

```
# script à utiliser dans R
# repo : https://github.com/Fatma1992/Nuage-de-Mots
# -*- coding: utf-8 -*-
# Installation des packages

install.packages ("NLP")
install.packages ("tm") #Pour le text mining
install.packages ("SnowballC") #Pour le text stemming
install.packages ("RColorBrewer") #Générateur Word-cloud
install.packages ("wordcloud") #Palettes de couleurs

# Chargement des packages:

library(NLP)
library(tm)
library(SnowballC)
library(RColorBrewer)
library(wordcloud)

# lire un fichier CSV
data <- read.csv("C:/Users/LENOVO PC/Desktop/corpus/corpus.txt", header = FALSE,
  sep = ",", quote= ",", fill = TRUE, encoding = "utf-8", stringsAsFactors =
FALSE,
  strip.white= TRUE, blank.lines.skip = TRUE, row.names = NULL, na.strings =
"NA")

#data <- readLines () #Optionnel

# Création d'un Corpus

data.corpus <- Corpus(VectorSource(data))

# La consultation du document
inspect (data.corpus)

# Réduire la taille de corpus en effectuant des transformations

toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
#data.corpus <- tm_map(data.corpus, toSpace, "http")

# Convertir le texte en minuscule
data.corpus <- tm_map(data.corpus, content_transformer(tolower))

# Supprimer les nombres
data.corpus <- tm_map(data.corpus, removeNumbers)

# Supprimer les ponctuations
#data.corpus <- tm_map(data.corpus, removePunctuation)

# Supprimer les mots vides français
data.corpus <- tm_map(data.corpus, removeWords, stopwords("french"))
```

```

# Supprimer votre propre liste de mots non désirés
data.corpus <- tm_map(data.corpus, removeWords,
c("laquel","http","chaque","autre","tôt","autres",
"tout","quand","http","etc","no
n","surtout","déjà",
"donc","où"))

#data.corpusPTD <- tm_map(data.corpus, PlainTextDocument)

# Supprimer les espaces vides supplémentaires
data.corpus<- tm_map(data.corpus, stripWhitespace)

# Text stemming
#data.corpus <- tm_map(data.corpus, stemDocument, language = "french")

# Construire la matrice des mots

data.tdm <- TermDocumentMatrix(data.corpus)
#data.tdm <- removeSparseTerms(data.tdm,0.004)

# Tranformer ce tdm en matrice
data.m <- as.matrix(data.tdm, byrow=TRUE)

# Classer en ordre décroissant les termes
data.v <- sort(rowSums(data.m), decreasing=TRUE)

# Calculer la fréquence de chaque mot
data.d <- data.frame(word = names(data.v), freq=data.v)

# Choisir combien de mots (optionnel)
head(data.d, 50)

# Création de Wordcloud (Nuage de mots)

# Générer une palette de 5 couleurs
pal <- rev (brewer.pal(200, "RdYlBu"))

# Mettre le fond de l'image grise
par(bg = "grey")

#Générer le nuage de mots
wordcloud(words = data.d$word, min.freq = 1,max.words=2000, random.order=TRUE,
rot.per=0.2,
          colors=brewer.pal(8, "Dark2"))

# words : les mots à dessiner

```

```
# min.freq : les mots avec une fréquence en dessous de min.freq ne seront pas
illustrés.
# max.words : nombre maximum de mots à dessiner.
# random.order : dessine les mots dans un ordre aléatoire.
    # Si false, ils seront dessinés par ordre décroissant de la
fréquence.
# rot.per : la proportion de mots verticaux sur le graphe
# colors : couleurs des mots du moins au plus fréquent.
    # Utiliser par exemple, colors ="black" pour une couleur unique.
```