**UNIVERSIDAD CENTRAL DEL ECUADOR**

**FACULTAD E INGENIERÍA, CIENCIAS FÍSICAS Y MATEMÁTICA**

**CARRERA DE INGENIERÍA INFORMÁTICA**

**ANÁLISIS DE DATOS**

**INTEGRANTES:** Alexis Bautista

Bryan Catucuamba

David Jácome

Alejandro Naranjo

Richard Quimbiulco

**ELIMINACIÓN DE REPETIDOS, RELEVANCIA Y FACTOR DE PONDERACIÓN EN R**

Para esta práctica de eliminación de repetidos, relevancia y factor de ponderación vamos a utilizar los tweets descargados de la emisora Radio La Bruja, su nombre de usuario en Twitter es @labrujaecuador. La cantidad de tweets descargados es 390, y los obtuvimos el día lunes 17 de julio del 2017.

Importamos el archivo de los tweets

> tweets <-

+ read.csv(

+ "C:/Users/David/Documents/AnalisisDatos/Maquinas/Datos\_David/labrujaecuador\_2017-07-17.csv",

+ header = TRUE,

+ sep = ";"

+ )

Verificamos la cantidad de tweets

> NROW(tweets)

[1] 390

**Eliminación de repetidos**

Para eliminar los valores duplicados o repetidos debemos de seguir los siguientes pasos:

1. Depuramos el texto de los tweets, para ello debemos de eliminar símbolos de retweets, el símbolo de @, signos de puntuación, números y enlaces, a ese resultado lo agregamos al dataframe en una nueva columna llamada “depurado”. Además se procede a eliminar cualquier tipo de espacio en blanco, y ese resultado lo agregamos al dataframe en una nueva columna llamada “sinEspacios”

> for (i in 1:NROW(tweets)) {

+ texto <- tweets[i,2]

+ #eliminar simbolos de retweets

+ sinRT <- gsub("(RT|via)((?:\\b\\W\*@\\w+)+)", "", texto)

+ sinCuentas <- gsub("@|#", "", sinRT)

+ sinSimbolos <- gsub("[[:punct:]].", "", sinCuentas)

+ sinNumeros <- gsub("[[:digit:]]", "", sinSimbolos)

+ sinEnlaces <- gsub("http\\w+", "", sinNumeros)

+ tweets$depurado[i] <- sinEnlaces

+ tweets$sinEspacios[i] <- gsub("[[:space:]]|","",tweets$depurado[i])

+ }

1. Ordenamos al dataframe de tweets por el texto sin espacios del tweet.

> tweets <- tweets[order(tweets$sinEspacios),]

1. Utilizamos el siguiente comando para eliminar los tweets que quedaron en blanco luego de la depuración.

> tweets <- tweets[!(tweets$sinEspacios==""),]

1. Utilizamos el siguiente comando para eliminar los tweets repetidos.

> tweets <- tweets[!duplicated(tweets$sinEspacios),]

1. Procedemos a verificar que los tweets se haya eliminado, para ello observamos el número de tweets restantes.

> NROW(tweets)

[1] 327

**Relevancia**

Para calcular la relevancia nos hemos guiado del índices de relevancia:

Los comandos para encontrar el ratio anterior son los siguientes:

> userInfo <- lookupUsers(tweets$screenName)

> userFrame <- twListToDF(userInfo)

> userFrame$relevancia1 <- userFrame$followersCount/userFrame$friendsCount

> for(i in 1:NROW(tweets)) {

+ usuarioT <- tweets[i,12]

+ for(j in 1:NROW(userFrame)) {

+ usuarioF <- userFrame[j,11]

+ rel <- userFrame[j,18]

+ if (usuarioT == usuarioF) {

+ tweets[i,21] <- rel

+ }

+ }

+ }

**Factor de ponderación**

Para calcular el factor de ponderación realizamos los siguientes pasos:

1. Convertimos la columna localidad a mayúsculas.

> for (i in 1:NROW(tweets)) {

+ tweets[i, 23] <- toupper(tweets[i, 18])

+ }

1. Verificamos a donde pertenece cada tweet mediante el uso de expresiones regulares, para ello usamos el comando grepl, asignándole a cada tweet un código de localidad.

Los códigos de las localidades son los siguientes:

|  |  |
| --- | --- |
| **Codigo** | **Localidad** |
| 1 | Azuay |
| 2 | Bolivar |
| 3 | Cañar |
| 4 | Carchi |
| 5 | Cotopaxi |
| 6 | Chimborazo |
| 7 | El Oro |
| 8 | Esmeraldas |
| 9 | Guayas |
| 10 | Imbabura |
| 11 | Loja |
| 12 | Los Rios |
| 13 | Manabi |
| 14 | Morona Santiago |
| 15 | Napo |
| 16 | Pastaza |
| 17 | Pichincha |
| 18 | Tungurahua |
| 19 | Zamora Chinchipe |
| 20 | Galápagos |
| 21 | Sucumbíos |
| 22 | Orellana |
| 23 | Santo Domingo |
| 24 | Santa Elena |
| 25 | Otros |

> for (i in 1:NROW(tweets)) {

+ if (!is.na(tweets[i, 23])) {

+ resultado <-

+ grepl(

+ "AZUAY|SIGSIG|SEVILLA DE ORO|SANTA ISABEL|SAN FERNANDO|PUCARA|PAUTE|OÑA|NABÓN|GUALACEO|GUACHAPALA|GIRÓN|EL PAN|CUENCA|CHORDELEG|CAMILO PONCE ENRÍQUEZ",

+ tweets[i, 23]

+ )

+ if (resultado == TRUE) {

+ tweets[i, 24] <- "1"

+ } else {

+ resultado <-

+ grepl(

+ "BOLIVAR|SAN MIGUEL|LAS NAVES|GUARANDA|ECHEANDÍA|CHIMBO|CHILLANES|CALUMA",

+ tweets[i, 23]

+ )

+ if (resultado == TRUE) {

+ tweets[i, 24] <- "2"

+ } else {

+ resultado <-

+ grepl("CAÑAR|SUSCAL|LA TRONCAL|EL TAMBO|DÉLEG|CAÑAR|BIBLIÁN|AZOGUES",

+ tweets[i, 23])

+ if (resultado == TRUE) {

+ tweets[i, 24] <- "3"

+ } else ...

1. Calculamos el número total de tweets por localidad. Para ello inicializamos los contadores en 0.

> for (i in 1:NROW(tweets)) {

+ if (tweets[i, 24] == "1") {

+ AZUAY <- AZUAY + 1

+ } else if (tweets[i, 24] == "2") {

+ BOLIVAR <- BOLIVAR + 1

+ } else if (tweets[i, 24] == "3") {

+ CAÑAR <- CAÑAR + 1

+ } else if (tweets[i, 24] == "4") {

+ CARCHI <- CARCHI + 1

+ } else if (tweets[i, 24] == "5") {

+ CHIMBORAZO <- CHIMBORAZO + 1

+ } else if (tweets[i, 24] == "6") {

+ COTOPAXI <- COTOPAXI + 1

+ } else if (tweets[i, 24] == "7") {

+ EL\_ORO <- EL\_ORO + 1

+ } else if (tweets[i, 24] == "8") {

+ ESMERALDAS <- ESMERALDAS + 1

+ } else if (tweets[i, 24] == "9") {

+ GUAYAS <- GUAYAS + 1

+ } else if (tweets[i, 24] == "10") {

+ IMBABURA <- IMBABURA + 1

+ } else if (tweets[i, 24] == "11") {

+ LOJA <- LOJA + 1

+ } else if (tweets[i, 24] == "12") {

+ LOS\_RIOS <- LOS\_RIOS + 1

+ } else if (tweets[i, 24] == "13") {

+ MANABI <- MANABI + 1

+ } else if (tweets[i, 24] == "14") {

+ MORONA\_SANTIAGO <- MORONA\_SANTIAGO + 1

+ } else if (tweets[i, 24] == "15") {

+ NAPO <- NAPO + 1

+ } else if (tweets[i, 24] == "16") {

+ PASTAZA <- PASTAZA + 1

+ } else if (tweets[i, 24] == "17") {

+ PICHINCHA <- PICHINCHA + 1

+ } else if (tweets[i, 24] == "18") {

+ TUNGURAHUA <- TUNGURAHUA + 1

+ } else if (tweets[i, 24] == "19") {

+ ZAMORA\_CHINCHIPE <- ZAMORA\_CHINCHIPE + 1

+ } else if (tweets[i, 24] == "20") {

+ GALAPAGOS <- GALAPAGOS + 1

+ } else if (tweets[i, 24] == "21") {

+ SUCUMBIOS <- SUCUMBIOS + 1

+ } else if (tweets[i, 24] == "22") {

+ ORELLANA <- ORELLANA + 1

+ } else if (tweets[i, 24] == "23") {

+ SANTO\_DOMINGO\_TSACHILAS <- SANTO\_DOMINGO\_TSACHILAS + 1

+ } else if (tweets[i, 24] == "24") {

+ SANTA\_ELENA <- SANTA\_ELENA + 1

+ } else if (tweets[i, 24] == "25") {

+ OTROS <- OTROS + 1

+ }

+ }

1. Ingresamos los datos del censo realizado en el 2010 para conocer el número de personas mayores a 11 años en el Ecuador a la fecha.

> datosCensales <-

+ read\_excel(

+ "C:/Users/David/Documents/AnalisisDatos/Maquinas/Datos\_David/datosCensales.xlsx"

+ )

> porcenCenso <- datosCensales$porcentaje

1. Calculamos el porcentaje de tweets por localidad.

> porcentajeT[1, 4] <- (AZUAY \* 100) / NROW(tweets)

> porcentajeT[2, 4] <- (BOLIVAR \* 100) / NROW(tweets)

> porcentajeT[3, 4] <- (CAÑAR \* 100) / NROW(tweets)

> porcentajeT[4, 4] <- (CARCHI \* 100) / NROW(tweets)

> porcentajeT[5, 4] <- (CHIMBORAZO \* 100) / NROW(tweets)

> porcentajeT[6, 4] <- (COTOPAXI \* 100) / NROW(tweets)

> porcentajeT[7, 4] <- (EL\_ORO \* 100) / NROW(tweets)

> porcentajeT[8, 4] <- (ESMERALDAS \* 100) / NROW(tweets)

> porcentajeT[9, 4] <- (GUAYAS \* 100) / NROW(tweets)

> porcentajeT[10, 4] <- (IMBABURA \* 100) / NROW(tweets)

> porcentajeT[11, 4] <- (LOJA \* 100) / NROW(tweets)

> porcentajeT[12, 4] <- (LOS\_RIOS \* 100) / NROW(tweets)

> porcentajeT[13, 4] <- (MANABI \* 100) / NROW(tweets)

> porcentajeT[14, 4] <- (MORONA\_SANTIAGO \* 100) / NROW(tweets)

> porcentajeT[15, 4] <- (NAPO \* 100) / NROW(tweets)

> porcentajeT[16, 4] <- (PASTAZA \* 100) / NROW(tweets)

> porcentajeT[17, 4] <- (PICHINCHA \* 100) / NROW(tweets)

> porcentajeT[18, 4] <- (TUNGURAHUA \* 100) / NROW(tweets)

> porcentajeT[19, 4] <- (ZAMORA\_CHINCHIPE \* 100) / NROW(tweets)

> porcentajeT[20, 4] <- (GALAPAGOS \* 100) / NROW(tweets)

> porcentajeT[21, 4] <- (SUCUMBIOS \* 100) / NROW(tweets)

> porcentajeT[22, 4] <- (ORELLANA \* 100) / NROW(tweets)

> porcentajeT[23, 4] <- (SANTO\_DOMINGO\_TSACHILAS\*100)/NROW(tweets)

> porcentajeT[24, 4] <- (SANTA\_ELENA \* 100) / NROW(tweets)

> porcentajeT[25, 4] <- (OTROS \* 100) / NROW(tweets)

1. Calculamos el porcentaje de tweets por localidad y el factor de ponderación.

|  |
| --- |
| > for (i in 1:NROW(porcentajeT)) {  + porcentajeT[i, 5] <-  + (porcentajeT[i, 4]) / (1 - (porcentajeT[25, 4] / 100))  + if (porcentajeT[i, 2] == "OTROS") {  + porcentajeT[i, 6] <- 1  + } else {  + fp <- porcentajeT[i, 3] / porcentajeT[i, 5]  + if (is.infinite(fp) == FALSE) {  + porcentajeT[i, 6] <- fp  + }  + }  + } |

1. Incluimos el factor de ponderación al dataframe de tweets.

> for (i in 1:NROW(tweets)) {

+ codLocT <- tweets[i, 24]

+ for (j in 1:NROW(porcentajeT)) {

+ codLocP <- porcentajeT[j, 1]

+ factPon <- porcentajeT[j, 6]

+ if (codLocT == codLocP) {

+ tweets[i, 25] <- factPon

+ }

+ }

+ }