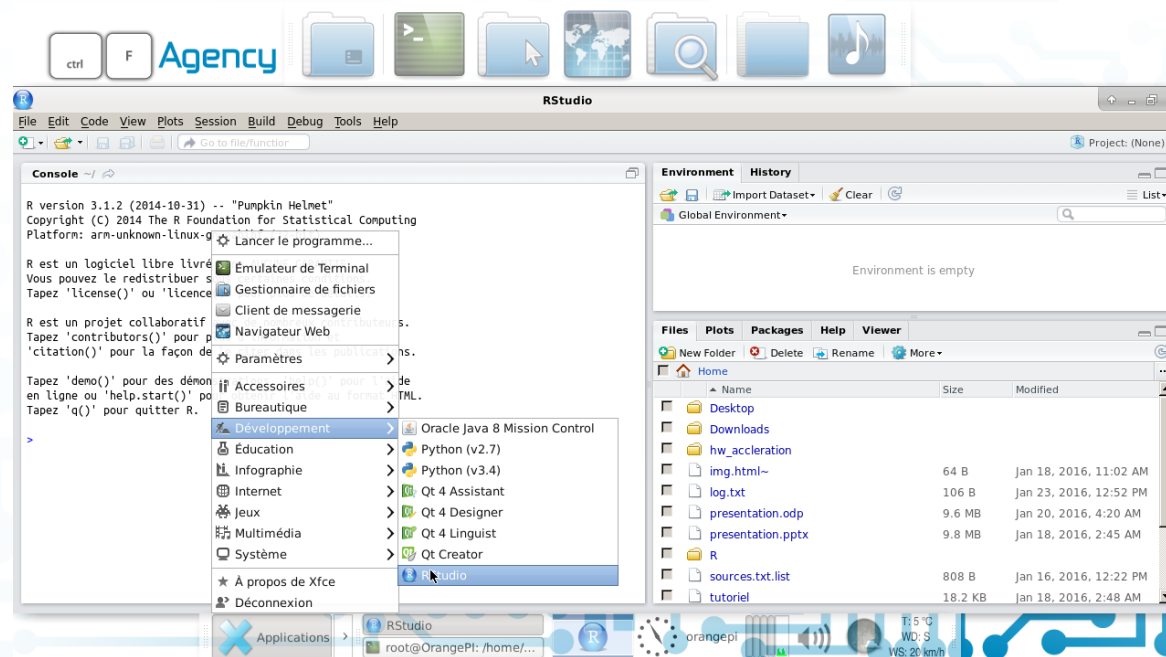
ctrl

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](#)



freqbarchart.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/')
  supply(paquets, require, character.only = TRUE)
}
packages(c("NLP", "tm", "ggplot2", "ggthemes", "dplyr"))
```

Rscript freqbarchart.r

```
# Ajoutons un Pangramme
# info : https://fr.wikipedia.org/wiki/Pangramme
monTexte <- c("Bâchez la queue du wagon-taxi avec les pyjamas du fakir.",
              "la matrice du wagon-taxi",
              "le fakir est dans la matrice",
              "le taxi fait le pyjamas sous la bâche",
              "le fakir est dans le wagon",
              "le sparse est conduit par le fakir",
              "vous êtes avec la matrice de pyjamas",
              "le wagon-taxi de pyjamas apprennent le machine learning")
monCorpus <- Corpus(VectorSource(monTexte))
tdm <- TermDocumentMatrix(monCorpus, control = list(minWordLength = 1))

# La somme des lignes en tant que matrices
tf <- rowSums(as.matrix(tdm))
# Sélectionne les fréquences >= 2
print('tf.2')
tf.2 <- subset(tf, tf>=2)
tf.2

# Créé un tableau avec comme nom de colonne term et freq
d <- data.frame(term = names(tf.2), freq = tf.2)
```

Simple diagramme en barres

```
ggplot(d, aes(x = term, y = freq)) + geom_bar(stat = "identity", color="blue", fill="steelblue") +  
ggtitle("Mots avec Fréquence +2") + labs(y='term',x='freq') + coord_flip() + theme_few()
```

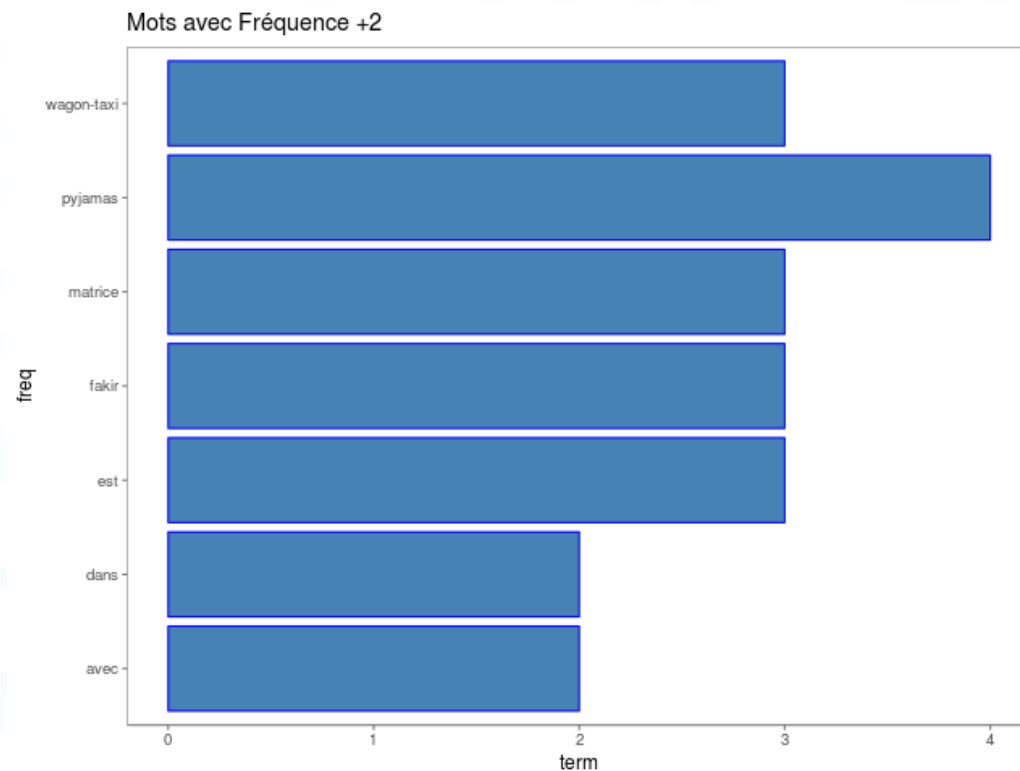


Diagramme en barres organisé par fréquence (factor)

```
d$term <- factor(d$term, levels = d$term[order(d$freq)])  
ggplot(d, aes(x = term, y = freq)) + geom_bar(stat = "identity", color="blue", fill="steelblue") +  
ggtitle("Mots avec Fréquence +2") + labs(y='term',x='freq') + coord_flip() + theme_few()
```

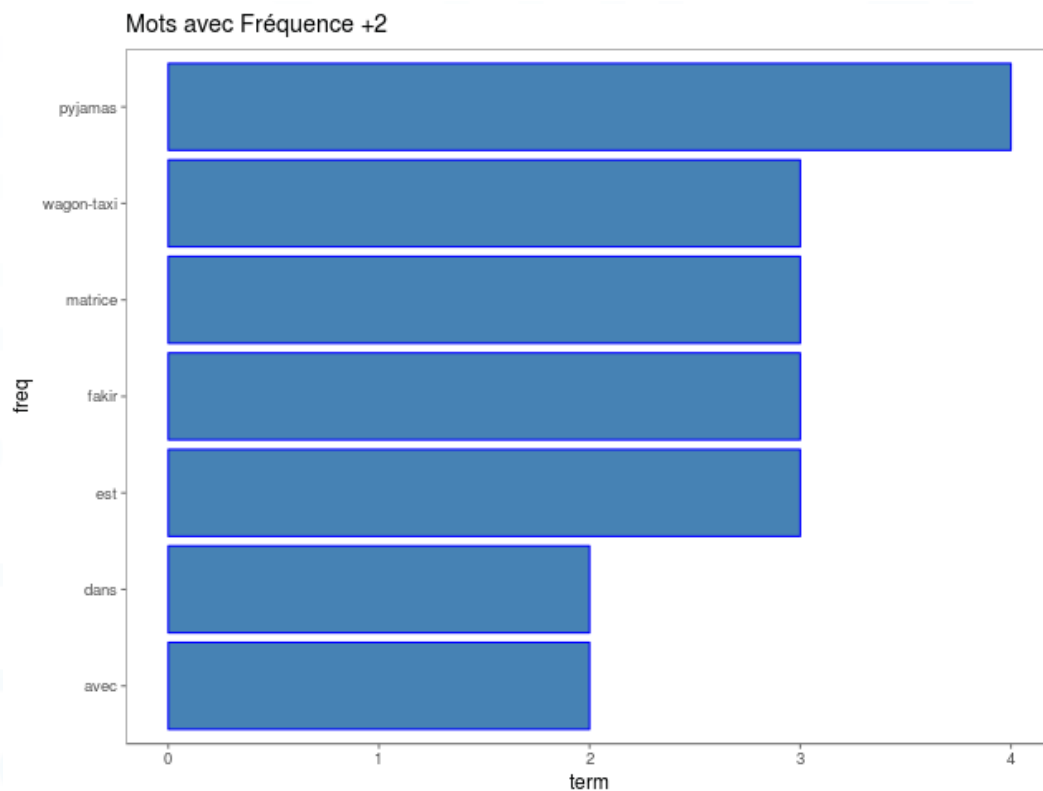


Diagramme en barres organisé par fréquence (inversé)

```
d$term <- factor(d$term, levels = d$term[rev(order(d$freq))])  
ggplot(d, aes(x = term, y = freq)) + geom_bar(stat = "identity", color="blue", fill="steelblue") +  
ggtitle("Mots avec Fréquence +2") + labs(y='term',x='freq') + coord_flip() + theme_few()
```

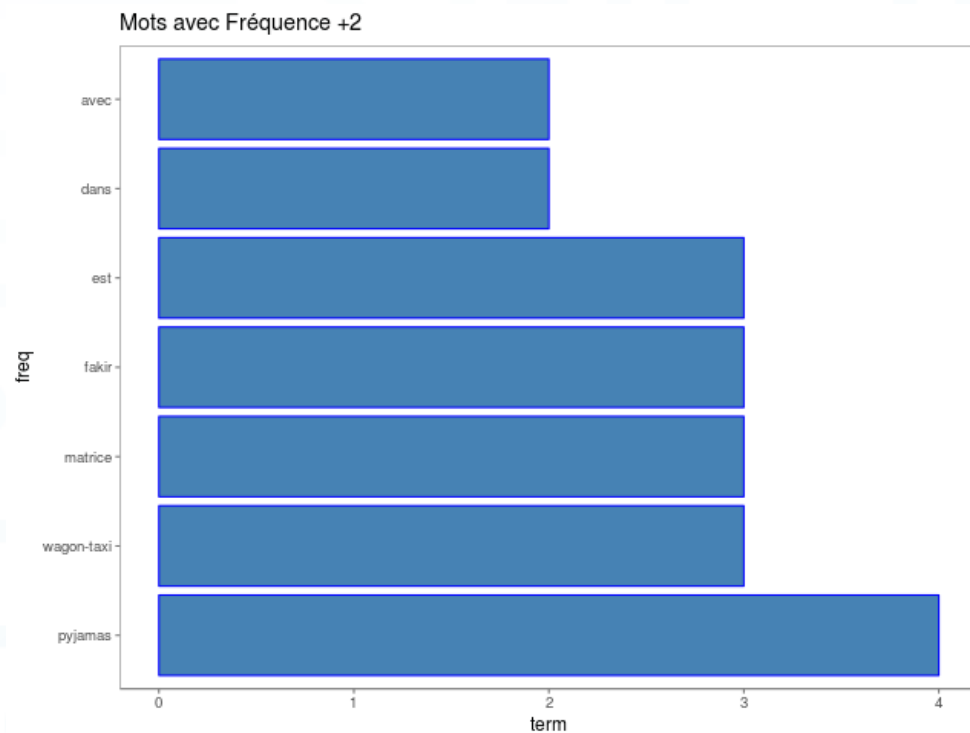


Diagramme en barres organisé par fréquence (within)

```
d2 <- within(d , term <- ordered(term, levels = rev(sort(unique(term)))))  
ggplot(d2, aes(x = term, y = freq)) + geom_bar(stat = "identity", color="blue", fill="steelblue") +  
ggtitle("Mots avec Fréquence +2") + labs(y='term',x='freq') + coord_flip() + scale_y_reverse() + theme_few()
```

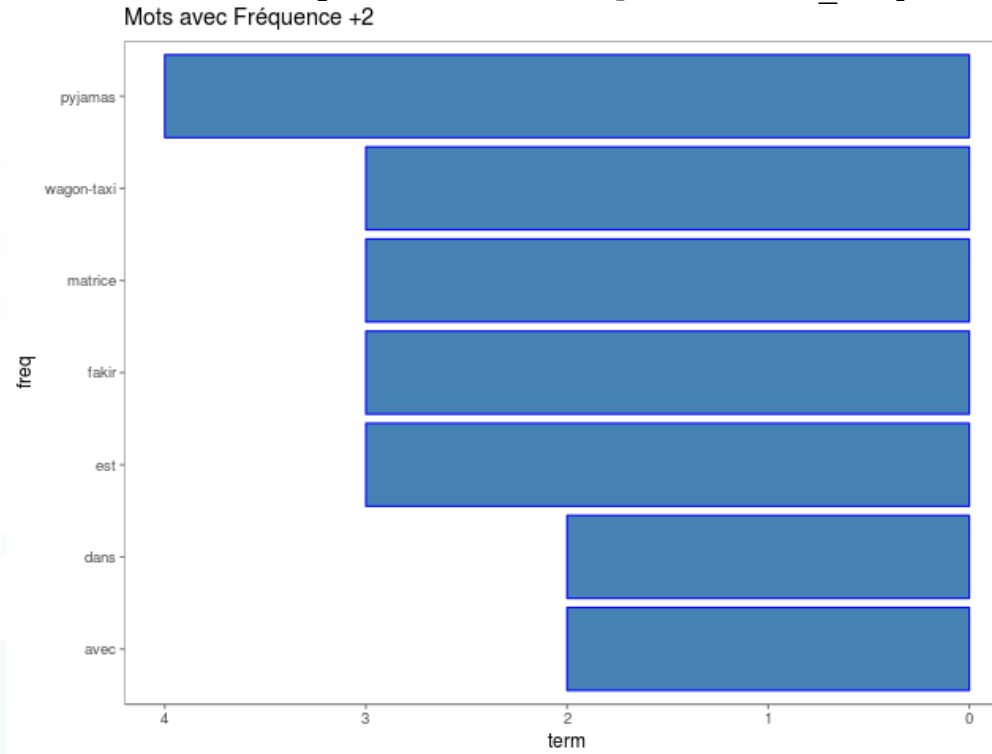
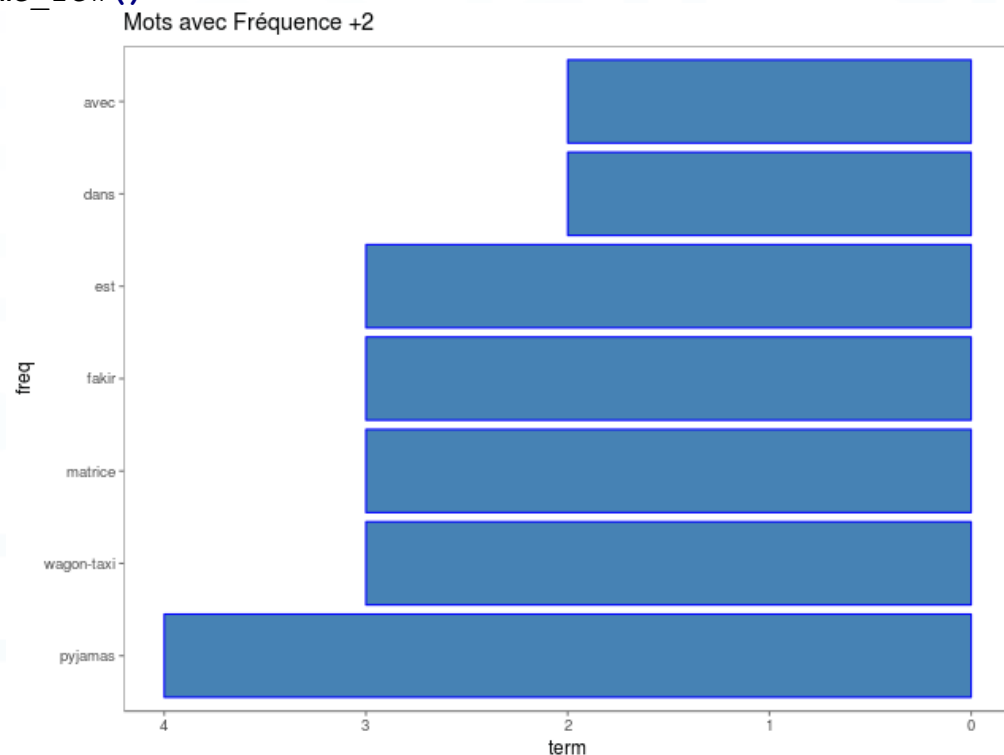


Diagramme en barres organisé par fréquences et termes (dplyr)

```
ggplot(d, aes(x = reorder(term, -freq), y = freq)) + geom_bar(stat = "identity", color="blue",  
fill="steelblue") + ggtitle("Mots avec Fréquence +2") + labs(y='term',x='freq') + coord_flip() +  
scale_y_reverse() + theme_few()
```



Liens connexes

- [GGplot2 – Barplot, guide de démarrage rapide](#)
- [Introduction à dplyr](#)
- [Documentation officielle ggplot2 – qplot](#)
- [How do i re-arrange ? Ordering a plot revisited](#)

ctrl

F

Agency

Q/A