Universidad del Valle de Guatemala
Lynnet Garcia
Jose Ramirez
Josue David Lopez Florian

3 de septiembre 2020
Data Science
carne:15441
carne: 17081

Laboratorio 6:
Análisis de redes sociales

Caso 1 : COVID-19 en Guatemala.

(15 puntos) Extracción de información:
- Está documentado el proceso de extracción de los datos de las redes sociales. Se extraen suficientes datos para descubrir información.

Se planea obtener la información de COVID-19 en Guatemala a partir de Twitter. Por tanto, se aplicó a una cuenta de desarrollador dentro de Twitter, esta cuenta de desarrollador nos otorgará el acceso a un API. Para facilitar el uso de esta API se tiene planeado utilizar librerías como Tweepy en caso de usar Python.

Luego, se realizará una búsqueda de tweets relacionados con el COVID-19 en Guatemala. Para ello se hará una búsqueda utilizando palabras claves o hashtags que nos proporcionen dicha información. Se utilizaran palabras clave como:

- CoronavirusGT
- Covid19GT
- CovidGT
- CovidGuatemala

(25 puntos) Análisis exploratorio:
- Se elaboró un análisis exploratorio en el que se explican los cruces de variables, hay gráficos explicativos y análisis que permiten comprender el conjunto de datos.

Las variables que tenemos en el Dataset son las siguientes:

- ID: identificador del tweet
- Date: fecha de publicación del tweet y la hora
- Tweet: texto del tweet publicado
- Favorites: la cantidad de favoritos dados al tweet
- Retweets: la cantidad de retweets dados al tweet

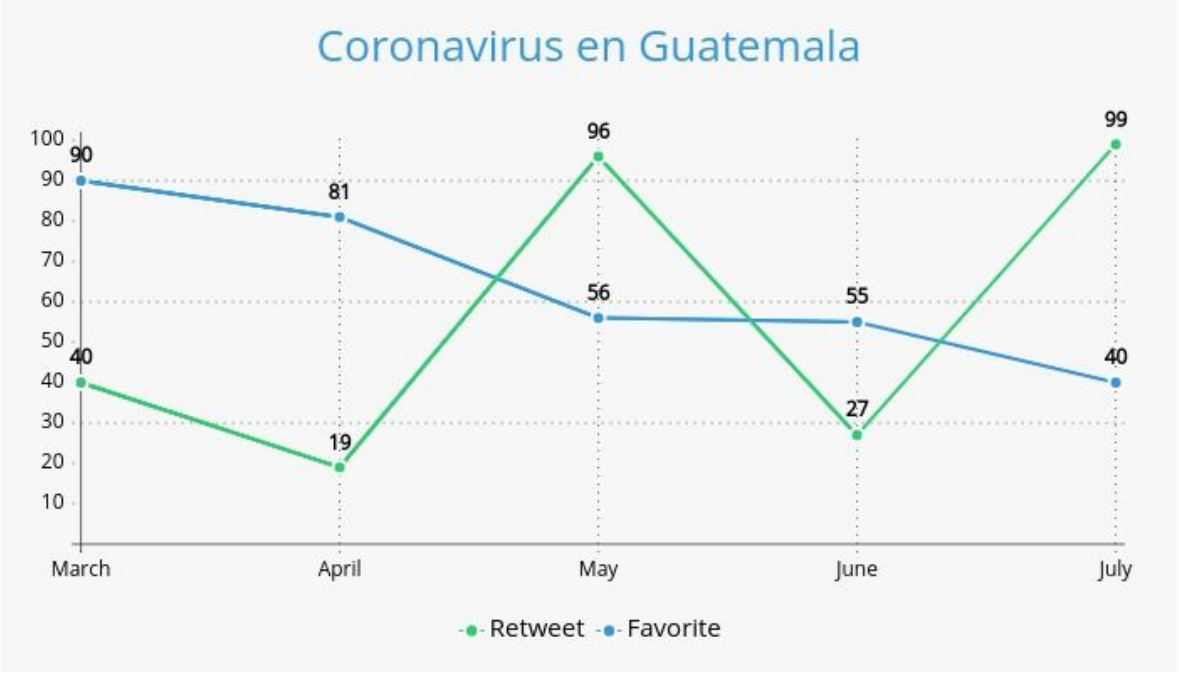

Figura 1: Gráfica de líneas sobre la correlación de Favoritos y Retweets a lo largo del tiempo

(20 puntos) Limpieza y preprocesamiento de los datos:
- Se documentan las tareas de limpieza, incluyendo los paquetes/módulos que se usaron.

Actualmente se tiene pensado utilizar el paquete de tweepy, página principal del paquete: [https://www.tweepy.org/].

La información obtenida por el paquete viene en formato json. Las parejas llave-valor que no proporcionen información relevante para nuestro análisis serán eliminadas. Por ejemplo, si el usuario es verificado o no, si existe mención o etiqueta a otro usuario, entre otros.

En R se llevó a cabo la limpieza de los textos en la columna de Tweets. Se llevó a cabo con el uso de las siguientes librerías:

```
library(stringi)
library(dplyr)
library(qdapRegex)
library(readr)
library("stringr")
install.packages("qdapRegex")
install.packages("stopwords")
library("stopwords")
library('tm')
```

Después de recolectar la data y de asignarla a una variable, con la cual se hará la limpieza del texto, se procedió a realizar los siguientes cambios:

- Limpiar las tildes de los tweets.(Aquí se reemplazaron los pares símbolos que eran su equivalente en ASCII)
- Volver todo el texto a minúsculas

- eliminar las Urls de los tweets y los urls que conectaban a otros tweets
- se eliminaron los stop words en español
- se hizo la eliminación de emoticones
- se eliminaron los caracteres especiales y los signos de puntuación
- quitamos caracteres especiales de la codificación, como saltos de línea y tabulaciones
- eliminamos el exceso de espacios en blancos
- se eliminaron los duplicados de la data

Esto fue lo que se hizo sobre la data que teníamos y solo fue modificada la columna de tweets, ya que era la única que requería de limpieza para trabajar con ella.

(30 puntos) Descubrimiento de información.
- Se describe de forma comprensible y amena los descubrimientos encontrados. Se calcular algunas métricas. Se detalla la información que es interesante.

Uno de los descubrimientos que saltan a la vista cuando vemos la figura 2 es que las palabras más usadas son coronavirusgt, guatemala y coronavirus. Estas 3 palabras lo más posible es que estén relacionadas a los hashtags que usa la gente a la hora de hacer sus tweets. lo que sí vemos es que se repite la palabra casos y Giammattei, lo cual nos habla que lo más mencionado y en tendencias están los casos nuevos y lo dicho por el presidente. podríamos creer que la situación del covid en Guatemala se hace de mayor interés para la gente cuando el presidente hace sus cadenas nacionales o se discute sobre el aumento de los caso de covid en Guatemala.

las figuras 3 y 4 nos muestra que en las nubes de palabras hay otras que también se repiten como nuevos, muertes, pruebas, salud y personas, lo cual nos da una idea que este sea uno de los tópicos más hablados. Según los visto en las gráficas podemos deducir que algo muy hablado en Guatemala sobre el tema del covid es el estado actual de los nuevos casos, las muertes ocurridas, las pruebas realizadas y sobre el sistema de salud o la salud de la gente.

Algo que nos da una buena retroalimentación sobre cómo se relacionan las palabras que vemos en las figuras 2, 3 y 4, son las figuras 5 y 6. Esto porque nos dan una clara forma de ver la relación entre palabras y grupos de palabras. Podemos ver como las palabras guatemala y coronavirus está directamente relacionada a casos, por lo que nos confirma la teoría de ser de lo que más se habla en el momento al mencionar los casos. También vemos como drgiammattei y statscovid19gt están directamente relacionados junto a la palabra más. Esto nos podría confirmar el aumento de tweets acerca del tema cuando el presidente habla sobre las estadísticas y el estado actual del covid en el país.

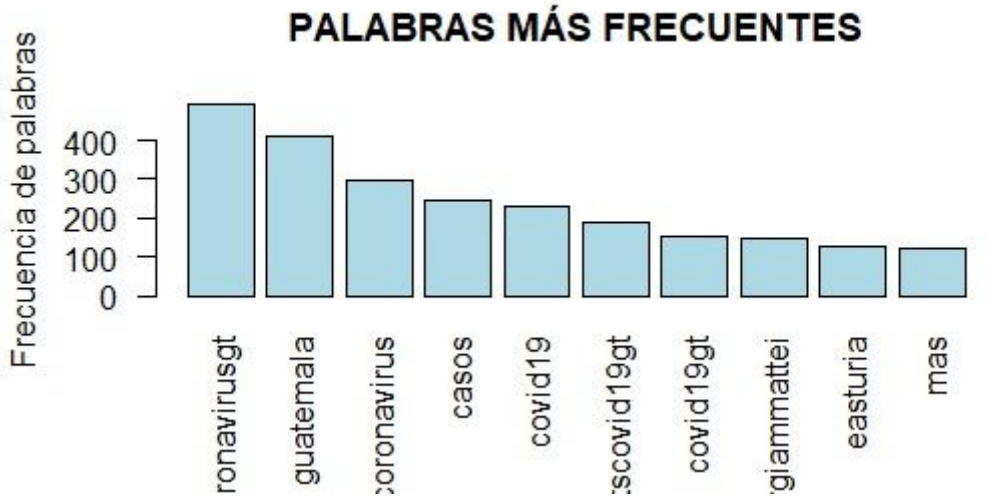

Figura 2: Histograma de palabras sobre tweets relacionados con el coronavirus

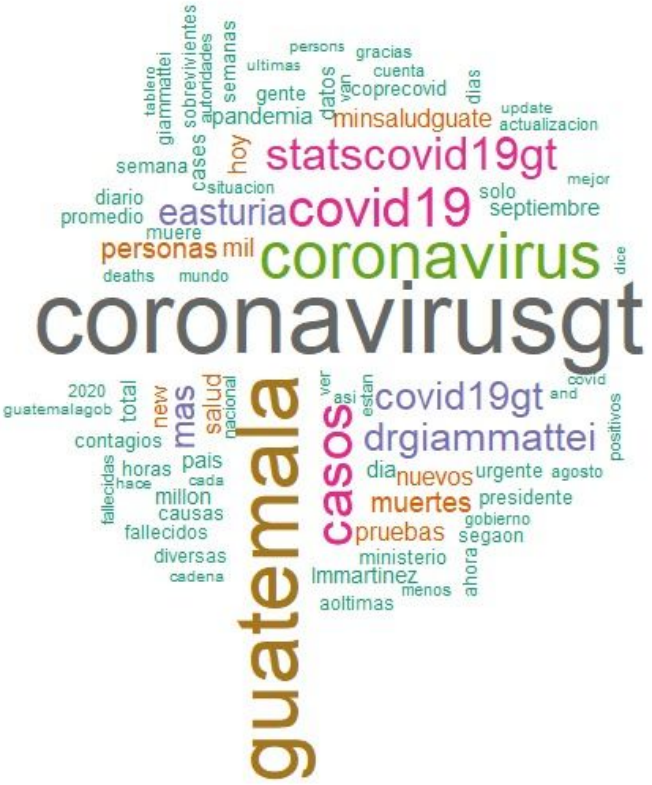

Figura 3: Nube 1 de palabras sobre tweets relacionados con el coronavirus.

Figura 4: Nube 2 de palabras sobre tweets relacionados con el coronavirus.

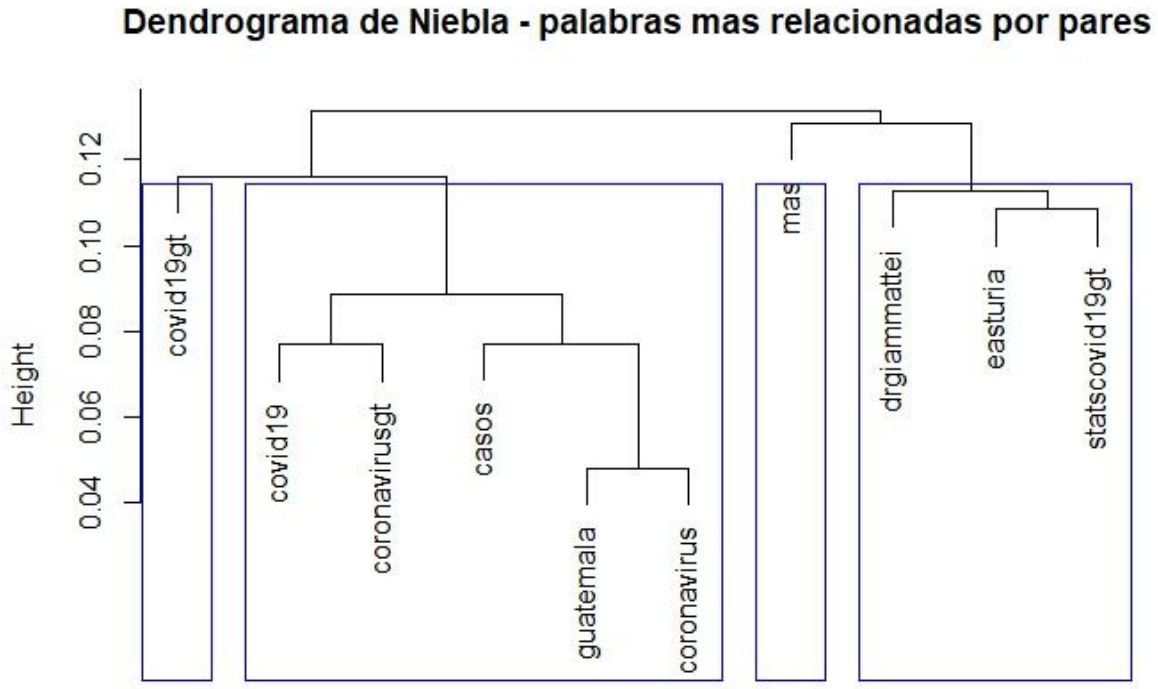

Figura 5: Dendrograma de Niebla de palabras relacionadas por pares de los tweets relacionados con el coronavirus.

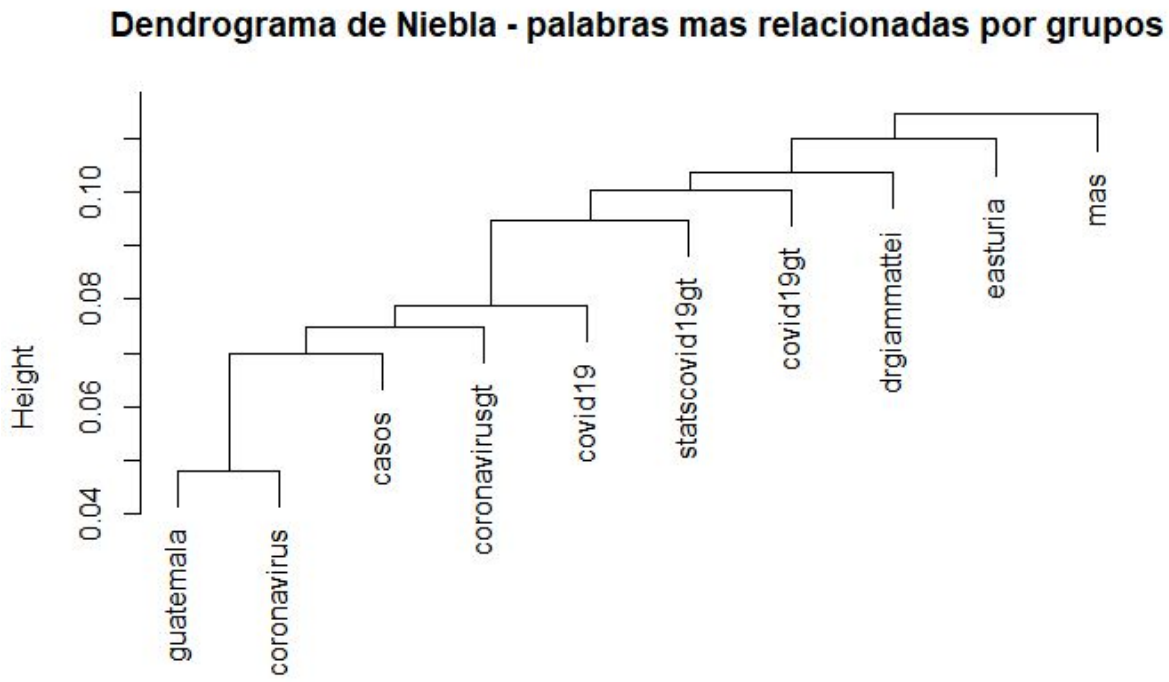

Figura 6: Dendrograma de Niebla de palabras relacionadas por grupos de los tweets relacionados con el coronavirus

(10 puntos) Conclusiones.

- Dados los resultados encontrados podemos observar que existe cierta cantidad de palabras que se repiten en los tweets con cierta relación entre ellos.
- Estos tweets mencionan al presidente, el país y la enfermedad, lo cual visto en los resultados obtenidos sabemos que puede estar relacionado a las cadenas nacionales.
- Fue interesante observar como la cantidad de favoritos y tweets no mantienen una relación.
- Se pudo deducir que uno de los temas más hablados en twitter con relación al Covid19 en nuestro país, se refiere a las estadísticas y situación actual.