Projet Chatbot Orange

Tom BLACHON Meyssa BEDDAR Matthieu SIMOES

02/05/2021

Présentation du projet

Nous disposons des réponses à un questionnaire sur les utilisateurs de Chatbots afin de mieux comprendre la vision qu'ils ont de cet outil. À partir de ces données, notre objectif est de réaliser un clusturing de ces individus, afin de déterminer des profils d'utilisateurs.

Installation des packages :

```
library(foreign)
library(tidyverse)
library(cluster)
library(factoextra)
library(clue)
library(WeightedCluster)
library(openxlsx)
library(knitr)
library(ggplot2)
```

Import des données :

```
setwd("D:/Documents/Master MEDAS - CNAM/Fouille de données/chatbot Orange")
data <- read.spss("BJ20818 Orange test Chatbot - Base SPSS/base bj20818 pour client.sav",
                  to.data.frame=TRUE)
```

Nettoyage des données

Afin de faire correspondre notre étude aux statistiques descriptives précédemment réalisées, nous commençons par supprimer certaines réponses de notre jeu de données :

- Les répondants n'ayant pas échangé avec un chatbot dans les 18 derniers mois.
- Les répondants ayant déjà utilisé l'option "Voix" d'un chatbot.

```
data_flt <- data %>% filter(q2 %in% c("Il y a moins de 6 mois", "Entre 6 et 12 mois",
                                      "Entre 12 et 18 mois"), q5 %in% c("Non", NA))
```

Pour réaliser nos clusters, nous avons choisi de ne sélectionner que certaines variables (questions) parmis nos données. En effet, le jeu de données brut contient 95 variables différentes. Si nous les conservions toutes, notre clustering ne serait pas pertinent car il y aurait trop de paramètres à prendre en compte, ce qui impacterait sur la qualité de notre modèle. Après discussion avec monsieur Thierry Curiale et étude des statistiques descriptives,nous avons donc sélectionné les 10 variables suivantes :

- Sexe du répondant (s1)
- Tranche d'âge du répondant (s2)
- Avec quel type de chatbot le répondant souhaiterait-il discuter ? (q7)

- Le répondant perçoit-il le chatbot comme une entitée vivante ? (q8)
- De quels attributs humains dispose le chatbot imaginé ? (q15a)
- Si le répondant imagine un corps, quel serait son genre ? (q16)
- Si le répondant imagine un visage, quel serait son genre ? (q17)
- Si le répondant imagine une voix, quelle serait son genre ? (q18)
- Quelle serait la personnalité du chatbot imaginé ? (q20)
- Comment le répondant aimerait-il communiquer avec son chatbot ? (q22a)

Nous avons donc choisi de nous concentrer sur l'étude de **la personnatilé et des attributs physiques des chatbots** imaginés par les répondants.

```
data_clean <- data.frame()
data_clean<- data_flt %>% select(s1, s2, q7, q8, q15a, q20, q22a, q17, q18, weight0)
```

Nous avons ensuite compilé les réponses des questions 16, 17 et 18, concernant le sexe du chatbot imaginé, en une seule variable "genre" :

```
data_clean$genre <- 0
data_clean$genre <- data_flt$q16
levels(data_clean$q18) <- c(levels(data_clean$q18), "Féminin", "Masculin", "Non genré",
                            "Non renseigné")
data_clean$q18[data_clean$q18 == "Féminine"] <- "Féminin"
data_clean$q18[data_clean$q18 == "Masculine"] <- "Masculin"
data_clean$q18[data_clean$q18 == "Non genrée"] <- "Non genré"
data_clean <- data_clean %>% mutate(genre = coalesce(genre, q17))
data_clean <- data_clean %>% mutate(genre = coalesce(genre, q18))
data_clean <- data_clean %>% replace_na(list(genre = "Non renseigné"))
data_clean <- data_clean %>% select(-c("q18", "q17"))
data_clean <- data_clean %>% relocate("weight0", .after = "genre")
```

Présentation du clustering

Choix de la méthode :

Pour réaliser un clustering et regrouper des ensembles d'individus selon des caractéristiques communes, plusieurs méthodes sont envisageables. Ici, notre jeu de données contient exclusivement des informations qualitatives, regrouppées en classes. Nous ne pouvons donc pas réaliser de clustering *k-means*, car nous ne pouvons pas calculer de moyenne sur ces données qualitatives. Nous avons donc choisi d'effectuer des clusters par *Partitionning Around Medoids*, également appellé "PAM", car cette méthode nous a semblé plus adaptée à la structure de nos données.

Présentation de la méthode PAM (Partitionning Around Medoids) :

Aussi appelé k-medoids, ce clustering vise à sélectionner des individus représentatifs, appelé "medoids". L'algorithme va ainsi sélectionner k individus représentatifs, puis va assigner les n individus de notre dataset à l'individu représentatif le plus proche. Nous obtiendrons ainsi k clusters, dont les points centraux sont les individus représentatifs.

Pour calculer la proximité entre chaque individu, l'algorithme calcule au préalable une matrice de distance entre les individus. Dans notre cas, comme nous disposons de données qualitatives, nous allons utiliser la dissimilarité de Gower pour estimer une distance entre nos objets.

Réalisation du clustering :

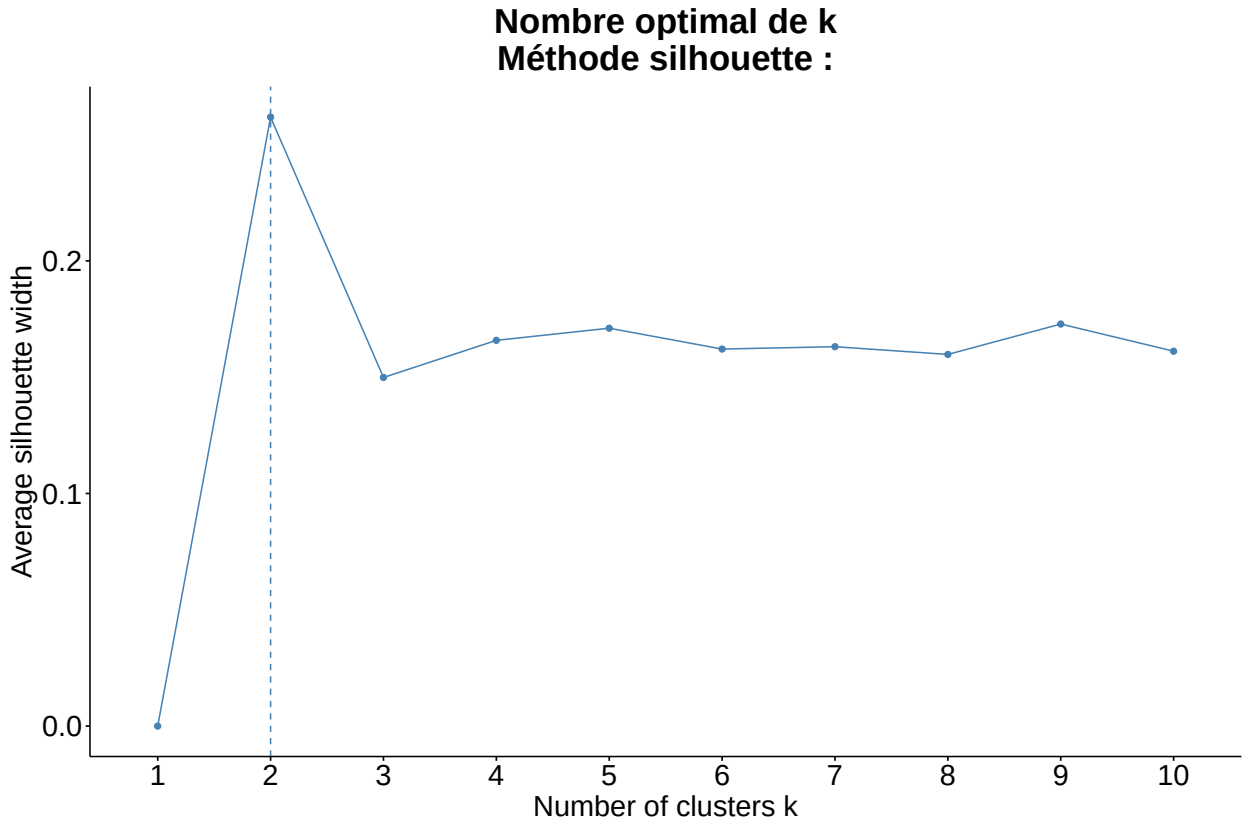
Matrice de distance : Nous commençons donc par réaliser une matrice de distance entre nos individus grâce à la dissimilarité de Gower. Pour celà, nous utilisons la fonction "daisy" :

```
dist<-daisy(data_clean[,1:8],metric = "Gower")
dist_matrix <- as.matrix(dist)
```

Nombre idéal de clusters : Avant de créer nos clusters, nous allons tout d'abbord essayer d'estimer le nombre idéal de clusters, c'est-à-dire le nombre k d'individus représentatifs.

Pour celà, nous avons utilisé la méthode silhouette :

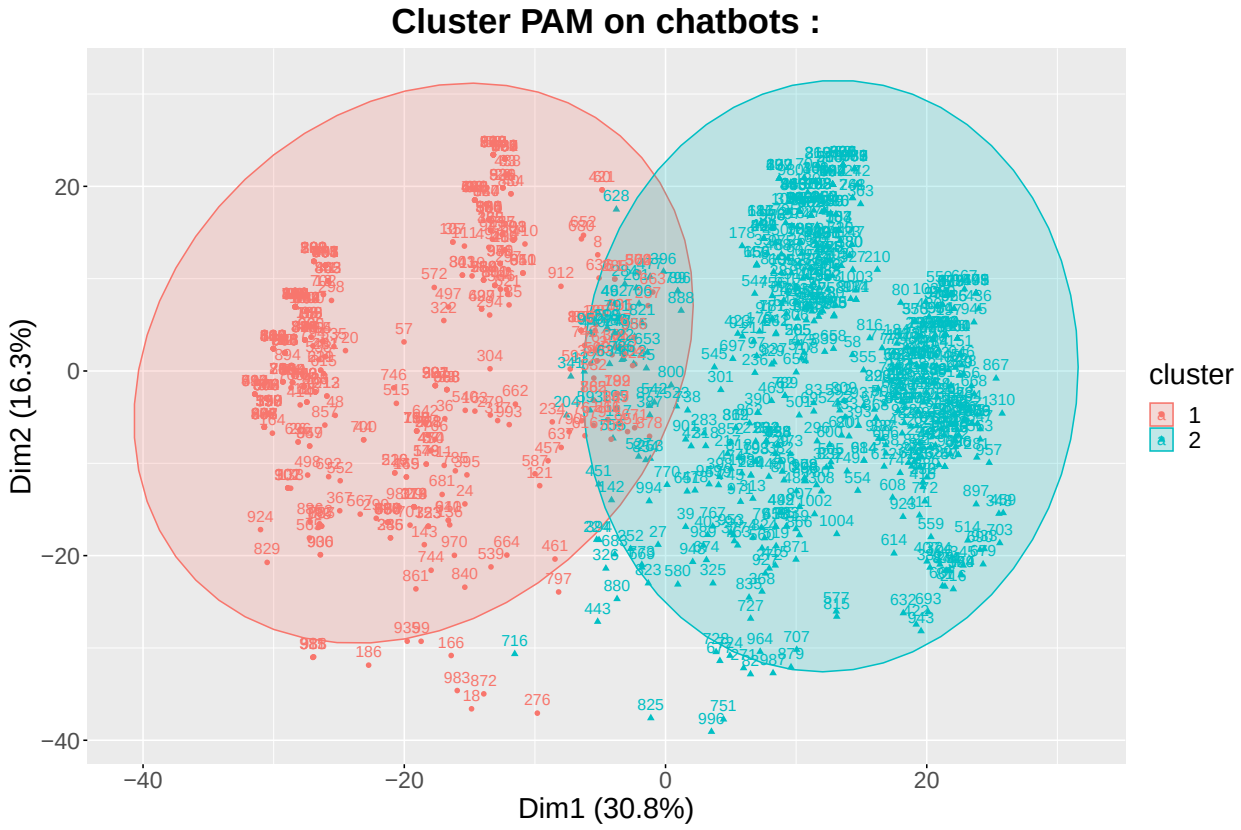
```
fviz_nbclust(dist_matrix, pam, method = "silhouette") +
  ggtitle("Nombre optimal de k\nMéthode silhouette :") +
  theme(text = element_text(size = 20), plot.title = element_text(face="bold",
                                                                  hjust = 0.5))
```



Ici, nous pouvons donc voir que notre nombre de clusters idéal est $k = 2$.

Clustering PAM : Nous avons ensuite utilisé la fonction "pam" afin de réaliser nos clusters. Nous appliquons cette fonction directement sur notre matrice de dissimilarité calculée plus tôt. De plus, nous indiquons le nombre de clusters souhaités : $k = 2$.

```
kmed <- pam(dist_matrix, 2)
fviz_cluster(kmed, data_clean, ellipse.type = "norm") +
  ggtitle("Cluster PAM on chatbots :") +
  theme(text = element_text(size = 20), plot.title = element_text(face="bold",
                                                                  hjust = 0.5))
```



Observations des résultats :

Isolation : Nous pouvons tout d'abbord constater que nos deux clusters sont relativement distincts, mais qu'ils s'intersectent en un point, si bien que les individus centraux pourraient appartenir à l'un ou l'autre des clusters.

```
kmed$isolation
```

```
##  1  2
## no no
## Levels: no L L*
```

Lorsque l'on teste l'isolation de nos clusters, nous pouvons en effet constater qu'ils ne sont ni de type L, ni de type L*. Nos clusters sont donc mal isolés. Ils ne respectent pas les caractéristiques permettant d'indiquer que les clusters sont isolés :

- L'isolation est de type L si : $\max_{j \in C} \ d(i,j) < \min_{h \notin C} \ d(i,h)$
- L'isolation est de type L* si : $\max_{i,\ j \in C} \ d(i,j) < \min_{l \in C,\ h \notin C} \ d(l,h)$

Dans notre cas, nous voyons que les points centraux peuvent appartenir aux deux clusters qui, par conséquent, ne sont pas isolés.

Analyse des clusters :

```
table(kmed$clustering)
```

```
##
##   1   2
## 397 607
```

```
rownames(kmed$medoids)
```

```
## [1] "794" "58"
```

Les codes ci-dessus nous informent sur la répartition de nos individus au sein des clusters. Ainsi, nous pouvons voir que celle-ci n'est pas équitable, car **le cluster n°1 (en rouge) est composé de 397 individus**, tandis que **le cluster n°2 (en bleu) est composé de 607 individus**.

Pour chaque cluster, rappellons que l'individu central est appelé "medoid" et se veut **représentatif** des autres individus du cluster. Ici, nous pouvons constater que les individus représentatifs sont :

- Le n°794 pour le cluster 1 (en rouge)
- Le n°58 pour le cluster 2 (en bleu)

Ainsi, observons les réponses données par ces deux répondants pour dégager les grandes tendances :

```
kable(rbind(data_clean[794,c(1, 2, 4:8)], data_clean[58,c(1, 2, 4:8)]),
      caption = "Medoids des clusters 1 et 2")
```

Table 1: Medoids des clusters 1 et 2

	s1	s2	q8	q15a	q20	q22a	genre
794	Un homme	18-24 ans	Non vivante (ex : pierre, peluche, fantôme...)	Aucun attribut humain	Neutre	Par l'écriture	Non renseigné
58	Un homme	35-49 ans	Vivante (ex : animale, humaine, végétale...)	Une voix	Neutre	Autre	Non genré

Nous pouvons voir que dans nos deux clusters, l'individu représentatif est un homme, imaginant le chatbot comme un conseiller clientèle, dont la personnalité serait neutre. Cependant, dans le premier cluster (n°1), les individus ont majoritairement tendance à percevoir le chatbot comme une entitée **"Non vivante"** et sans attributs humains, tandis que le second cluster (n°2) va le percevoir comme **"Vivant"**, avec une voix humaine.

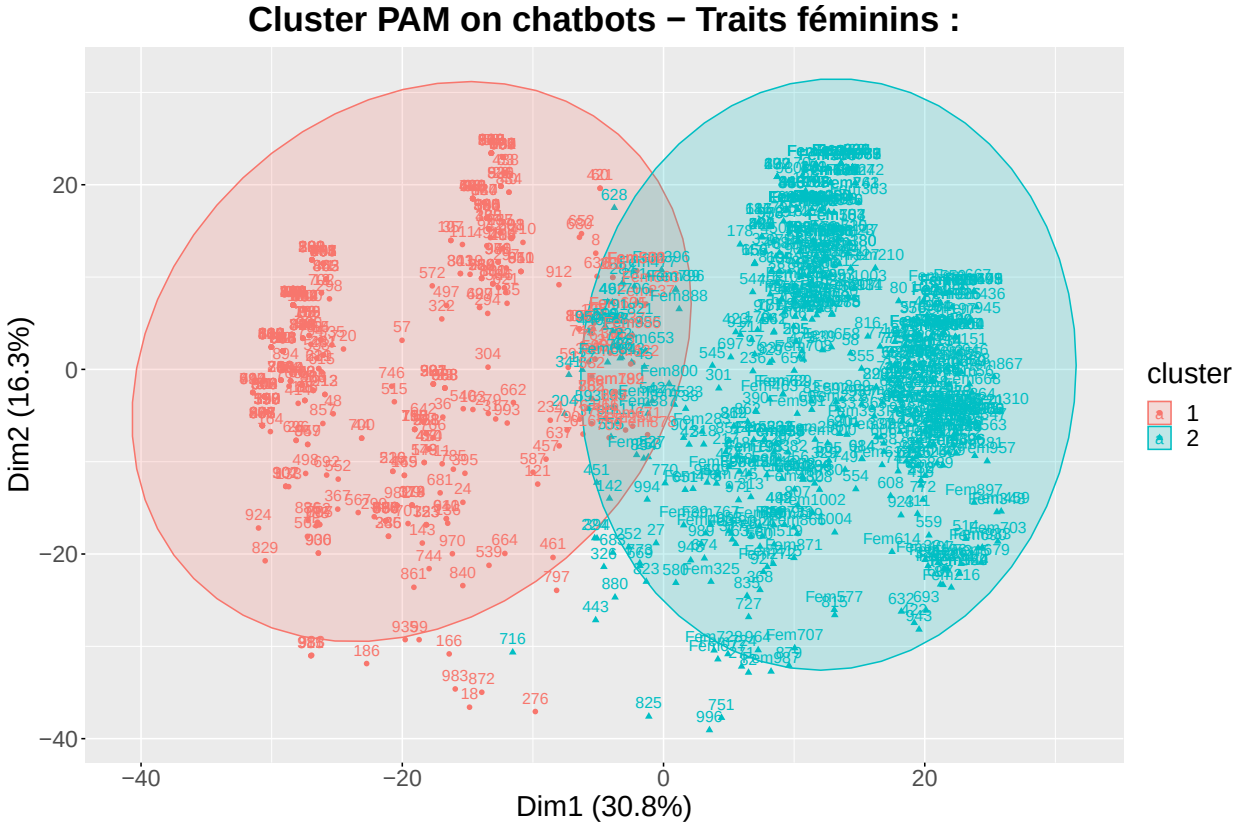
Cette distinction marque bien les divergences d'opinions entre nos deux groupes d'individus.

Perception féminine du chatbot : Malgré le constat énoncé plus haut, nous avons souhaité vérifier une hypothèse : l'étude menée par monsieur Thierry Curiale sur les répondants avait déjà mis en avant cette différence de percéption entre entitée "vivante" ou "non vivante". Cependant, celui-ci avait soumis l'idée que même certains répondants percevant le chatbot comme "non vivant" l'associaient instinctivement à **une entitée disposant de carractéristiques féminines**.

Nous avons donc entrepris de vérifier cette théorie, en réalisant un graphique mettant en avant les répondants attribuant des traits féminins au chatbot :

```
for(i in 1:nrow(data_clean)){
  if(data_clean$genre[i] == "Féminin"){
    rownames(data_clean)[rownames(data_clean) == i] <- paste0("Fem",
                                                              rownames(data_clean[i,]))
  }
}
dist<-daisy(data_clean[,1:8],metric = "Gower")
dist_matrix <- as.matrix(dist)
kmed_fem <- pam(dist_matrix, 2)
fviz_cluster(kmed_fem, data_clean, ellipse.type = "norm") +
  ggtitle("Cluster PAM on chatbots - Traits féminins :") +
  theme(text = element_text(size = 20), plot.title = element_text(face="bold",
```

```
                                                                hjust = 0.5))
```



Dans ce graphique, les individus associant des caractéristiques féminines au chatbot possèdent le préfixe “Fem”. Logiquement, nous pouvons constater que ceux-ci sont très largement majoritaires au sein du cluster n°2, percevant le chatbot comme une entitée “vivante”. Cependant, nous pouvons également remarquer qu’à la jointure des deux groupes, certains répondants rattachés au cluster n°1 attribuent également des caractéristiques féminines au chatbot.

Ces individus pourraient ainsi vérifier notre théorie : **Le point de jonction entre les répondants percevant le chatbot comme “non vivant” et ceux le percevant comme “vivant” serait donc l’attributions de traits féminin par une partie des individus.** Explicant au passage l’impossible indépendance de nos clusters.

Incorporation des poids : Enfin, afin de perfectionner la représentativité de notre cluster, nous avons souhaité incorporer les poids des individus à notre modèle. En effet, compte-tenu du fait que les répondants ne sont pas représentatifs de la population française, un rééquilibrage a été effectué en aval en attribuant un poids à chaque idividu.

```
round(sum(data_clean$weight0),0)
```

```
## [1] 1019
```

Nous pouvons d’ailleurs remarquer que ce rééquilibrage entraine une hausse de répondants “fictifs” à notre questionnaire. En effet, la somme des poids de chaque individu est égale à 1019. Donc pour 1004 répondants “réels”, nous obtenons 1019 répondants “fictifs”.

Nous avons donc utilisé une formule nous permettant de réaliser un clustering PAM en prenant en compte le poids des individus :

```
weighted_kmed <- wcKMedoids(dist_matrix, 2, weights = data_clean[,9], method = "PAM")
```

Cependant, nous ne sommes pas parvenus à afficher nos résultats dans un graphique, car la fonction "fviz_cluster" ne reconnait pas les objets de type "kmedoidslist" :

```
fviz_cluster(weighted_kmed, data_clean[,1:8], scale = T)
```

```
## Error in fviz_cluster(weighted_kmed, data_clean[, 1:8], scale = T): Can't handle an object of class l
```

Mais nous pouvons tout de même étudier ce nouveau clustering :

```
table(weighted_kmed$clustering)
```

```
##
##   1 691
## 571 433
```

Ici, nous disposons donc de deux clusters :

- Un premier cluster composé de 571 répondants et dont l'individut représentatif est le répondant n°1.
- Un second cluster composé de 433 répondants et dont l'individut représentatif est le répondant n°691.

Aussi, si nous nous intéressons à ces individus :

```
kable(rbind(data_clean[1,c(1, 2, 4:8)], data_clean[691,c(1, 2, 4:8)]),
      caption = "Medoids des clusters 1 et 2 (avec poids)")
```

Table 2: Medoids des clusters 1 et 2 (avec poids)

	s1	s2	q8	q15a	q20	q22a	genre
1	Un homme	50-64 ans	Non vivante (ex : pierre, peluche, fantôme...)	Aucun attribut humain	Neutre	Par l'écriture	Non renseigné
Fem691	Une femme	35-49 ans	Vivante (ex : animale, humaine, végétale...)	Une voix	Neutre	Par la parole	Féminin

Plusieurs changement avec notre clustering précédent sont alors observables :

- Nous retrouvons la dychotomie "Vivant"/"Non vivant", mais la tendance s'est inversé car il y a désormais plus de personnes percevant le chatbot comme "non vivant".
- La prise en compte du poids semble avoir entrainé une meilleure représentativité des femmes dans le questionnaire, car l'individut représentatif du second cluster est une femme.
- Cette fois, l'attrubution de caractéristiques féminines au chatbot est plus directement visible au sein du cluster percevant le chatbot comme une entitée "vivante".
