# Proyecto

Pedro Paiva, Rodolfo Miquilarena y Krystle Salazar

August 12, 2015

**Abstract**

En este proyecto trataremos un conjunto de posts de facebook de 5 usuarios, los cuales nos encargaremos de limpiar para sacarle el mayor provecho y buscaremos que usuarios tienen similitudes con los demás de acuerdo a sus palabras en los posts.

## 1 Introduccion

Para este proyecto nos fue proporcionado un conjunto de posts de facebook, en el cual debimos limpiar los textos y procesarlos para su posterior analisis.

El archivo proporcionado es *data.csv* el cual contiene 5000 líneas de datos de las cuáles hay que extraer la mayor cantidad de información posible, por lo cual debimos hacer un proceso de limpieza.

## 2 Limpiando la data

El primer paso a realizar para analizar esta data y colocarla en la estrutura de un data frame, llamado "df". Este proceso fue difícil de realizar puesto que el archivo de entrada presentaba más de un tipo de codificación.

```
# Data Frame de entrada
posts <- read.table("data2.csv",header=TRUE,sep=";",dec=",",row.names=1)
df <- do.call("rbind", lapply(posts$post, as.data.frame))
```

Luego de esto creamos nuestro corpus el cual usamos para limpiar los datos. Las letras de las palabras fueron cambiadas todas a minúscula, se quitaron los signos de puntuación y palabra muy comunes que no tienen valor nuestro análisis, tales como artículos, conjunciones, etc.

```
myCorpus <- Corpus(VectorSource(df$X))
tm_map(myCorpus, function(x) iconv(enc2utf8(x), sub = "byte"))
myCorpus <- tm_map(myCorpus, PlainTextDocument)
myCorpus <- tm_map(myCorpus, tolower,lazy=TRUE)
myCorpus <- tm_map(myCorpus, removePunctuation,lazy=TRUE)
myCorpus <- tm_map(myCorpus, removeNumbers,lazy=TRUE)
removeURL <- function(x) gsub("http[[:alnum:]]*", "", x)
removeURL2 <- function(x) gsub("www[[:alnum:]]*", "", x)
removejaja <- function(x) gsub("\b(?:a*(?:ja)+j?|(?:l+o+)+l+)\b", "", x)
removehaha <- function(x) gsub("\b(?:a*(?:ha)+h?|(?:l+o+)+l+)\b", "", x)
```

```
myCorpus <- tm_map(myCorpus, removeURL,lazy=TRUE)
myCorpus <- tm_map(myCorpus, removeURL2,lazy=TRUE)
myCorpus <- tm_map(myCorpus, removejaja,lazy=TRUE)
myCorpus <- tm_map(myCorpus, removehaha,lazy=TRUE)
myStopwords <-
    c(stopwords('english'),stopwords('spanish'),"NA","xd","xD","like","RT","etc",
"csm","para","ser","wtf","sin","mas","una","los","nos")
myCorpus <- tm_map(myCorpus, removeWords, myStopwords,lazy=TRUE)
myCorpus <- tm_map(myCorpus, stemDocument,language="english",lazy=TRUE)

myCorpus <- tm_map(myCorpus, stemDocument,language="english",lazy=TRUE)
myStopwords2 <-
    c("est","esto","xd","jajajaja","y","un","una","te","a","el","d","jajaja",
"like","para","que","the","mira","the")

myCorpus <- tm_map(myCorpus, PlainTextDocument)
myCorpus <- tm_map(myCorpus, removeWords2, myStopwords2,lazy=TRUE)
```

Hicimos lo mismo para cada usuario, creamos un data frame y un corpus para cada uno, y tomamos toda la data para hacer el proceso de clusterización.

# 3 Procesamiento y análisis de los datos

Luego de que limpiamos la data nos enfocamos en hacer las matrices de frecuencia de documento, para sacar la frecuencia de palabras de cada usuario y en general de toda la data, lo hicimos de la siguiente manera:

```
myTdm <- TermDocumentMatrix(myCorpus,control=list(wordLengths=c(1,Inf)))
termFrequency <- rowSums(as.matrix(myTdm))
termFrequency <- subset(termFrequency, termFrequency>=15)
```

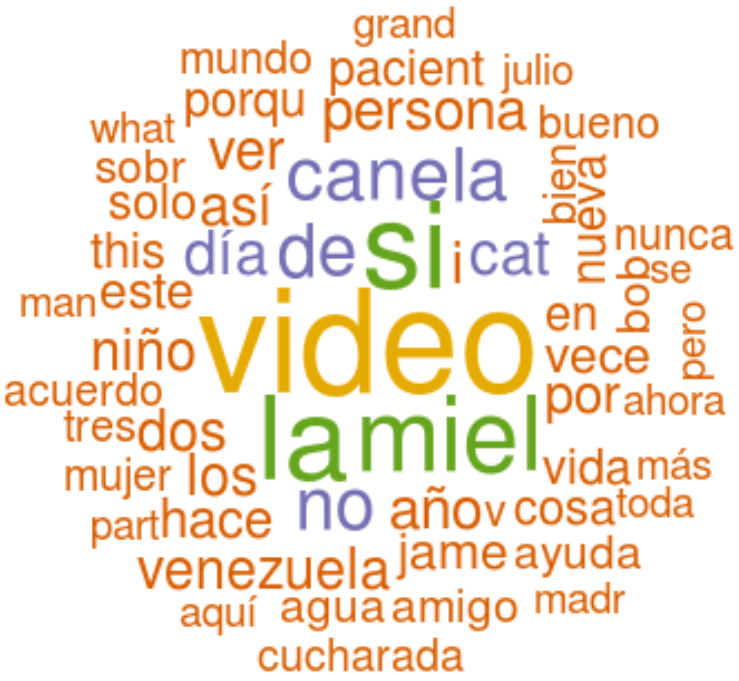
Para la impresion de la nube hicimos lo siguiente:

```
#Impresion Grafica todo el mundo
barplot(termFrequency, las=2)
library(wordcloud)
m <- as.matrix(myTdm)
wordFreq <- sort(rowSums(m), decreasing=TRUE)
set.seed (375)
#grayLevels <- gray( (wordFreq+10) / (max(wordFreq)+10))
wordcloud(words=names(wordFreq), freq=wordFreq,
    min.freq=10,random.order=F, colors=brewer.pal(6, "Dark2"))
```

Lo cual nos genero el siguiente gráfico:


Toda la data

Hicimos lo mismo con cada usuario:


grand
mundo
pacient
julio
porqu
persona
bueno
what
ver
canela
sobr
bien
nueva
solo
así
nunca
this
día
de
si
¡cat
se
bob
man
este
video
en
niño
vece
pero
acuerdo
por
ahora
tres
dos
la
miel
vida
más
mujer
los
no
año
v
cosa
toda
part
hace
jame
ayuda
venezuela
aquí
agua
amigo
madr
cucharada

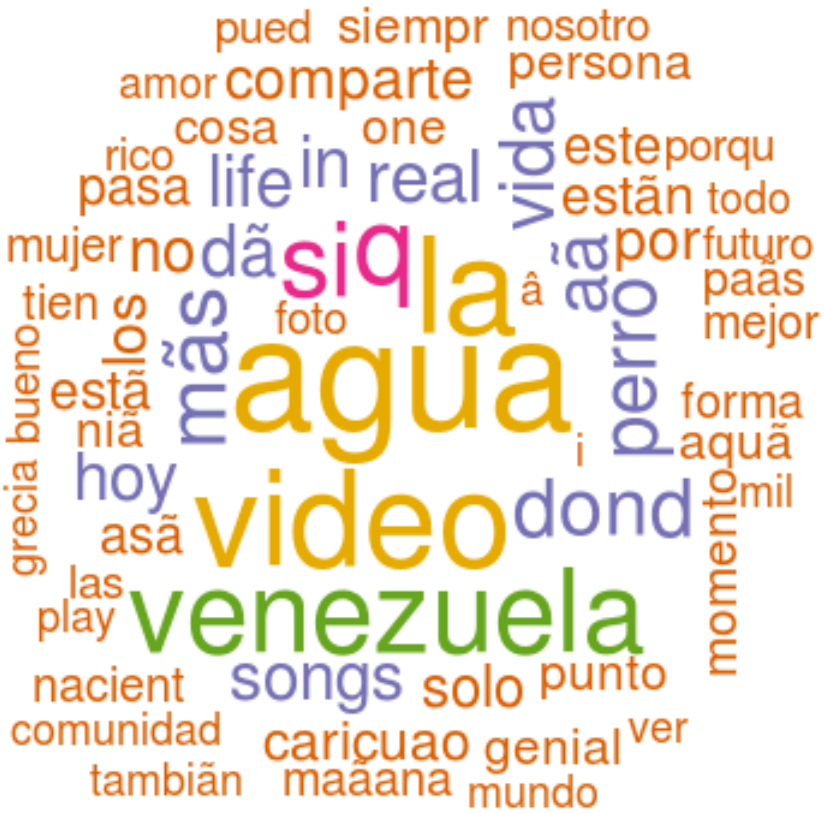
Usuario 1


hahaha
olvides
sabe
ucv
jose
momento
más
minuto
cute
tu
compartir
babi
quién
si
eso
yo
video
solo
vida
de
mar
ml
amigos
hace
página
ball
entr
no
la
favor
mucho
día
niño
playa
así
darle
videos
nunca
este
tan
foto
ivan
familiar
también
animal
ola
victor
mejor
llama

Usuario 2


obra
artãculo
salir
cristo
bien
arriba
tambiãn
dar
rezar
oraciãn
aquã
rep
catãlico
hoy
gracia
accião
llorã
hacer
aã
se
grand
está
sã
francisco
cosa
vio
viaj
joven
pozo
mãs
entr
us
video
san
fe
dã
da
por
los
en
de
vida
la
si
dio
va
niã
del
toda
â
amor
santo
cancião
paso
tierra
divina
no
papa
cada
chist
sobr
dios
visita
persona
âquã
burro
bs
lo
hacia
este
to
corazãn
part
campesino
cristiana
jornada
tãº
dos
estãn
conductor
gran

Usuario 3


pued
siempr
nosotro
amor
comparte
persona
cosa
one
rico
vida
este
porqu
pasa
life
in
real
estãn
todo
mujer
no
dã
siq
la
por
futuro
aa
â
perro
paãs
tien
los
foto
mejor
grecia
bueno
ãas
agua
estã
forma
niã
aquã
hoy
video
dond
mil
momento
asã
las
play
venezuela
nacient
songs
solo
punto
comunidad
caricuao
genial
ver
tambiãn
maãana
mundo

Usuario 4


Usuario 5

Como podemos observar la palabra que mas se repite en casi todos los usuarios es "Video" lo cual quiere decirnos que lo que probablemente mas se comparte en los posts de facebook son videos.

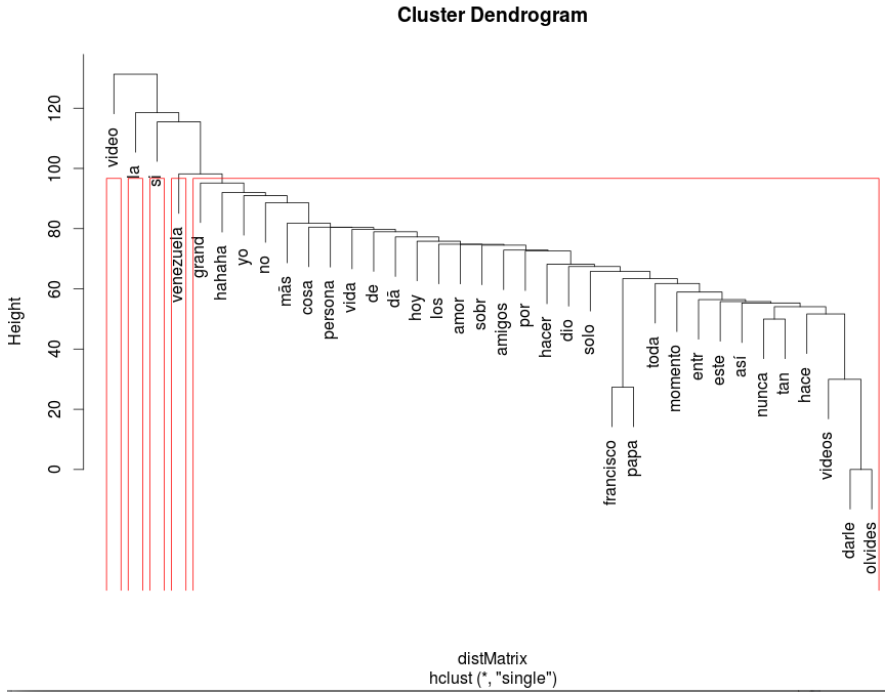
Ahora viendo el analisis de cada usuario:

- Usuario 1: Aqui podemos observar que la palabra mas frecuente en los posts de éste usuario es video, lo que indica que ve muchos videos y puede que esten relacionados con la cocina, ya que se ven palabras de dulces como "miel" o "canela".
- Usuario 2: En éste usuario observamos que se repiten mucho las palabras amigos, playa, compartir, video, esto puede indicarnos que esta persona es muy social.
- Usuario 3: Aqui observamos que las palabras mas repetidas son papa, tierra, vida, amor, francisco, gracia, puede indicarnos la tendencia religiosa de este usuario el cual parece ser católico practicante.
- Usuario 4: Con este usuario es un poco dificil encontrar algún patrón a simple vista, ya que las palabras mas repetidas son agua, video, venezuela, songs, tal vez le gusten los videos de surf o deportes acuáticos ?.
- Usuario 5: Bueno este usuario sigue la tendencia de la mayoria en la cual se observa que casi todos sus posts son de videos.

Para la clusterizacion de las palabras en la matriz hicimos lo siguiente, utilizamos clusterizacion jerárquica, lo hicimos de esta manera con la data total para luego observar en que grupo cae cada usuario dependiendo de cuales palabras tienen mas frecuencia me dirán en que grupo están

```
# remove sparse terms
myTdmAux <- removeSparseTerms(myTdm, sparse=0.99)
m2 <- as.matrix(myTdmAux)
frequency <- colSums(m2)
frequency <- sort(frequency, decreasing=TRUE)
#frequency
# cluster terms
distMatrix <- dist(scale(m2))
fit <- hclust(distMatrix, method="single")
plot(fit)
# cut tree into 9 clusters
rect.hclust(fit, k=5)
groups <- cutree(fit, k=5)
```

El código anterior genero el siguiente gráfico:




Para contar la frecuencia de cada palabra hicimos esto:

```
# count frequency
temp <- inspect(myTdmAux)
FreqMat2 <- data.frame(ST = rownames(temp), Freq = rowSums(temp))
row.names(FreqMat2) <- NULL
FreqMat2
```

A partir de los grupos que nos genero el grafico anterior podemos ahora clusterizar a los usuarios dependiendo de sus palabras, pero antes nos dimos a la tarea de agrupar cada usuario entre sí para ver cuales tenían palabras en común y ver quienes se parecían mas.

Lo hicimos haciendo la matriz de frecuencia para cada usuario(detalles en el source.R) y luego lo que hicimos fue crear vectores con las intersecciones de cada uno, para ello utilizamos la función **"intersect"** de la siguiente manera:

```
inters12 <- paste(intersect(FreqMat1$ST, FreqMat2$ST), collapse = " ")
inters13 <- paste(intersect(FreqMat1$ST, FreqMat3$ST), collapse = " ")
inters14 <- paste(intersect(FreqMat1$ST, FreqMat3$ST), collapse = " ")
inters15 <- paste(intersect(FreqMat1$ST, FreqMat5$ST), collapse = " ")

inters23 <- paste(intersect(FreqMat2$ST, FreqMat3$ST), collapse = " ")
inters24 <- paste(intersect(FreqMat2$ST, FreqMat4$ST), collapse = " ")
inters25 <- paste(intersect(FreqMat2$ST, FreqMat5$ST), collapse = " ")

inters34 <- paste(intersect(FreqMat3$ST, FreqMat4$ST), collapse = " ")
inters35 <- paste(intersect(FreqMat3$STm, FreqMat5$ST), collapse = " ")

inters45 <- paste(intersect(FreqMat4$ST, FreqMat5$ST), collapse = " ")

user1 <- c(NA,     inters12,  inters13,  inters14,  inters15)
user2 <- c(inters12, NA,   inters23,  inters24,  inters25)
user3 <- c(inters13, inters23, NA,       inters34,  inters35)
user4 <- c(inters14, inters24, inters34, NA,       inters45)
user5 <- c(inters15, inters25, inters35, inters45, NA)

# Matriz de individuo x individuo de las intersecciones
users.df <- cbind(user1, user2, user3, user4, user5)
```

Las intersecciones fueron asi:

- Usuario 1 con el 2: inters12 [1] "la no si video".
- Usuario 1 con 3: [1] "la no si video"
- Usuario 1 con 4:[1] "la no si video
- Usuario 1 con 5:[1] "la no si video
- Usuario 2 con 3:"cristo entr este la no si vida video"
- Usuario 2 con 4: "este la no si solo vida video"
- Usuario 2 con 5: "este la no si vida video"
- Usuario 3 con 4: "cosa dã está este hoy la mãs no persona por si vida video"
- Usuario 3 con 5: ""
- Usuario 4 con 5: "dã este la no si venezuela vida video"

Podemos concluir que el usuario 2 y 3 tienen bastantes palabras en comun, al igual que el 3 y 4, aunque la mayoria de esas palabras son basura que no se pudo limpiar correctamente con el stopwords.

El usuario 3 y 5 no tiene nada en común, por lo menos con sus palabras más frecuentes.

### 3.1 Clusterización

En esta sección buscamos clusterizar a los usuarios, observando cuales palabras son sus mas frecuentes y en base a ello compararlos con el grafico de la clusterizacion jerarquica de toda la data, para ello obsevamos sus matrices de frecuencia, sin embargo con la nube de palabras nos fue mas fácil analizarlos.

- El primer usuario caeria dentro de la categoria del grupo 1, ya que la palabra que mas se repite es video, por lo tanto caen en el grupo de video.
- El segundo usuario tiene bastantes palabras repetidas, por lo tanto su matriz de frecuencia es mas grande, sin embargo gracias al wordcloud podemos ver cuales palabras son mas frecuentes, en su caso es "amigos" cae en el grupo 5 el cual contiene todas las palabras que no son ni "video" ni "la" ni "si" ni "venezuela".
- El tercer usuario caeria en el grupo 2 por tener como palabra mas repetida "la" sin embargo al ser esta palabra un articulo que no pudo ser eliminado en la limpieza, esto me puede indicar que este usuario caeria en el grupo 5 tambien, a ser su otras palabras mas frecuentes, "tierra, papa, amor".
- El cuarto usuario tambien se clusteriza en el grupo 5 ya que si palabra mas repetida es agua, aunque tambien tiene una relacion con el grupo 1, que es video, por lo tanto lo pondriamos en el grupo 1.
- El quinto usuario fácilmente se observa que su palabra más frecuente es video, por lo tanto esta en el grupo 1.

De esta forma clusterizamos a los usuarios basándonos en la frecuencia de sus palabras y la clusterizacion que hicimos de toda la data.

## 4 CONCLUSION

Pudimos observar que definitivamente en base a los posts de la muestra, y la palabra mas frecuente de la muestra limpia, los usuarios hoy en día tienden a compartir videos más que cualquier otro tipo de dato, por lo menos en la red social de Facebook. También pudimos observar ciertas similitudes entre los usuarios, como palabras en común, esto podria sernos útil a la hora de aplicar lo que se conoce como Homofilia en una red social, para hacer sugerencias mas efectivas. Todo este tema de las redes sociales, y de su analisis viene en auge y encontramos muy interesante y entretenido el desarrollo del presente proyecto, a pesar de que nos encontramos con ciertas dificultades, como en el caso de la limpieza, y la manera en la que analizariamos a los usuarios en base a sus posts.