Rodrigo samayoa 17332
Diego Sevilla 17238
Alejandro Tejada 17584

Guatemala, septiembre del 2020
Data Science
Lynette Garcia

# Laboratorio no. 6

Análisis de redes sociales

## Los pasos involucrados para el analisis de texto fueron:

***Removidos:***

- columnas vacias
- URLs.
- caracteres especiales: “~@#$%&-_=<>”.
- emojis.
- puntuación, acentos, tildes.
- números / dígitos
- espacios en blanco

***Procedimientos:***

- Texto a  minúsculas
- Dividir líneas en palabras según ""
- Crear vector con las palabras
- Vector limpio de palabras
- Vector a data frame para facilitar el análisis y el uso de gráficos
- Eliminar palabras con menos de 3 letras
- Crear data frame con número total de palabras significativas
- Gráficos de frecuencia de las palabras
- Nube de palabras

***Problema a analizar***

**PROBLEMA 2.**

Se tomaron los datos del hasthag de tweeter #TraficoGT y  @amilcarmontejo. Para conocer tendencias y elementos interesantes del tráfico en la ciudad de Guatemala.

## Análisis general de 1000 tweets

### Gráfico de introducción del dataset

Rodrigo samayoa 17332
Diego Sevilla 17238
Alejandro Tejada 17584

Guatemala, septiembre del 2020
Data Science
Lynette Garcia


## Estructura del dataset

```
> str(data)
'data.frame':   1000 obs. of  10 variables:
 $ X             : int  1 2 3 4 5 6 7 8 9 10 ...
 $ text          : chr  " @soy_:  mltiple colisin en el anillo perifrico complica movilidad\n\nel acci
dente fue protagonizado por varios automviles" " @melvinnoguera: las ciclovas son as... no pintura en
 el pavimento. si van a hacer algo, hganlo bien. las motos"| __truncated__ " @melvinnoguera: las cicl
ovas son as... no pintura en el pavimento. si van a hacer algo, hganlo bien. las motos"| __truncated_
_ "ya hay trfico de motos en el carril no existente.   " ...
 $ favorited     : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ favoriteCount: int  0 0 0 0 0 0 0 0 0 0 ...
 $ created       : chr  "2020-09-04 23:33:59" "2020-09-04 23:28:14" "2020-09-04 23:24:51" "2020-09-04
 23:22:43" ...
 $ truncated     : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
 $ retweetCount : int  3 22 22 0 0 22 0 22 22 22 ...
 $ isRetweet     : logi  TRUE TRUE TRUE FALSE FALSE TRUE ...
 $ retweeted     : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ created2      : chr  "2020-09-04 17:33:59" "2020-09-04 17:28:14" "2020-09-04 17:24:51" "2020-09-04
 17:22:43" ...
```

Rodrigo samayoa 17332
Diego Sevilla 17238
Alejandro Tejada 17584

Guatemala, septiembre del 2020
Data Science
Lynette Garcia

## Resumen de los datos

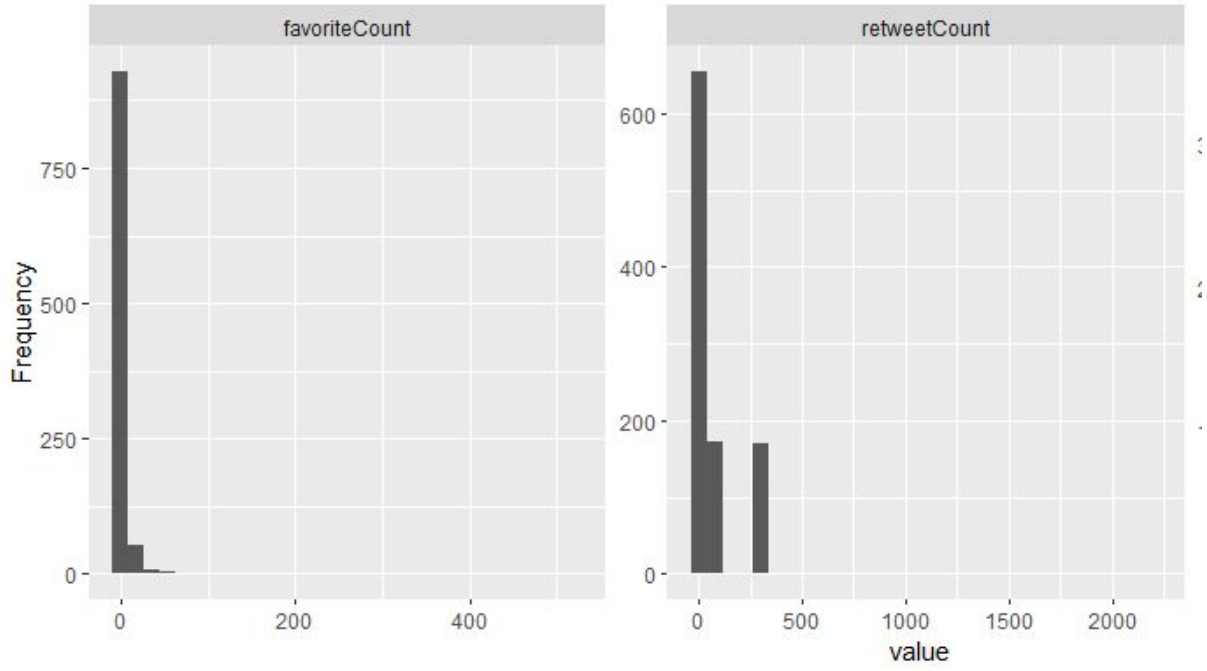
```
> summary(data)
       X              text             favorited       favoriteCount
 Min.   :   1.0   Length:1000        Mode :logical   Min.   :  0.000
 1st Qu.: 250.8   Class :character   FALSE:1000      1st Qu.:  0.000
 Median : 500.5   Mode  :character                   Median :  0.000
 Mean   : 500.5                                     Mean   :  2.411
 3rd Qu.: 750.2                                     3rd Qu.:  0.000
 Max.   :1000.0                                     Max.   :521.000
   created          truncated       retweetCount     isRetweet       retweeted
 Length:1000        Mode :logical   Min.   :   0.0   Mode :logical   Mode :logical
 Class :character   FALSE:783       1st Qu.:   1.0   FALSE:329       FALSE:1000
 Mode  :character   TRUE :217       Median :   7.0   TRUE :671
                                    Mean   :  75.5
                                    3rd Qu.:  58.0
                                    Max.   :2199.0

   created2
 Length:1000
 Class :character
 Mode  :character
```

Rodrigo samayoa 17332
Diego Sevilla 17238
Alejandro Tejada 17584

Guatemala, septiembre del 2020
Data Science
Lynette Garcia

## Histogramas de las variables cuantitativas




## Conteo de tweets

```
> head(data$text)
[1] " @soy_:  mltiple colisin en el anillo perifrico complica movilidad\n\nel accidente fue
 protagonizado por varios automviles"
[2] " @melvinnoguera: las ciclovas son as... no pintura en el pavimento. si van a hacer alg
o, hganlo bien. las motos y los carros jams van"
[3] " @melvinnoguera: las ciclovas son as... no pintura en el pavimento. si van a hacer alg
o, hganlo bien. las motos y los carros jams van"
[4] "ya hay trfico de motos en el carril no existente.  "

[5] "entrate de cmo ser parte de la comunidad traeguate, la comunidad del jaln.\nuna comuni
dad solidaria que ha dejado "
[6] " @melvinnoguera: las ciclovas son as... no pintura en el pavimento. si van a hacer alg
o, hganlo bien. las motos y los carros jams van"
> intLineCount <- length(data$text)
> intLineCount
[1] 1000
```

Promedio de palabras por tweet

```
> mean(vciUNPrfWperL)
[1] 19.503
```

Conteo de palabras

```
> intWordCount
[1] 19503
```

Primer vector con palabras

Rodrigo samayoa 17332
Diego Sevilla 17238
Alejandro Tejada 17584

Guatemala, septiembre del 2020

Data Science

Lynette Garcia

```
> head(vcsUNPrfWords,100)
  [1] "soy"           "mltiple"       "colisin"       "en"            "el"
  [6] "anillo"        "perifrico"     "complica"      "movilidadel"   "accidente"
 [11] "fue"           "protagonizado" "por"           "varios"        "automviles"
 [16] "melvinnoguera" "las"           "ciclovas"      "son"           "as"
 [21] "no"            "pintura"       "en"            "el"            "pavimento"
 [26] "si"            "van"           "a"             "hacer"         "algo"
 [31] "hganlo"        "bien"          "las"           "motos"         "y"
 [36] "los"           "carros"        "jams"          "van"           "melvinnoguera"
 [41] "las"           "ciclovas"      "son"           "as"            "no"
 [46] "pintura"       "en"            "el"            "pavimento"     "si"
 [51] "van"           "a"             "hacer"         "algo"          "hganlo"
 [56] "bien"          "las"           "motos"         "y"             "los"
 [61] "carros"        "jams"          "van"           "ya"            "hay"
 [66] "trfico"        "de"            "motos"         "en"            "el"
 [71] "carril"        "no"            "existente"     "entrate"       "de"
 [76] "cmo"           "ser"           "parte"         "de"            "la"
 [81] "comunidad"     "traeguate"     "la"            "comunidad"     "del"
 [86] "jalnuna"       "comunidad"     "solidaria"     "que"           "ha"
 [91] "dejado"        "melvinnoguera" "las"           "ciclovas"      "son"
 [96] "as"            "no"            "pintura"       "en"            "el"
```

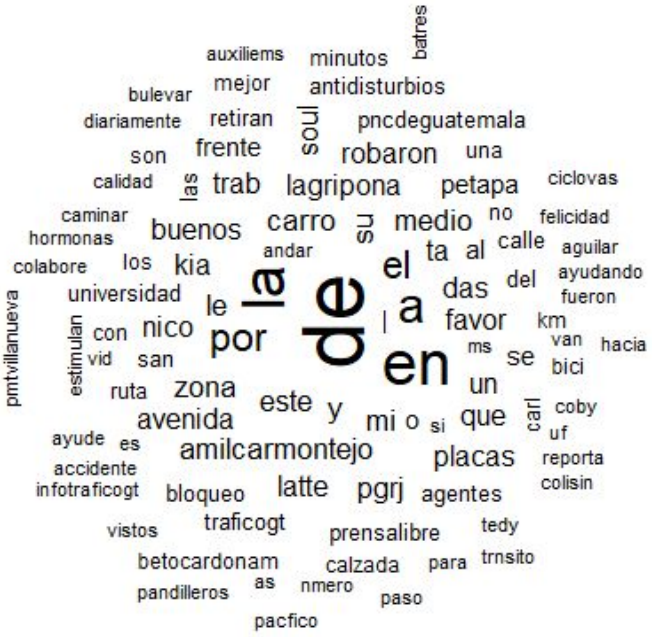
Ese vector a dataframe

```
> head(dfrUNPrfWords,10)
         words
1          soy
2       mltiple
3       colisin
4            en
5            el
6       anillo
7     perifrico
8      complica
9   movilidadel
10    accidente
```

Resumen

```
> head(dfrUNPrfFreq)
# A tibble: 6 x 2
  Words  Freq
  <chr> <int>
1 de      964
2 en      673
3 la      586
4 a       470
5 por     315
6 el      305
```

Primera nube de palabras (f>100)




Resumen de palabras significativas

Palabras con muy poca frecuencia antes y despues de ser removidas

Rodrigo samayoa 17332
Diego Sevilla 17238
Alejandro Tejada 17584

Guatemala, septiembre del 2020
Data Science
Lynette Garcia

```
> head(dfrUNPrfFreq)
# A tibble: 6 x 2
  Words           Freq
  <chr>          <int>
1 este             195
2 carro            190
3 amilcarmontejo   181
4 zona             178
5 medio            177
6 favor            171
```

```
# A tibble: 6 x 2
  Words          Freq
  <chr>         <int>
1 volante           1
2 volver            1
3 vuele             1
4 whatsapp          1
5 xinicosandra      1
6 youtube           1
```

```
> tail(dfrUNPrfFreq)
# A tibble: 6 x 2
  Words           Freq
  <chr>          <int>
1 todo               6
2 trabajos           6
3 transmetroguate    6
4 ufabuaufef         6
5 ufddos             6
6 vichoguate         6
```

Conteo de palabras significativas

```
> intWordCountFinal
[1] 300
```

Frecuencias comunes en el dataset

```
> head(dfrUNPrfFocf,10)
# A tibble: 5 x 2
  Fcat        Rfrq
  <ord>      <int>
1 "  10"      151
2 "  20"       59
3 "  50"       59
4 " 100"       13
5 " 500"       18
```

# RESUMEN PARA LOS 100 TWEETS

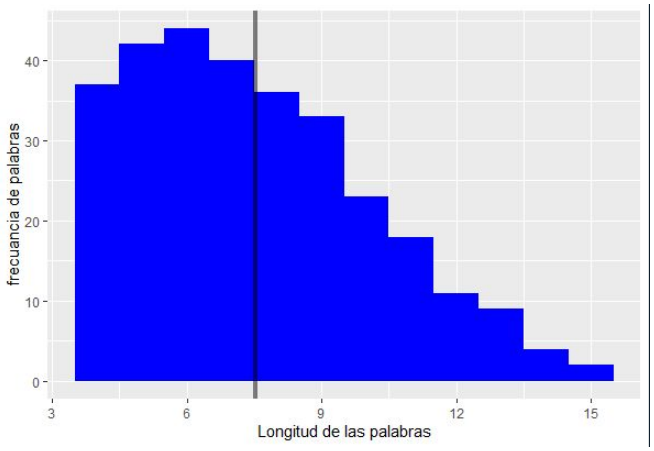
Nube de palabras significativas (f>50)

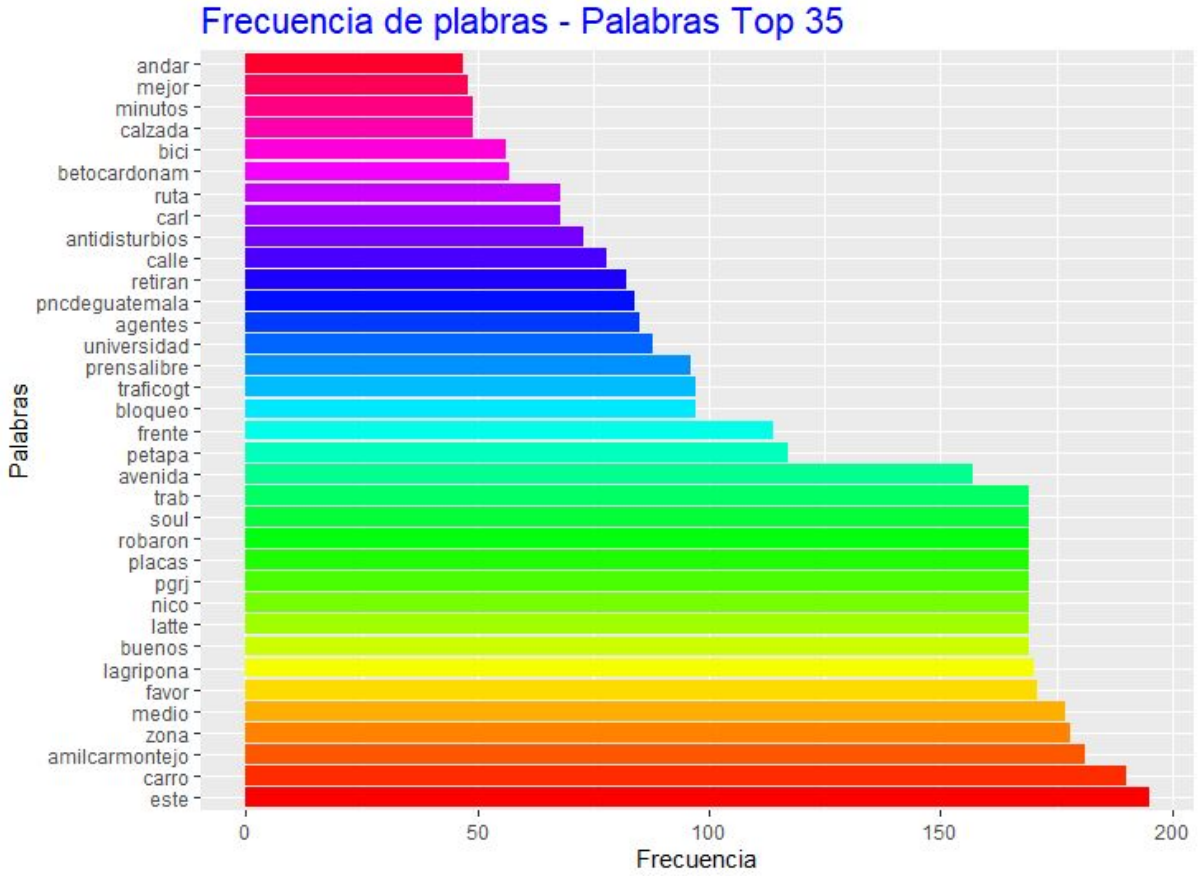
Rodrigo samayoa 17332
Diego Sevilla 17238
Alejandro Tejada 17584

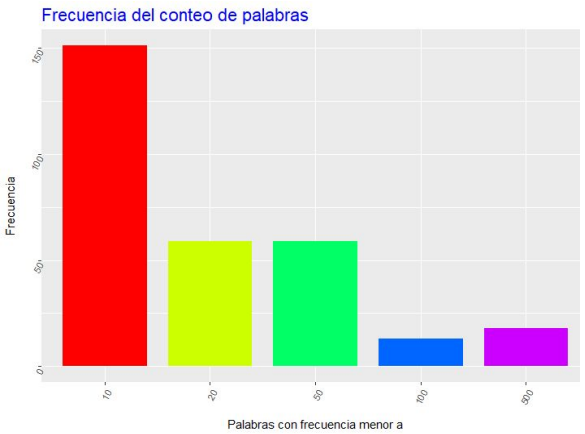
Guatemala, septiembre del 2020

Data Science

Lynette Garcia



Rodrigo samayoa 17332
Diego Sevilla 17238
Alejandro Tejada 17584

Guatemala, septiembre del 2020
Data Science
Lynette Garcia

# RESUMEN PARA TWEETS CON MÁS DE 100 RETWEETS

Tweets con mas retweets

```
> intLineCount
[1] 157
```

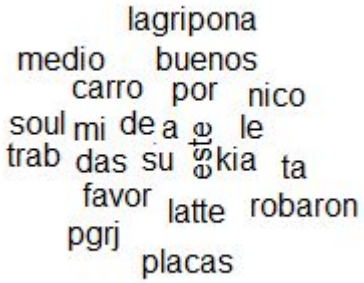
Palabras promedio de los tweets

```
> mean(vciUNPrfWperL)
[1] 23.92357
```

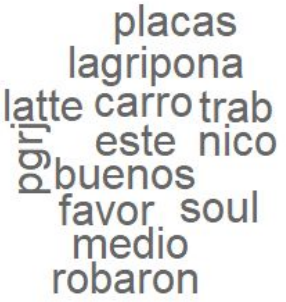
Conteo de palabras

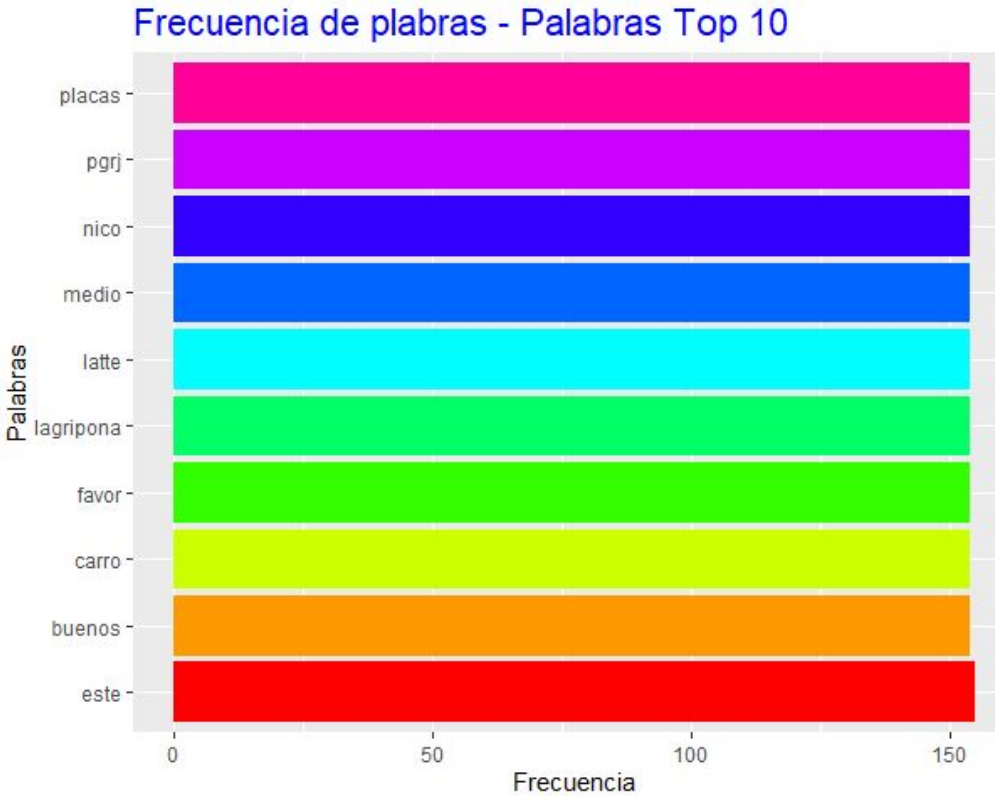
```
> intWordCount
[1] 3756
```

Nubes de palabras



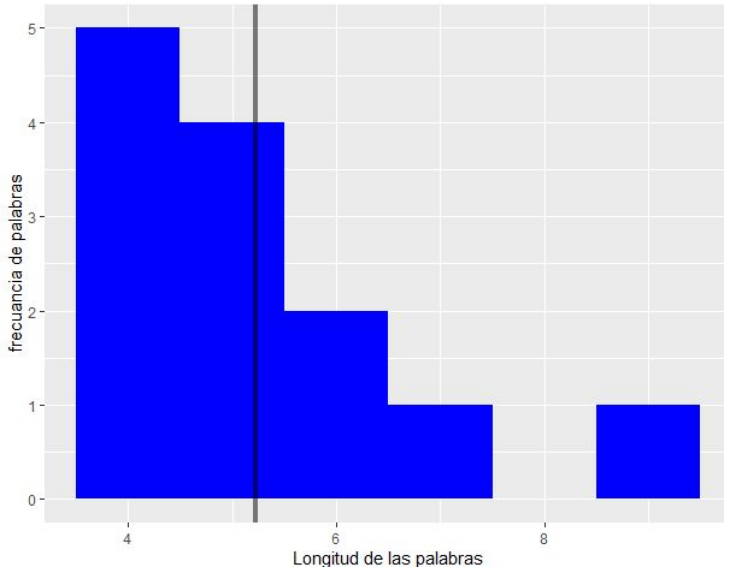

Nube de palabras significativas



Rodrigo samayoa 17332
Diego Sevilla 17238
Alejandro Tejada 17584

Guatemala, septiembre del 2020
Data Science
Lynette Garcia




# Discusión

Según los resultados observados para el análisis general de los tweets realizado es posible afirmar que muchos padecen de sueño al momento de hacer tráfico debido a que la palabra **latte** es una de las palabras más frecuentes en el espacio de tweets. Todos sabemos que una buena taza de café puede quitarnos el sueño y, en este caso, salvarnos de quedar debajo de un camión. Los **accidente**s no le gustan a nadie.

Es de esperarse encontrar la palabra **calzada** dentro del top 30 de las palabras más frecuentes en el análisis, ya que todas las calzadas de la ciudad de guatemala tienen fama de padecer bastante tráfico. Hay distintas calzadas con fama de tránsito lento en algunas horas del día como la Aguilar Batres, Roosevelt, La Paz. Sin embargo, también hay **avenida**s que llegan a tener mucho tránsito en ciertas horas como la avenida **Petata**. Cabe destacar que Petapa se encuentra en el top 20 de las palabras más frecuentes por lo que es acertado decir que esta es de las avenidas más transitadas en la ciudad de Guatemala.

Los universitarios conocen muy bien lo que es estar varados en el tráfico por horas para llegar a sus casas luego de estudiar con empeño y dedicación durante todo el día. Esto se multiplica para los dedicados estudiantes de data science. La palabra **universidad** se encuentra en el top 30 de las palabras más frecuentes, lo cual no es extraño porque los centros educativos, sean de nivel primario, secundario o de estudios superiores son focos de tránsito lento en las mañanas y en horas de la tarde.

Una palabra interesante es **lagripona.** Al parecer en esta época de tráfico han habido muchos conductores con síntomas de gripe al manejar. Esto no es del todo aceptable, ya que la gripe provoca agotamiento y puede llevar a causar accidentes. Puede asociarse al covid qué se ha vivido recientemente entre aglomeramientos de personas, por la necesidad de los ciudadanos de continuar con su empleo.

Rodrigo samayoa 17332 Guatemala, septiembre del 2020
Diego Sevilla 17238 Data Science
Alejandro Tejada 17584 Lynette Garcia

Algo común que ocurre en horas de tráfico lento son los asaltos. Los robos se dan cuando un conductor está distraído o simplemente porque el conductor debe detenerse por un semáforo. La palabra **robaron** se encuentra en el top 30.

Por último, a pesar que a muchos no nos gusta el tráfico y tratamos de evitarlo lo más posible la palabra **felicidad** se encuentra en la nube del top 50 de palabras con más frecuencia en el dataset. Puede parecernos extraño, ya que existen muchos conductores que pueden volverse irritables en situaciones como esta, pero este resultado nos dice lo contrario. Es muy probable, también, que esta palabra sea algo que siente el protagonista principal de los hechos de tránsito en la ciudad de Guatemala **Amilcar Montejo**, pero en este análisis no se ahondó tanto en los detalles para poder afirmarlo o negarlo con certeza.

# Conclusiones

1. Los centros educativos son focos de tránsito en ciertas horas del día.
2. Muchos emplean el café, más específicamente el latte, para mantenerse despiertos al momento de manejar.
3. Enfrentarse al **tráfico** es como enfrentarse a cualquier otra cosa, se debe tener una buena actitud y una sonrisa grande en el rostro.
4. El análisis de textos puede ayudarnos a encontrar aspectos interesantes que no encontraríamos de otra forma.

Referencias


- https://twitter.com/amilcarmontejo?lang=en
- https://github.com/trinker/textclean
- https://www.mjdenny.com/Text_Processing_In_R.html
- https://acadgild.com/blog/text-mining-using-r