Laboratorio 7

Grupo 10

9/10/2019

Librerías a utilizar

```
# Instalación de paquetes neuevos en caso no estén descargados
if(!require(twitteR)) {install.packages("twitteR")}
```

```
## Loading required package: twitteR
```

```
if(!require(ROAuth)) {install.packages("ROAuth")}
```

```
## Loading required package: ROAuth
```

```
if(!require(readr)) {install.packages("readr")}
```

```
## Loading required package: readr
```

```
if(!require(tm)) {install.packages("tm")}
```

```
## Loading required package: tm
```

```
## Loading required package: NLP
```

```
if(!require(wordcloud)) {install.packages("wordcloud")}
```

```
## Loading required package: wordcloud
```

```
## Loading required package: RColorBrewer
```

```
if(!require(ggplot2)) {install.packages("ggplot2")}
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:NLP':
##
##     annotate
```

```
if(!require(base64enc)) {install.packages("base64enc")}
```

```
## Loading required package: base64enc
```

```
if(!require(openssl)) {install.packages("openssl")}
```

```
## Loading required package: openssl
```

```
if(!require(httpuv)) {install.packages("httpuv")}
```

```
## Loading required package: httpuv
```

```
if(!require(httr)) {install.packages("httr")}
```

```
## Loading required package: httr
```

```
##
## Attaching package: 'httr'
```

```
## The following object is masked from 'package:NLP':
##
##     content
```

```
if(!require(qdapRegex)) {install.packages("qdapRegex")}
```

```
## Loading required package: qdapRegex
```

```
##
## Attaching package: 'qdapRegex'
```

```
## The following object is masked from 'package:ggplot2':
##
##     %+%
```

```
# Se mandan a traer las librerías
library(qdapRegex)
library("twitteR") # Análisis de Twitter
library("ROAuth") # Accese al API
library("readr") # Lee los archivos
library("tm")  # Tranformaciones para el TM
library("wordcloud") # Nube de palabras
library(ggplot2) # Gráficos
library(base64enc)
library(openssl)
library(httpuv)
library(httr)
library("SentimentAnalysis")
```

```
##
## Attaching package: 'SentimentAnalysis'
```

```
## The following object is masked from 'package:base':
##
##     write
```

```
library("SnowballC") # Sentiment analysis
library(tidytext) # Otro análisis de sentimientos.
library("syuzhet")
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:qdapRegex':
##
##     explain
```

```
## The following objects are masked from 'package:twitteR':
##
##     id, location
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(plyr)
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following object is masked from 'package:twitteR':
##
##     id
```

Obteniendo el acceso a Twitter

Para la realización de este laboratorio se decidió utilizar Twitter, para ello se creó una APP desde la página para developers de Twitter, donde se debía de obtener la autorización de Twitter para poder tener una APP de developer. Luego, con la autorización fue posible obtener el API y el token para poder obtener los tweets.

```
## [1] "Using direct authentication"
```

Obtención de los datos para el Hashtag TráficoGT

Se utilizaron las librerías *TwitterR* y *ROAuth* para la obtención de los datos y análisis. Se tomó la decisión de utilizar el $\#TraficoGt$, con ayuda de las librerias importadas se buscaron los tweets y se creó un data frame a partir de los resultados obtenidos.

```
search.string <- "#TraficoGT"
result.term <- searchTwitter(search.string, n = 1000)
head(result.term)
```

```
## [[1]]
## [1] "pablonet1984: RT @amilcarmontejo: Trailer colisiona contra separadores viales en km 12.5 #Transi
##
## [[2]]
## [1] "QONTROL: Parece que siguen abriendo carros en #Zona2 #Guatemala @Marti7 #Marti7 @PNCdeGuatemala
##
## [[3]]
## [1] "TransitoCa: RT @libertadgarrido: ¡Cuidado con el semáforo de la Avenida Elena y 9calle! Ambas vi
##
## [[4]]
## [1] "clara_illescas: RT @amilcarmontejo: Ataque armado. \n\nReportan un hombre fallecido dentro del a
##
## [[5]]
## [1] "YoLaPulga: RT @libertadgarrido: ¡Cuidado con el semáforo de la Avenida Elena y 9calle! Ambas vía
##
## [[6]]
```

```
## [1] "dubonpined: RT @amilcarmontejo: Trailer colisiona contra separadores viales en km 12.5 #Transito
```

```
#Se almacena todo en un Data Frame
```

```
dataTweets<- twListToDF(result.term)
write.csv(dataTweets, "tweets.csv")
tweets.text <- sapply(result.term, function(x) x$getText())
```

Limpieza y preprocesamiento

Se comenzó eliminando la palabra "rt" debido a que se observó que los usuarios realizaban "rt" de tipo comentario. Ademas, cuando se hace esto aparece el nombre del usuario al cual rwtweetearon, entonces se prosiguió a eliminar también los usuarios. Acontinuación se eliminaron los hastags debido a que era lo que estabamos utilizando para filtrar los tweets que teníamos. Por último se realizó una limpieza normal de texto. Cabe mencionar que se eliminaron las tildes para evitar palabras repetidas escritas de diferente manera, dado que algunas personas escriben con tildes y otras no.

```
# Quitar ("rt")
tweets.text <- gsub("rt", "", tweets.text)
# Quitar @Usuario
tweets.text <- gsub("@\\w+", "", tweets.text)
# Quitar #Hashtag
tweets.text <- gsub("#\\S+", "", tweets.text)
# Quitar puntuaciones
tweets.text <- gsub("[[:punct:]]", "", tweets.text)
# Quitar links
tweets.text <- gsub("http\\w+", "", tweets.text)
# Quitar tabs
tweets.text <- gsub("[ |\t]{2,}", "", tweets.text)
# Quitar espacios en blanco del principio
tweets.text <- gsub("^ ", "", tweets.text)
# Quitar espacios en blanco del final
tweets.text <- gsub(" $", "", tweets.text)
# Quitar emojis
tweets.text <- gsub('\\p{So}|\\p{Cn}', "", tweets.text, perl = TRUE)
# Convertir a minúsculas
tweets.text <- tolower(tweets.text)
# Cambio de tildes
tweets.text <- gsub("á", "a", tweets.text)
tweets.text <- gsub("é", "e", tweets.text)
tweets.text <- gsub("í", "i", tweets.text)
tweets.text <- gsub("ó", "o", tweets.text)
tweets.text <- gsub("ú", "u", tweets.text)
```

Una vez limpio el texto, se prosiguió a realizar el corpus. Cabe mencionar que cuando se creó el corpus se eliminaron las stopwords en español.

```
#create corpus
#clean up by removing stop words
tweets.corpus <- Corpus(VectorSource(tweets.text))
tweets.corpus <- tm_map(tweets.corpus, content_transformer(removeNumbers))
```

```
## Warning in tm_map.SimpleCorpus(tweets.corpus,
## content_transformer(removeNumbers)): transformation drops documents
```

```
tweets.corpus <- tm_map(tweets.corpus, function(x) removeWords(x,stopwords("spanish")))
```

```
## Warning in tm_map.SimpleCorpus(tweets.corpus, function(x) removeWords(x, :
## transformation drops documents
```

Análisis exploratorio

Nube de palabras

En el worldcloud podemos observar en forma de nube las palabras más repetidas. Se encotró que la palabra con más frecuencia fue "ruta" y "zona". Por otro lado, se obtuvo un error que indicaba que habían varias palabras que no se pudieron plottear, entonces no se sabe que tanto se repitieron en los textos. Es por eso que se realizó un histograma de frecuencias de palabras, donde se obtuvieron la palabras que más se repetían y las que menos se repetían.

```
#generate wordcloud
wordcloud(tweets.corpus,min.freq = 2, scale=c(3,0.5), colors=brewer.pal(8, "Dark2"),
          random.color= TRUE, random.order = FALSE, max.words = 150)
```

Histograma de palabras con más frecuencia

Para el histograma de palabras con más frecuencia, se plottearon la palabras que tenían una frecuencia mayor a 50. Como se mencionó en los comentarios de worlcloud, hubieron palabras que no se pudieron plottear. Sin embargo, en este histograma se puede observar una de ellas, "atlántico". Cabe mencionar que la palabra atlántico hace referencia a ruta el atlántico.

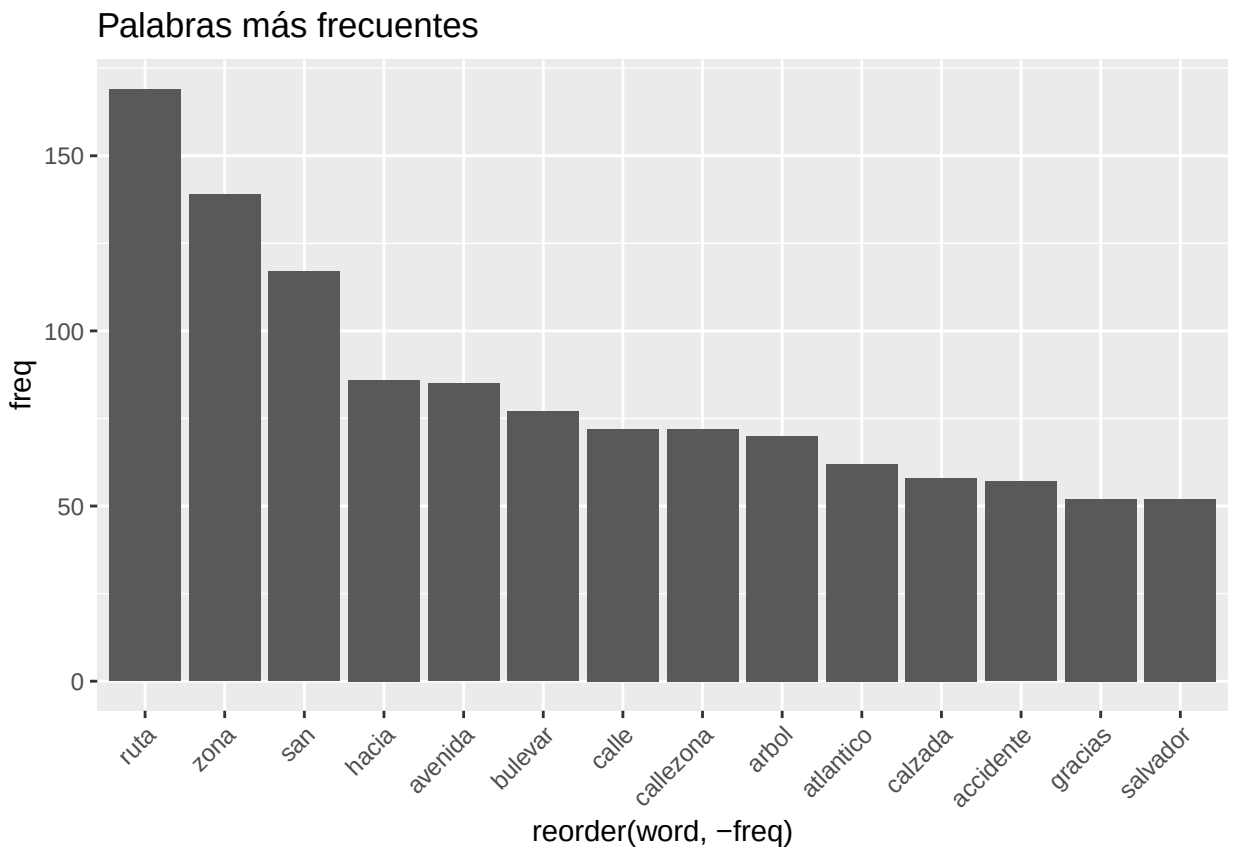
```
dtm <- DocumentTermMatrix(tweets.corpus)

freq <- colSums(as.matrix(dtm))
```

```
wfTweet <- data.frame(word = names(freq), freq = freq)
head(wfTweet)
```

```
##                 word freq
## afectando afectando   13
## ambas         ambas   17
## colisiona colisiona   13
## rttrailer rttrailer   11
## ruta           ruta  169
## salvador   salvador   52
```

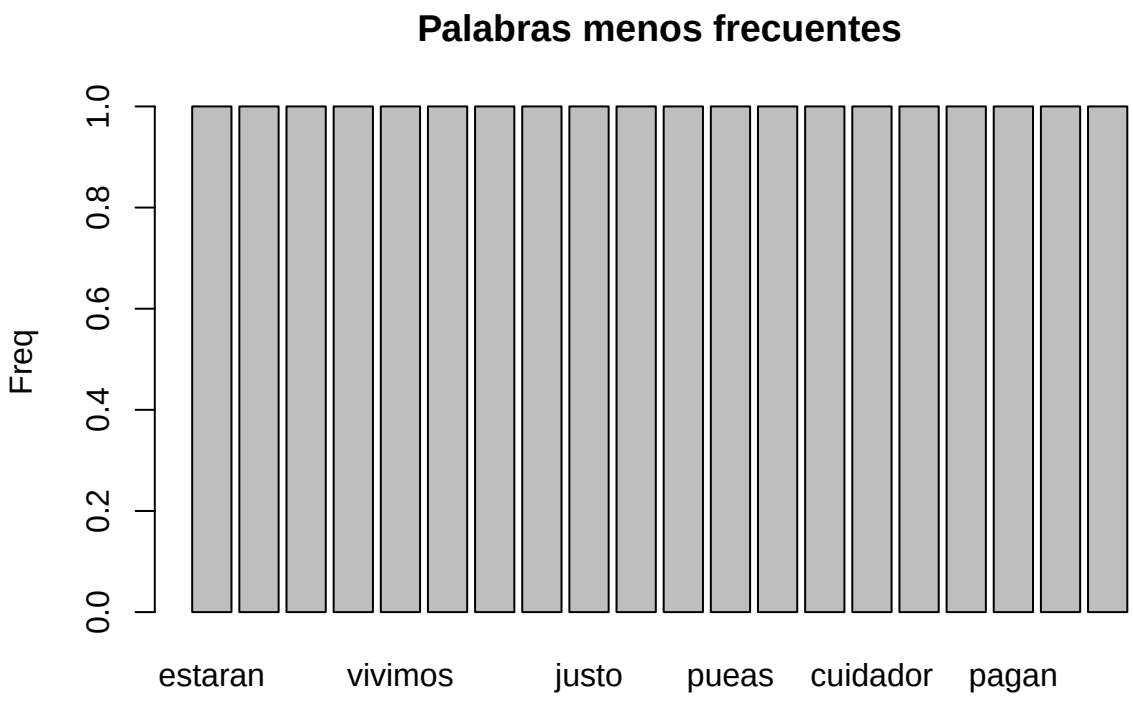
```
HistoR <- ggplot(subset(wfTweet, freq>50), aes(x = reorder(word, -freq), y = freq)) +
          geom_bar(stat = "identity") + ggtitle("Palabras más frecuentes") +
          theme(axis.text.x=element_text(angle=45, hjust=1))
HistoR
```



Histograma de palabras con menos frecuencia

En el histograma con menos frecuencia encontramos palabras como “pincha”, “cuota” y “mezcla”.

```
freqWords <- apply(dtm,2,sum)
freqData <- data.frame(word=names(freqWords), frecuency=as.numeric(freqWords),
                       stringsAsFactors = FALSE)
freqData<- freqData[order(freqData$frecuency, decreasing=TRUE), ]
barplot(tail(freqData$frecuency, n = 20),
        names.arg = tail(freqData$word, n = 20),
        main = "Palabras menos frecuentes", ylab = "Freq")
```



Asociasión de palabras

Según las palabras más frecuentes, se encontró la palabra 'zona', vamos a indagar que palabras se encuentran asociadas a esta con el fin de poder determinar algún tipo de relación.

```
findAssocs(dtm, 'zona', 0.2)
```

```
## $zona
##        bloquean       atlantico         entrada          rafael
##            0.51            0.50            0.49            0.49
## rtmanifestantes         colonia   manifestantes        circulan
##            0.41            0.37            0.37            0.36
##        limpieza       ocurridos         pinulas             ano
##            0.36            0.36            0.36            0.33
##      rtusuarios     conductores          guate     imprudentes
##            0.33            0.32            0.32            0.32
##         pedrera       boulevard           rtmi             col
##            0.32            0.31            0.31            0.30
##           rumbo       derrumbes            san            ruta
##            0.27            0.26            0.24            0.23
##          afecta        realizan         cruzado        estanque
##            0.23            0.22            0.22            0.22
##   semirremolque        usuarios         rtdejan
##            0.22            0.20            0.20
```

```
findAssocs(dtm, 'avenida', 0.2)
```

```
## $avenida
##        callezona           calle        derribada           motor
##             0.54            0.40            0.39            0.39
##            señal          volcado        automovil          bolivar
##             0.39            0.39            0.35            0.35
##        chinautla         inscrito         cableado        circuito
```

```
##             0.35             0.35             0.35             0.35
##              coo            poste             tuit        conductor
##             0.35             0.35             0.35             0.33
##           detuvo           rtabra          rtcarro             taxi
##             0.33             0.33             0.32             0.29
##       precaucion              blo        rtdormido        electrico
##             0.28             0.28             0.28             0.28
##            colon         banqueta      verbenazona          mercado
##             0.27             0.25             0.25             0.25
##        provienen        desviados        mecanicas        rtautobus
##             0.25             0.25             0.23             0.23
##          autobus       ministerio rtrepoabloqueado    manifestacion
##             0.23             0.23             0.23             0.22
##           fallas         fallecio         vehiculos              cal
##             0.21             0.21             0.21             0.21
```

Vemos que la palabra "avenida" se relaciona con advertencias o incidentes tipo "volcado", "precaucion" y lugares como "montufar", "verbena", "chinautla" y nos llama la atención la palabra "manifestaciongt". Por otro lado se menciona la palabra "bolivar" que hace referencia a la avenida Bolivar, una de las rutas con más tráfico en Guatemala.

```
findAssocs(dtm, 'ruta', 0.2)
```

```
## $ruta
##        atlantico         bloquean          entrada           rafael
##             0.56             0.42             0.41             0.41
##    manifestantes   interamericana              col              san
##             0.39             0.36             0.36             0.33
##         salvador         pacifico         usuarios        afectando
##             0.30             0.30             0.29             0.25
##        colisiona     separadores           viales         viajeros
##             0.25             0.25             0.25             0.24
##        rttrailer             zona          cruzado         estanque
##             0.23             0.23             0.23             0.23
##    semirremolque         afectade         multiple        kilometro
##             0.23             0.23             0.23             0.23
##           genera            ambas          rtdejan derrumbestrabaja
##             0.22             0.21             0.21             0.21
##        transitas
##             0.21
```

También decidimos analizar la palabra "ruta", la palabra que presenta mayor correlación es la de "atlántico", lo cual nos lleva a cuestionarnos: ¿acaso se realizan mayor cantidad de reportes de tránsito sobre la Ruta al Atlántico?

Así también vemos que se relaciona bastante con otras carreteras como la Interamericanta, Pacífico y Salvador. Otro hecho que nos llama la atención es su relación con las palabras "bloquean", "entrada" y "manifestantes" lo cual nos lleva a inferir que también se reportan con mayor frecuencia los bloqueos que realizan los manifestantes.

Analizando los retweets

Para ello, se creó un data frame que contuviera solamente el texto y la frecuencia de Retweets que obtuvo, a partir de el con la función *order* se colocaron en orden descendente y se mostrarán los primeros 10.

```
txandrt <- dataTweets[c(1,12)]
rt<-txandrt[order(txandrt$retweetCount, decreasing = TRUE),]
```

```r
rtMost <-rt[!duplicated(rt),]
head(rtMost,10)
```

```
##
## 109   RT @ConsejoMontejo: Se robaron automóvil negro, hyundai accent, placa P-778HGZ en zona 11.\n\n
## 152                                                  RT @avec_tats: Y hoy cómo en todas las quincenas
## 108     RT @ConsejoMontejo: "Porfa difundir. Se lo robaron a un amigo el fin de semana. Si lo ven av
## 18   RT @ConsejoMontejo: Esta placa la encontró Charly Hernández. Si es suya por favor comunicarse a
## 141  Esta placa la encontró Charly Hernández. Si es suya por favor comunicarse al número de teléfono
## 847                                              RT @SomosChapinas: Y hoy cómo en todas las quincenas
## 72  RT @VosDaviid: Mi Guate y sus conductores imprudentes.\nZona 6, Boulevard La Pedrera.\n#traficog
## 592   Mi Guate y sus conductores imprudentes.\nZona 6, Boulevard La Pedrera.\n#traficogt #TransitoGT
## 151  RT @amilcarmontejo: #UrgenteGT.\n\nEmulsión asfáltica se derrama desde cisterna involucrado en
## 224      #UrgenteGT.\n\nEmulsión asfáltica se derrama desde cisterna involucrado en colisión de trans
##     retweetCount
## 109          126
## 152          106
## 108           87
## 18            61
## 141           61
## 847           30
## 72            29
## 592           29
## 151           28
## 224           28
```

Observamos que la mayoría son retweets de ConsejoMontejo y que contienen la palabra "UrgenteGt"

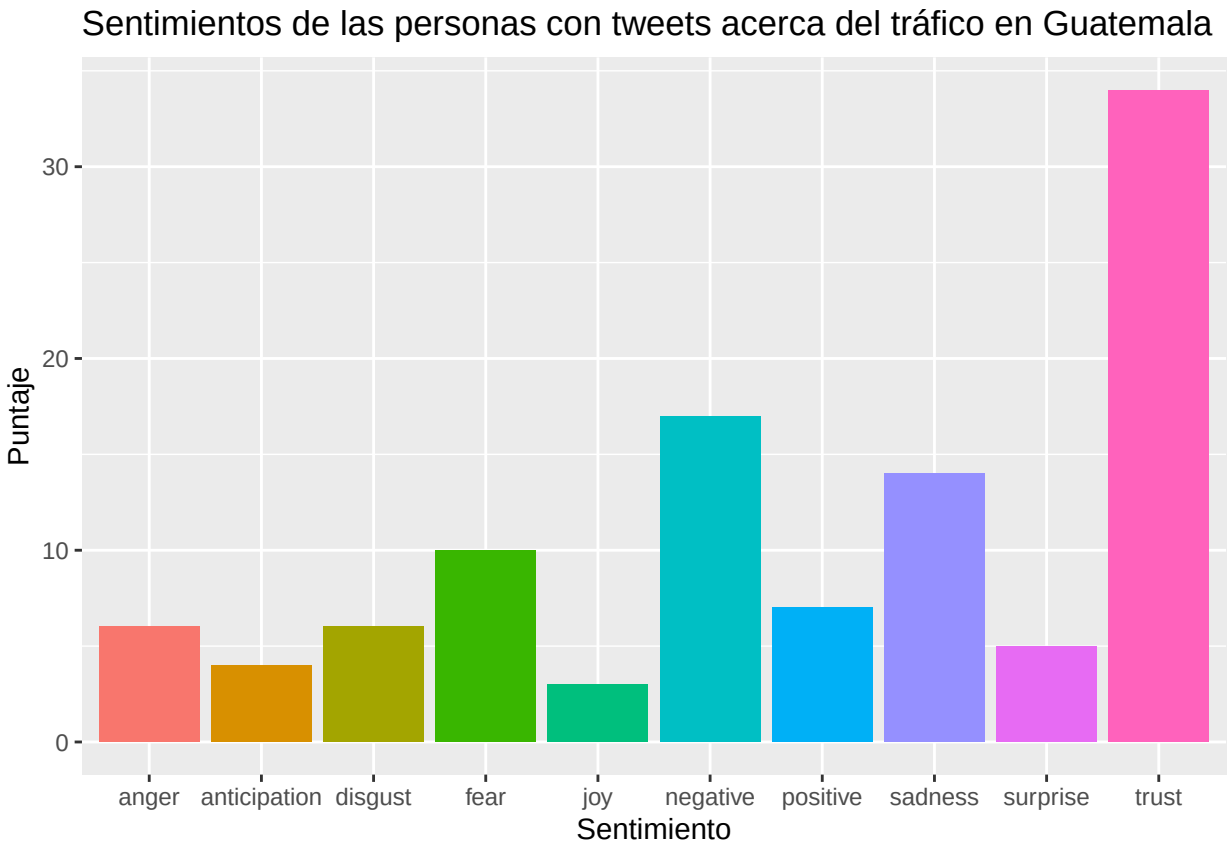
Análisis de sentimientos

Con el fin de determinar la razón o la actitud de los usuarios que publican tweets usando el hashtag #TraficoGT, se decidió realizar un análisis de sentimientos y así determinar si la mayoría de tweets son positivos o negativos. En este caso, vemos que la mayoría de tweets muestran confianza/seguridad y una segunda parte muestra negativismo. Esto posiblemente se debe a que la mayoría de guatemaltecos saben con certeza que en ciertas áreas de la ciudad o del país hay bastante embotellamiento (por lo que muestran seguridad en sus tweets); por otra parte, el negativismo se debe a que nadie le gusta el tráfico y usualmente se expresa en esta red social para desahogar su frustramiento.

```r
senti.tweets <-get_nrc_sentiment(as.vector(dataTweets$text))
# Total de cada sentimiento
Sentimentscores <-data.frame(colSums(senti.tweets[,]))

names(Sentimentscores)<-"Puntaje"
Sentimentscores <-cbind("Sentimiento"=rownames(Sentimentscores),Sentimentscores)
rownames(Sentimentscores)<-NULL

# Se grafica los puntajes de cada sentimiento
ggplot(data=Sentimentscores,aes(x=Sentimiento,y=Puntaje))+geom_bar(aes(fill=Sentimiento),stat = "identi
  theme(legend.position="none")+
  xlab("Sentimiento")+ylab("Puntaje")+ggtitle("Sentimientos de las personas con tweets acerca del tráfi
```



Ahora bien, se analizan cuáles son los comentarios positivos, negativos y neutros dentro de nuestra data de tweets. Se observa lo siguiente:

- *Tweets positivos*: Se habla la alegría que hay en las nuevas rutas, mensajes de precaución, piden ayuda para que los policías revisen pasos, rutas o desniveles en las carreteras.
- *Tweets negativos*: Se habla acerca de accidentes automovilísticos, fallas mecánicas que obstruyen el paso en las carreteras.
- *Tweets neutros*: No se observa nada relevante, solo alertas o avisos, tales como el tráfico usual de cada quincena y autos que se suben a las banquetas.

```
# Valor de los sentimientos
valor <- get_sentiment(dataTweets$text)

# Asignando los valores de sentimientos
max <- dataTweets$text[valor == max(valor)]
min <- dataTweets$text[valor == min(valor)]
```

Tweets positivos

```
positivo <- dataTweets$text[valor>0]
tail(positivo)
```

```
## [1] "@amilcarmontejo #Trafico #TraficoGT #TransitoGt #RedBull #Xtrem #Extrem https://t.co/Spa9sehC3s
## [2] "@EmixtraPablo por favor le solicito que revisen los tiempos del policía que está bajo el paso a
## [3] "Continúa el ingreso del primer frente frío norte el cuál impide la formación de lluvias intensas
## [4] "Lee todo lo que ocurre en #Guatemala.  Artículos de #noticias y #revistas haciendo #tendecia.
## [5] "Excelente noticia y que bueno que se abran nuevas rutas. Ideal sería que cuando pase esto, se sa
```

```
## [6] "#OjoDelLector | #TráficoGT lento en carretera al Atlántico hacia la capital. Maneje con precauci
```

Tweets negativos

```
negativo <- dataTweets$text[valor<0]
head(negativo)
```

```
## [1] "RT @libertadgarrido: ¡Cuidado con el semáforo de la Avenida Elena y 9calle! Ambas vías están en
## [2] "RT @libertadgarrido: ¡Cuidado con el semáforo de la Avenida Elena y 9calle! Ambas vías están en
## [3] "¡Cuidado con el semáforo de la Avenida Elena y 9calle! Ambas vías están en verde #TraficoGT @ami
## [4] "RT @amilcarmontejo: Auxilian a un hombre con herida en la mano, la cual se provocó al trabajar c
## [5] "Auxilian a un hombre con herida en la mano, cuando trabajaba con una máquina.\n\nBomberos reali
## [6] "Auxilian a un hombre con herida en la mano, la cual se provocó al trabajar con una máquina. \n\n
```

Tweets neutros

```
neutral <- dataTweets$text[valor==0]
tail(neutral,2)
```

```
## [1] "RT @avec_tats: Y hoy cómo en todas las quincenas. #TraficoGT #Guatemala https://t.co/eFEEZW452a
## [2] "RT @SomosChapinas: Y hoy cómo en todas las quincenas. #TraficoGT #Guatemala https://t.co/d2uzuI0
```

Conclusiones

Según el análisis realizado a los tweets del tráfico en Guatemala, se llegaron a las siguientes conclusiones:

- Como Ruta al Atlántico es una de las palabras con mucha frencuencia, se puede decir que es una de las rutas con más tráfico en Guatemala.
- El segundo sentimiento más alto dentro de los tweets realizados, se muestra un comportamiento negativo ya que se cree que las personas usan este medio para mostrar las molestias acerca del embotellamiento capitalino.
- La mayoría de retweets son de ConsejoMontejo, cuyos tweets contienen la palabra UrgenteGT.
- La mayoría de tweets negativos informan accidentes o vehículos en las carreteras, mostrando molestia y desesperación.
- Nos llamó la atención que una de las palabras asociadas a “avenida” era “Bolívar”, lo cual muestra coherencia, ya que es una de las rutas con más embotellamiento, al igual que la Atanasio Tzul.
- Las rutas siempre son bloqueadas por manifestaciones.