UNIVERSITÉ CÔTE D'AZUR

ESTIA INSTITUTE OF TECHNOLOGY

DATUM ACADEMY

BIG DATA & DATA SCIENCE

USE CASE 3 : Etude sur la portée des Tweets publiés en rapport avec la propagation du Coronavirus aux Etats-Unis

Co-financé par :

FUN FRANCE UNIVERSITÉ NUMÉRIQUE

THEME

OPCO Atlas

CFA ÉPURE MÉDITERRANÉE

Use cases réalisés par les masters :

MIAGE RÉSEAU DES MIAGE DE FRANCE Nice

MBDS

BIMAR

Sommaire

- Présentation du use case
- Chargement des librairies
- Authentification Twitter (maintenant appelé X) depuis R
- Récupération des tweets
- Preprocessing des données
- Tokenisation
- Visualisation des données : wordcloud

1. Présentation du use case

Ce use case porte sur l'analyse de texte, NLP (natural language processing) afin de visualiser les mots qui reviennent le plus dans les tweets liés au Covid.

La suite de ce use case est à réaliser avec R sur RStudio.

2. Chargement des librairies

Installer et charger les librairies suivantes :

```
install.packages("magrittr")
install.packages("rtweet")
install.packages("tidyverse")
library(tidyverse)
library(magrittr)
library(rtweet)
```

3. Authentification Twitter (maintenant appelé X) depuis R

Dans R, lancer la commande suivante :

```
auth_setup_default()
```

Une page internet va s'ouvrir afin de renseigner son adresse mail et mot de passe Twitter (maintenant appelé X).

Attention, pour se connecter il faut déjà avoir créé un compte Twitter (maintenant appelé X).

Une fois la boîte de dialogue fermée, l'authentification sera faite depuis R.

4. Récupération des tweets

On récupère 3000 tweets émis aux USA, en anglais et on met le texte de ces tweets dans un dataframe « rt »

```
coords <- lookup_coords("usa")

rt_covid <- search_tweets("covid lang:en",n= 3000,include_rts = FALSE)

rt = rt_covid["full_text"]
```

5. Preprocessing des données

On nettoie les données pour les mettre au bon format :

```
#convert all text to lower case
ndf30 <- tolower(rt$full_text)
# Replace blank space
ndf30 <- gsub("rt", "", ndf30)
# Replace @UserName
ndf30 <- gsub("@\\w+", "", ndf30)
# Remove punctuation
ndf30 <- gsub("[[:punct:]]", "", ndf30)
# Remove links
ndf30 <- gsub("http\\w+", "", ndf30)
# Remove tabs
ndf30 <- gsub("[ |\t]{2,}", "", ndf30)
# Remove blank spaces at the beginning
ndf30 <- gsub("^ ", "", ndf30)
# Remove blank spaces at the end
ndf30 <- gsub(" $", "", ndf30)
#on enleve les emojis
ndf30 <- sapply(ndf30,function(row) iconv(row, "latin1","ASCII", sub=""))
```

6. Tokenisation

La Tokenisation est le processus de décomposition d’un texte donné en unités appelées jetons. Les jetons peuvent être des mots individuels, des parties de phrases ou même des phrases entières. Dans le processus de tokenisation, certains caractères comme les signes de ponctuation peuvent être écartés. Les jetons deviennent généralement l’entrée pour les processus comme l’analyse ou l’extraction de texte.

6. Tokenisation

```
install.packages("keras")
library(keras)
install.packages("tensorflow")
library(tensorflow)

num_words <- 10000
max_length <- 50
text_vectorization <- layer_text_vectorization(max_tokens = 10000, output_sequence_length = 50,)

text_vectorization %>% adapt(ndf30)

text_vectorization(matrix(ndf30[141:280]))
```

On tokenise nos données avec la fonction layer_text_vectorization de Keras.
Nous avons désormais une séparation par mots, et une valeur numérique unique est affecté à chaque mot.

7. Visualisation des données : wordcloud

On installe et charge les librairies nécessaires :

```
install.packages("tm")
install.packages("wordcloud")
install.packages("wordcloud2")
library("tm")
library("RColorBrewer")
library("wordcloud2")
```

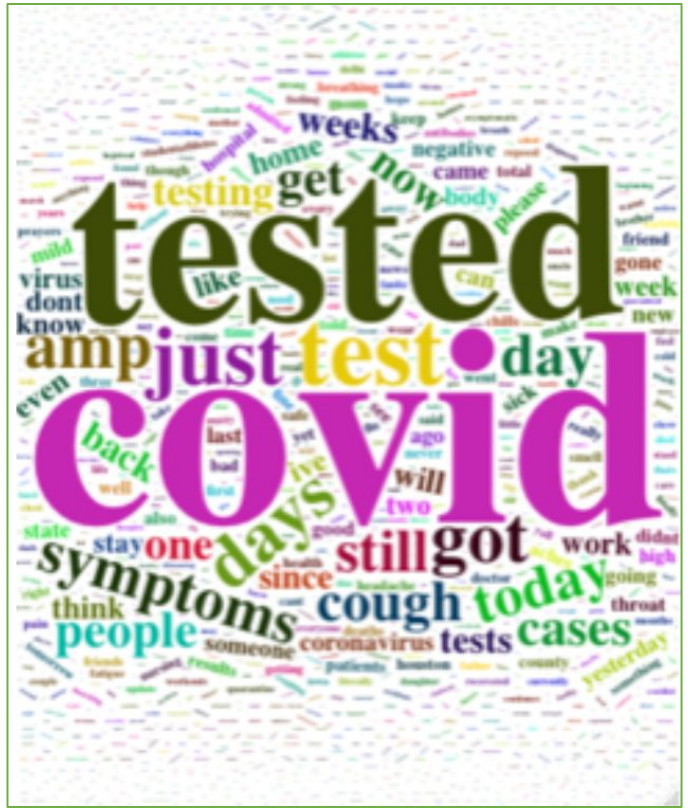
7. Visualisation des données : wordcloud

On crée une matrice des mots fréquents qu'on représente graphiquement en un nuage de mots. Plus les mots représentés sont gros, plus ils sont fréquents dans nos données.

```
text <- ndf30
docs <- Corpus(VectorSource(text))
#on enleve les stop words (les mots communs inutiles)
docs <- tm_map(docs, removeWords, stopwords("english"))
#construction de la matrice des mots fréquents
dtm <- TermDocumentMatrix(docs)
matrix <- as.matrix(dtm)
words <- sort(rowSums(matrix),decreasing=TRUE)
dfwc <- data.frame(word = names(words),freq=words)
#representation graphique du nuage de mots clé
set.seed(1234)
wordcloud2(data=dfwc, size = 1.2,shape = 'pentagon')
```

7. Visualisation des données : wordcloud

Nous constatons que les mots les plus fréquents sont : « **covid** », « **tested** », « **symptoms** », « **cough** », « **test** ».



Références

- [RapportHallaci] Rapport de projet tutoré en machine learning et deep learning, Chahineze HALLACI, 2020