Proyecto semestral: Hito 1

Nicolás Herrera, Yesenia Marulanda, Franco Migliorelli, Samuel Sánchez, Sebastián Urbina

Octubre 2020

Table of Contents

# Motivacion:

Twitter es una de las redes sociales más utilizadas para comentar, compartir o debatir temas de actualidad y tendencias. El primer trimestre de 2020 tuvo un aumento de un 23% en comentarios diarios con respecto al mismo periodo de tiempo del año 2019 [1], siendo una de sus principales causas la pandemia del COVID-19.

Es por este contexto que analizar datos de twitter es interesante, además de que se puede obtener información en tiempo real de sucesos que están ocurriendo, se pueden ocupar métodos de la minería de datos sobre esta información y es fácil manipular grandes volúmenes de información. Por estas razones, nace la motivación por explorar tweets relacionados con coronavirus en un intervalo de tiempo acotado, desde la perspectiva de los sentimientos y relacionandolos con el contexto pais desde donde se emiten; buscando establecer si este ùltimo influencia la percepcion de las personas acerca de la pandemia.

Dicho lo anterior, los puntos claves a analizar en el desarrollo del proyecto son:

* Identificar cómo se relaciona el sentimiento identificado con el contexto país (según mayor o menor presencia del sentimiento).
* Categorizar países según mayor o menor presencia de sentimiento y relacionarlos con algun índice de felicidad publicado en el último año.
* Identificar palabras que son clave a la hora de categorizar el sentimiento.
* Establecer algoritmos para predecir sentimientos de forma sistematizada.
* Entrenar modelos de clasificaciòn en base a tweets usando un dataset de entrenamiento y un dataset de evaluacion.

# Descripción de base de datos

La base de datos, extraída desde la plataforma *Kaggle*[2], esta compuesta en principio por 44.955 tweets relacionados con el tema COVID 19 y que fueron publicados desde el 2 de marzo al 14 de abril de 2020. Además de encontrarse el texto publicado se encuentran en la base de datos los atributos de fecha exacta de publicación, ubicación desde la cual se realizla publicación, un identificador para el usuario y la asignación de sentimiento para cada tweet. La asignación de sentimiento a cada Tweet fue realizada de forma manual por el propietario de la base de datos. Esta variable de sentimiento sería un punto de comparación para un posible modelo de clasifiación de los sentimientos de los tweets.

# Exploración de datos

El objetivo de esta sección es describir la base de datos seleccionada, mostrando estadísticas de resumen, gráficos relevantes para la descripción y un breve análisis sobre estos.

El primer paso consiste en cargar todas las librerías que se utilizarán en este trabajo, las cuales permiten realizar gráficos y trabajar con los datos de una forma más simple y eficiente.

library(ggplot2)  
library(dplyr)  
library(tidyverse)  
library(tidytext)  
library(stopwords)  
library(wordcloud)  
library(wordcloud2)  
library(stringr)

A continuación se procece a cargar la base de datos para poder realizar el análisis respectivo.

train <- read.csv("data/Corona\_NLP\_train.csv", encoding="Latin-1")  
test <- read.csv("data/Corona\_NLP\_test.csv", encoding="Latin-1")  
df <- rbind(train,test)

Tal y como se mencionó anteriormente, esta base de datos contiene 6 columnas y 44.955 filas de datos. Para poder tener una referencia de cómo son estos campos, a continuación se puede observar una vista previa de las primeras filas del set de datos.

head(df)

## UserName ScreenName Location TweetAt  
## 1 3799 48751 London 16-03-2020  
## 2 3800 48752 UK 16-03-2020  
## 3 3801 48753 Vagabonds 16-03-2020  
## 4 3802 48754 16-03-2020  
## 5 3803 48755 16-03-2020  
## 6 3804 48756 ÃœT: 36.319708,-82.363649 16-03-2020  
## OriginalTweet  
## 1 @MeNyrbie @Phil\_Gahan @Chrisitv https://t.co/iFz9FAn2Pa and https://t.co/xX6ghGFzCC and https://t.co/I2NlzdxNo8  
## 2 advice Talk to your neighbours family to exchange phone numbers create contact list with phone numbers of neighbours schools employer chemist GP set up online shopping accounts if poss adequate supplies of regular meds but not over order  
## 3 Coronavirus Australia: Woolworths to give elderly, disabled dedicated shopping hours amid COVID-19 outbreak https://t.co/bInCA9Vp8P  
## 4 My food stock is not the only one which is empty...\n\nPLEASE, don't panic, THERE WILL BE ENOUGH FOOD FOR EVERYONE if you do not take more than you need. \nStay calm, stay safe.\n\n#COVID19france #COVID\_19 #COVID19 #coronavirus #confinement #Confinementotal #ConfinementGeneral https://t.co/zrlG0Z520j  
## 5 Me, ready to go at supermarket during the #COVID19 outbreak.\n\nNot because I'm paranoid, but because my food stock is litteraly empty. The #coronavirus is a serious thing, but please, don't panic. It causes shortage...\n\n#CoronavirusFrance #restezchezvous #StayAtHome #confinement https://t.co/usmuaLq72n  
## 6 As news of the regionÂ’s first confirmed COVID-19 case came out of Sullivan County last week, people flocked to area stores to purchase cleaning supplies, hand sanitizer, food, toilet paper and other goods, @Tim\_Dodson reports https://t.co/cfXch7a2lU  
## Sentiment  
## 1 Neutral  
## 2 Positive  
## 3 Positive  
## 4 Positive  
## 5 Extremely Negative  
## 6 Positive

Para poder trabajar con los datos es necesario convertir algunos formatos de las variables, en particular en este caso, se procede a transformar las variables “TweetAt” a un formato de fecha, dado que corresponde al momento en donde se realizó el tweet, y la variable “OriginalTweet” que corresponde al contenido del tweet realizado.

df$TweetAt <- as.Date(df$TweetAt, format="%d-%m-%y")  
df$OriginalTweet <- as.character(df$OriginalTweet)

Los formatos de cada variable del set de datos son mostrados a continuación:

str(df)

## 'data.frame': 44955 obs. of 6 variables:  
## $ UserName : int 3799 3800 3801 3802 3803 3804 3805 3806 3807 3808 ...  
## $ ScreenName : int 48751 48752 48753 48754 48755 48756 48757 48758 48759 48760 ...  
## $ Location : chr "London" "UK" "Vagabonds" "" ...  
## $ TweetAt : Date, format: "2020-03-16" "2020-03-16" ...  
## $ OriginalTweet: chr "@MeNyrbie @Phil\_Gahan @Chrisitv https://t.co/iFz9FAn2Pa and https://t.co/xX6ghGFzCC and https://t.co/I2NlzdxNo8" "advice Talk to your neighbours family to exchange phone numbers create contact list with phone numbers of neigh"| \_\_truncated\_\_ "Coronavirus Australia: Woolworths to give elderly, disabled dedicated shopping hours amid COVID-19 outbreak htt"| \_\_truncated\_\_ "My food stock is not the only one which is empty...\n\nPLEASE, don't panic, THERE WILL BE ENOUGH FOOD FOR EVERY"| \_\_truncated\_\_ ...  
## $ Sentiment : chr "Neutral" "Positive" "Positive" "Positive" ...

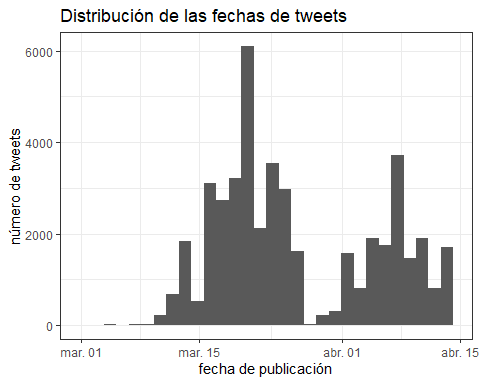
Además, es importante mencionar con qué periodos se estará trabajando, por lo que el resultado de la siguiente línea de código arroja la fecha mínima y máxima del set de datos, siendo el 2 de marzo de 2020 y el 14 de abril de 2020 respectivamente.

summary(df$TweetAt)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## "2020-03-02" "2020-03-19" "2020-03-23" "2020-03-26" "2020-04-06" "2020-04-14"

Para ver cómo están distribuidos los tweets en el tiempo, a continuación se realiza y muestra un histograma con las distribuciones de las fechas de los tweets realizados y registrados en esta base de datos.

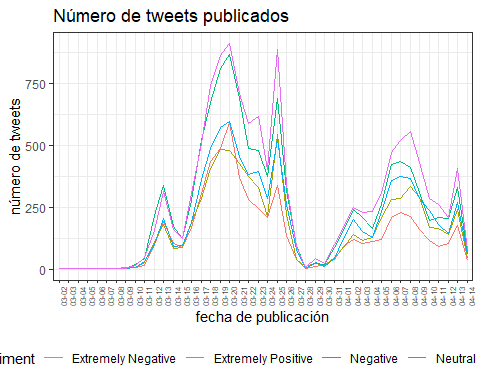
ggplot(data=df, aes(x=TweetAt)) + geom\_histogram(position="identity", bins=30) +  
 labs(title = "Distribución de las fechas de tweets", x = "fecha de publicación",  
 y = "número de tweets") + theme\_bw()



Un comportamiento particular de los datos es que a finales de marzo la cantidad de tweets registrados disminiye considerablemente llegando a ser nulo, mientras que la mayor cantidad de tweets se encuentra concentrada durante la tercera semana de marzo y la segunda semana de abril.

Un campo importante que posee esta base de datos es la columna “Sentiment”, la cual representa el sentimiento asociado al tweet registrado. A continuación se presenta la cantidad de tweets segun tipo de sentimiento (“Extremely Negative”, “Extremely Postive”, “Negative”, “Neutral”, “Positive”) durante el periodo registrado en el set de datos.

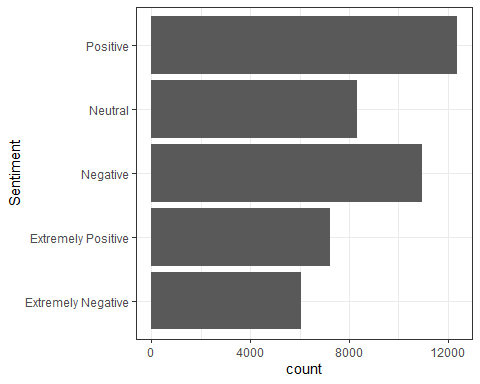
tweets\_mes\_dia <- df %>% mutate(mes\_dia = format(TweetAt, "%m-%d"))  
tweets\_mes\_dia %>% group\_by(Sentiment, mes\_dia) %>% summarise(n = n()) %>%  
 ggplot(aes(x = mes\_dia, y = n, color = Sentiment)) +  
 geom\_line(aes(group = Sentiment)) +  
 labs(title = "Número de tweets publicados", x = "fecha de publicación",  
 y = "número de tweets") +  
 theme\_bw() +  
 theme(axis.text.x = element\_text(angle = 90, size = 6),  
 legend.position = "bottom")



Se puede observar en este último gráfico que en general aquellos sentimientos que prevalecen son los de “Positive” y “Negative”. Por otro lado, todos los tipos de sentimientos registrados tienen un comportamiento similar a lo largo del tiempo. Mientras que el sentimiento menos registrados en el tiempo corresponde al de “Extremely Negative”, lo cual es interesante dado el contexto de la base de datos, en donde se esperaría que hubiese una mayor cantidad de tweets asociado a un sentimiento negativo o extramademente negativo.

En el siguiente gráfico se puede observar la cantidad de tweets por cada tipo de sentimiento durante todo el periodo de observación.

# library(tidyverse)  
#tweets\_sentiment <- df %>% group\_by(Sentiment) %>% summarise(n = n())  
df %>% ggplot(aes(x = Sentiment)) + geom\_bar(stat="count") + coord\_flip() + theme\_bw()



Se observa que tal y como se mencionó anteriormente, los sentimientos que destacan son “Positive” y “Negative”, alcanzando un total aproximado de 13000 y 11000 tweets respectivamente. El sentimiento con una menor cantidad de registros es “Extremely Negative”, con un aproximado de 6000 tweets.

Si se comparación la proporción de estos tweets en comparación al total del periodo de observación, se tiene que el porcentaje asociado a cada sentimiento son los siguientes:

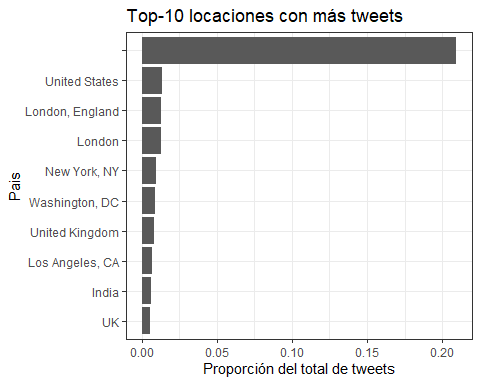
df %>% group\_by(Sentiment) %>% summarise(Proporcion = n()/nrow(df)) %>% arrange(-Proporcion)

## # A tibble: 5 x 2  
## Sentiment Proporcion  
## <chr> <dbl>  
## 1 Positive 0.275  
## 2 Negative 0.244  
## 3 Neutral 0.185  
## 4 Extremely Positive 0.161  
## 5 Extremely Negative 0.135

Es decir, el sentimiento “Positive” representa al 27,5% de los tweets del set de datos, mientras que el 13,5% de los tweets están categorizados bajo el sentimiento “Extremely Negative”.Es importante notar que los sentimientos extremos tanto positivo como negativo se presentan en menor proporcion.

Otro punto importante a caracterizar es la locación en donde se emiten los tweets. En la siguiente tabla, se puede observar la cantidad total de tweets registrados para el top 10 de localidades.

top\_10 <- df %>% group\_by(Location) %>% summarise(N=n()/nrow(df)) %>% arrange(-N)  
top\_10 <- top\_10[1:10,]  
top\_10$N <- round(top\_10$N,4)  
ggplot(top\_10,aes(x=reorder(Location,N), y =N )) +   
 geom\_bar(stat="identity") +   
 coord\_flip() +   
 labs(x="Pais", y = "Proporción del total de tweets", title="Top-10 locaciones con más tweets") +  
 theme\_bw()



En este último gráfico se puede observar que un 20% de los tweets no registran alguna locación, lo que se puede atribuir a que dentro de la aplicacion de twitter no todas las personas comparten su ubicacion. Del porcentaje restante, los lugares más comunes son de países como Estados Unidos, Reino Unido e India. Se puede observar tambien que la estandarizacion de las localidades no es la mejor, pues no siempre estan relacionadas a un pais de origen.

Para poder analizar los contenidos de los tweets es necesario realizar una limpieza y normalización de estos. A continuación se crea una función que permite corregir algunos patrones de los textos tales como números, puntuación y espacios en blanco.

limpiar\_texto <- function(texto){  
 # Se convierte todo el texto a minúsculas  
 nuevo\_texto <- tolower(texto)  
 # Eliminación de páginas web (palabras que empiezan por "http." seguidas   
 # de cualquier cosa que no sea un espacio)  
 nuevo\_texto <- str\_replace\_all(nuevo\_texto,"http\\S\*", "")  
 # Eliminación de signos de puntuación  
 nuevo\_texto <- str\_replace\_all(nuevo\_texto,"[[:punct:]]", " ")  
 # Eliminación de números  
 nuevo\_texto <- str\_replace\_all(nuevo\_texto,"[[:digit:]]", " ")  
 # Eliminación de espacios en blanco múltiples  
 nuevo\_texto <- str\_replace\_all(nuevo\_texto,"[\\s]+", " ")  
 # Tokenización por palabras individuales  
 nuevo\_texto <- str\_split(nuevo\_texto, " ")[[1]]  
 # Eliminación de tokens con una longitud < 2  
 nuevo\_texto <- keep(.x = nuevo\_texto, .p = function(x){str\_length(x) > 1})  
 return(nuevo\_texto)  
}

Para entender como procede esta última función, en la siguiente línea se muestra un ejemplo junto con los resultados de la aplicación de esta, en donde se puede observar que se extrajo cada palabra del objeto *text*.

text = "Hola mi nombre es https://www.google.cl. Como. no sé xd6666 ASDA"  
limpiar\_texto(text)

## [1] "hola" "mi" "nombre" "es" "como" "no" "sé" "xd"   
## [9] "asda"

En el siguiente paso, se aplica la función *limpiar\_texto* al contenido de los tweets de la base de datos, en donde cada resultado de cada tweet es almacenado en un vector de palabras, por lo que cada tweet tendría asociado uno de estos vectores.

tweets <- df %>% mutate(texto\_vector = map(.x = OriginalTweet, .f = limpiar\_texto))

tweets %>% select(texto\_vector) %>% head()

## texto\_vector  
## 1 menyrbie, phil, gahan, chrisitv, and, and  
## 2 advice, talk, to, your, neighbours, family, to, exchange, phone, numbers, create, contact, list, with, phone, numbers, of, neighbours, schools, employer, chemist, gp, set, up, online, shopping, accounts, if, poss, adequate, supplies, of, regular, meds, but, not, over, order  
## 3 coronavirus, australia, woolworths, to, give, elderly, disabled, dedicated, shopping, hours, amid, covid, outbreak  
## 4 my, food, stock, is, not, the, only, one, which, is, empty, please, don, panic, there, will, be, enough, food, for, everyone, if, you, do, not, take, more, than, you, need, stay, calm, stay, safe, covid, france, covid, covid, coronavirus, confinement, confinementotal, confinementgeneral  
## 5 me, ready, to, go, at, supermarket, during, the, covid, outbreak, not, because, paranoid, but, because, my, food, stock, is, litteraly, empty, the, coronavirus, is, serious, thing, but, please, don, panic, it, causes, shortage, coronavirusfrance, restezchezvous, stayathome, confinement  
## 6 as, news, of, the, regionâ, first, confirmed, covid, case, came, out, of, sullivan, county, last, week, people, flocked, to, area, stores, to, purchase, cleaning, supplies, hand, sanitizer, food, toilet, paper, and, other, goods, tim, dodson, reports

#Cada valor de la columna texto\_vector es un vector con cada palabra del texto  
tweets$texto\_vector[1]

## [[1]]  
## [1] "menyrbie" "phil" "gahan" "chrisitv" "and" "and"

En donde cada valor de la columna texto\_vector es un vector con cada palabra del texto

#unnest() nos permite realizar una expansión de los vectores de palabras que creamos, esto aumenta la dimension de filas considerablemente  
tweets\_expand <- tweets %>% select(-OriginalTweet) %>% unnest()

## Warning: `cols` is now required when using unnest().  
## Please use `cols = c(texto\_vector)`

tweets\_expand <- tweets\_expand %>% rename(word = texto\_vector)  
head(tweets\_expand)

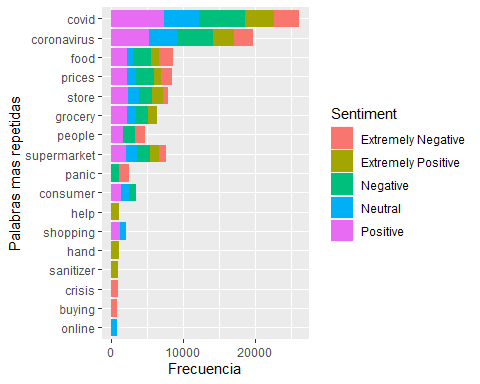
## # A tibble: 6 x 6  
## UserName ScreenName Location TweetAt Sentiment word   
## <int> <int> <chr> <date> <chr> <chr>   
## 1 3799 48751 London 2020-03-16 Neutral menyrbie  
## 2 3799 48751 London 2020-03-16 Neutral phil   
## 3 3799 48751 London 2020-03-16 Neutral gahan   
## 4 3799 48751 London 2020-03-16 Neutral chrisitv  
## 5 3799 48751 London 2020-03-16 Neutral and   
## 6 3799 48751 London 2020-03-16 Neutral and

Se utilizan *stopwords* para filtrar algunas palabras propias del inglés (lenguaje dominio de los comentarios) como artículos, pronombres, preposiciones, adverbios e incluso algunos verbos. Palabras que no tienen un significado por sí solas, sino que modifican o acompañan a otras.

# "word" %in% vector -> true or false  
lista\_stopwords <- stopwords("english")  
lista\_stopwords <- c(lista\_stopwords, "amp","can")

Luego se representa en un gráfico de barras sobre las 10 palabras más repetidas en los comentarios según el sentimiento asignado al Tweet en el que se encuentran.

tweets\_expand <- tweets\_expand %>% filter(!(word %in% lista\_stopwords))   
  
tweets\_expand %>% group\_by(Sentiment, word) %>%   
 count(word) %>%   
 group\_by(Sentiment) %>%   
 top\_n(10,n) %>%   
 arrange(Sentiment, desc(n)) %>%   
 ggplot(aes(x=reorder(word,n),y=n,fill=Sentiment)) +   
 geom\_col() +   
 labs(y = "Frecuencia", x = "Palabras mas repetidas") +   
 coord\_flip()



Se puede observar que, como era de esperarse, las palabras más comentadas en todas las categorias de sentimiento son las referentes directamente a la pandemia: “covid” y “coronavirus”. Se destaca de igual forma la alta popularidad de las palabras “food” y “prices”, posiblemente debido al parcial desabastecimiento de productos y la subida de precios producto de la cuarentena.

Un análisis similar se presenta a continuación, pero esta vez en nubes de palabras:

#Listas de colores utilizadas en las nubes de palabras  
pal\_neg <- c("#FC9272", "#FB6A4A", "#EF3B2C", "#CB181D", "#A50F15")  
pal\_neu <- c("#A1D99B", "#74C476", "#41AB5D", "#238B45", "#006D2C")  
pal\_pos <- c("#9ECAE1", "#6BAED6", "#4292C6", "#2171B5", "#08519C")  
  
most\_frec <- c("covid","coronavirus","supermarket","grocery","store")  
  
top <- 400

#Palabras más repetidas en comentarios negativos y extremadamente negativos  
neg\_tweets <- tweets\_expand %>%   
 filter(Sentiment == "Negative" | Sentiment == "Extremely Negative") %>%  
 count(word)  
neg\_tweets\_top <- neg\_tweets%>%   
 arrange(desc(n)) %>%  
 top\_n(top,n)   
  
  
set.seed(1234)  
wordcloud(words = neg\_tweets\_top$word, scale=c(5,0.7), freq = neg\_tweets\_top$n, min.freq = 1,max.words=100, random.order=FALSE, rot.per=0.35, colors=pal\_neg)

## Warning in wordcloud(words = neg\_tweets\_top$word, scale = c(5, 0.7), freq =  
## neg\_tweets\_top$n, : government could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = neg\_tweets\_top$word, scale = c(5, 0.7), freq =  
## neg\_tweets\_top$n, : every could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = neg\_tweets\_top$word, scale = c(5, 0.7), freq =  
## neg\_tweets\_top$n, : economy could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = neg\_tweets\_top$word, scale = c(5, 0.7), freq =  
## neg\_tweets\_top$n, : weeks could not be fit on page. It will not be plotted.

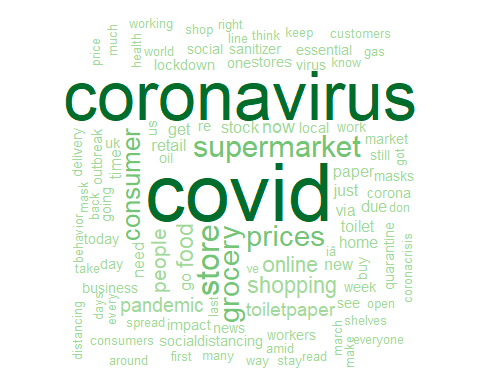
## Warning in wordcloud(words = neg\_tweets\_top$word, scale = c(5, 0.7), freq  
## = neg\_tweets\_top$n, : supermarkets could not be fit on page. It will not be  
## plotted.



#Palabras más repetidas en comentarios neutrales  
neu\_tweets <- tweets\_expand %>%   
 filter(Sentiment == "Neutral") %>%  
 count(word)  
neu\_tweets\_top <- neu\_tweets%>%  
 arrange(desc(n)) %>%  
 top\_n(top,n)   
   
  
set.seed(1234)  
wordcloud(words = neu\_tweets\_top$word, scale=c(5,0.7), freq = neu\_tweets\_top$n, min.freq = 1,max.words=100, random.order=FALSE, rot.per=0.35, colors=pal\_neu, fixed.asp = TRUE)

## Warning in wordcloud(words = neu\_tweets\_top$word, scale = c(5, 0.7), freq =  
## neu\_tweets\_top$n, : coronaviruspandemic could not be fit on page. It will not be  
## plotted.

## Warning in wordcloud(words = neu\_tweets\_top$word, scale = c(5, 0.7), freq =  
## neu\_tweets\_top$n, : stayhome could not be fit on page. It will not be plotted.



#Palabra más repetidas en comentarios positivos y extremadamente positivos  
pos\_tweets <- tweets\_expand %>%   
 filter(Sentiment == "Positive" | Sentiment == "Extremely Positive") %>%  
 count(word)  
pos\_tweets\_top <- pos\_tweets%>%   
 arrange(desc(n)) %>%  
 top\_n(top,n)   
  
  
set.seed(1234)  
wordcloud(words = pos\_tweets\_top$word, scale=c(5,0.7), freq = pos\_tweets\_top$n, min.freq = 1,max.words=100, random.order=FALSE, rot.per=0.35, colors=pal\_pos, fixed.asp = TRUE)

## Warning in wordcloud(words = pos\_tweets\_top$word, scale = c(5, 0.7), freq =  
## pos\_tweets\_top$n, : socialdistancing could not be fit on page. It will not be  
## plotted.

## Warning in wordcloud(words = pos\_tweets\_top$word, scale = c(5, 0.7), freq =  
## pos\_tweets\_top$n, : much could not be fit on page. It will not be plotted.

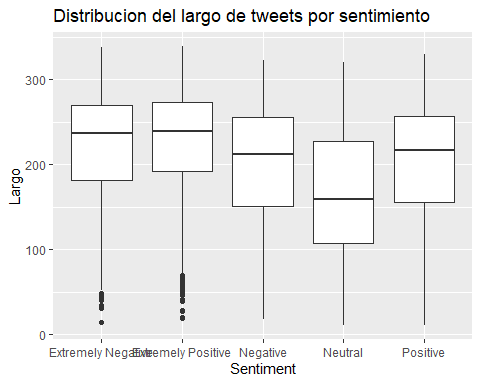
## Warning in wordcloud(words = pos\_tweets\_top$word, scale = c(5, 0.7), freq =  
## pos\_tweets\_top$n, : times could not be fit on page. It will not be plotted.



Se puede destacar que el tipo de palabras utilizadas en los 3 contextos de sentimiento (negativo, neutral y positivo) no varía en gran manera, repitiendose típicamente las mismas dentro de las mas populares: “covid”, “coronavirus”, “supermarket”, “food”, “prices”, “store” y “grocery”. Dejando de lado la palabra “covid” las palabras más repetidas se relacionan con el abastecimiento de productos basicos, lo que permite inferir de forma preliminar que este tema fue relevante para los usuarios durante la pandemia (entre marzo y abril).

En base a los datos obtenidos categorizados por sentimiento, puede analizarse su largo promedio, y así ver si existe alguna relación entre esas variables. Para esto se hace uso de los boxplot, con el fin de comparar promedios y distribuciones, y detectar algunos outliers segun el sentimiento. Es importante destacar aquí la presencia de hashtags (#), que suponen una mayor extensión al largo del tweet.

#Promedio en el largo de tweets por sentimiento.  
  
ggplot(tweets, aes(x=Sentiment, y =str\_length(OriginalTweet) ),) + geom\_boxplot() + labs(y = "Largo", title= "Distribucion del largo de tweets por sentimiento")



ggsave(plot = last\_plot(), width = 10, height = 10, dpi = 300, filename = "boxplot\_len\_words.png")

Se puede observar que el largo de las palabras no parece ser determinante para el valor de sentimiento del comentario en general, pues en promedio todas miden muy parecido: aproximadamente 8 caracteres.

# Propuesta de hipótesis

A partir del análisis exploratorio se proponen las siguientes hipótesis y preguntas que se podrían abordar con este set de datos:

1. ¿El sentimiento general sobre el COVID-19 varía por la localidad registrada?
2. Dada las características del COVID-19 y sus consecuencias, más del 50% de los tweets están asociados a un sentimiento negativo o extremadamente negativo.
3. ¿Los sentimientos mayormente expresados en los Tweets tienen relación con el contexto social del lugar desde donde son publicados?
4. ¿Se puede asociar un sentimiento a una palabra dependiendo de las otras palabras mencionadas en un tweet?
5. ¿Existe un alza o descenso de comentarios positivos al avanzar los días?¿Si es así, a qué se debe?
6. ¿Se puede generar un clasificador que se pueda generalizar en base a los datos que tenemos?
7. ¿El sentimiento de los comentarios se ha visto modificado con respecto a tiempos anteriores al Covid?

# Proximos pasos.

Como se mencionó antes, los datos manejados no tienen un estándar en la categorización por localidad, por lo que resulta muy importante tratar con nuevos dataset para agrupar mejor esas localizaciones.

Es necesario también mejorar el filtro de palabras, esta vez considerando el uso de hashtags, para realizar un análisis más detallado acerca de la correlación entre la extensión de estas y el sentimiento asociado al tweet.

Y para entrar a comparar el sentimiento promedio de tweets en distintas fechas, se requerirá elaborar y extender un algorítmo de clasificación a nuevos dataset.

# Referencias

[1] Twitter suma 166 millones de usuarios durante el Coronavirus. (2020, 3 junio). REBOLD, Data-Driven Marketing & Communication. <https://letsrebold.com/es/blog/twitter-suma-166-millones-de-usuarios-frente-al-coronavirus/#:%7E:text=Asimismo%2C%20Twitter%20comunic%C3%B3%20en%20la,el%20primer%20trimestre%20de%202019>.

[2] Coronavirus tweets NLP - Text Classification. (2020, 8 septiembre). Kaggle. <https://www.kaggle.com/datatattle/covid-19-nlp-text-classification?select=Corona_NLP_test.csv>

# Contribuciones del equipo

Existen tareas que fueron realizadas en conjunto por el grupo, tales como la selección del tema de interés y la búsqueda de bases de datos que se ajustarán a los requerimientos del grupo. A continuación se detallan las tareas que realizó cada miembro del equipo:

1. Nicolás Herrera: Construcción de presentación, grabación, edición del vídeo, construcción de tablas y análisis de los datos.
2. Yesenia Marulanda: Redacción de introducción, motivación e hipótesis del trabajo y construcción de presentación.
3. Franco Migliorelli: Redacción de introducción, motivación y análisis de gráficos generados, además de redacción de próximos pasos.
4. Samuel Sánchez: Análisis de gráficos, tratamiento de datos y redacción de hipótesis y preguntas.
5. Sebastián Urbina: Creación de gráficos, limpieza y tratamientos de datos.