

## LIMPIEZA Y VALIDACIÓN DE LOS DATOS

1. Descripción del dataset ¿Por qué es importante y qué pregunta/problema pretende responder?

### A) ¿Qué pregunta/problema pretende responder?

**¿Qué valen las mujeres?**<sup>1</sup> En los últimos quince años, más de 900 hombres han asesinado en España a sus parejas o ex parejas. No estamos hablando de un problema residual, ni puntual.

Si nos fijamos en las estadísticas oficiales, cuya fuente es el Consejo General del Poder Judicial, y ofrecidas en el Portal Estadístico de la Delegación del Gobierno para la Violencia de Género (fecha de referencia: 30/09/2017), observamos:<sup>2</sup>

| Año      | Núm. de denuncias por violencia de género | Año      | Núm. de órdenes de protección | Año      | Número de víctimas mortales |
|----------|---|----------|-------------------------------|----------|-----------------------------|
| Año 2009 | 135.539                                   | Año 2009 | 41.081                        | Año 2003 | 71                          |
| Año 2010 | 134.105                                   | Año 2010 | 37.908                        | Año 2004 | 72                          |
| Año 2011 | 134.002                                   | Año 2011 | 35.813                        | Año 2005 | 57                          |
| Año 2012 | 128.477                                   | Año 2012 | 34.537                        | Año 2006 | 69                          |
| Año 2013 | 124.893                                   | Año 2013 | 32.831                        | Año 2007 | 71                          |
| Año 2014 | 126.742                                   | Año 2014 | 33.167                        | Año 2008 | 76                          |
| Año 2015 | 129.193                                   | Año 2015 | 36.292                        | Año 2009 | 56                          |
| Año 2016 | 143.535                                   | Año 2016 | 37.958                        | Año 2010 | 73                          |
| Año 2017 | 125.769                                   | Año 2017 | 29.455                        | Año 2011 | 62                          |
|          |   |          |                               | Año 2012 | 52                          |
|          |   |          |                               | Año 2013 | 54                          |
|          |   |          |                               | Año 2014 | 55                          |
|          |   |          |                               | Año 2015 | 60                          |
|          |   |          |                               | Año 2016 | 44                          |
|          |   |          |                               | Año 2017 | 48                          |

Cuando se trata de abordar un problema tan grave y tan serio como los asesinatos machistas que se suceden año tras año, mes tras mes, semana tras semana, cabría preguntarse, ¿Qué valor le damos como sociedad a la vida de una mujer?

Salvo honrosas excepciones, los hombres siguen considerando el combate por la igualdad y contra la violencia de género como algo que atañe únicamente a las mujeres; no obstante, se trata de una lucha en la que nos tenemos que implicar todos para que tenga un éxito asegurado.

El engranaje de la violencia machista es muy complejo y se compone de múltiples piezas. Su versatilidad es casi infinita porque el machismo es capaz de mutar y adaptarse al ritmo de los tiempos. Pero el daño que produce es siempre el mismo y tiene siempre como objetivo el sometimiento de la mitad de la sociedad, que son las mujeres.

<sup>1</sup> ¿Qué valen las mujeres? Lidia Guinart Moreno. Periodista y escritora.  
<http://www.tribunafeminista.org/2018/01/que-valen-las-mujeres/>

<sup>2</sup> Portal Estadístico de la Delegación del Gobierno para la Violencia de Género.  
<http://estadisticasviolenciagenero.msssi.gob.es/>

Es necesario recorrer el camino hacia la plena igualdad, hacia la igualdad real entre hombres y mujeres, un camino que pase por la educación y que facilite el destierro del machismo de la sociedad.


**Necesitamos conocer**, si existe en la sociedad una actitud proactiva en contra de la violencia machista y a favor del feminismo -principio de igualdad de derechos de la mujer y el hombre y que cuestiona la dominación y violencia de los varones sobre las mujeres-, que pueda cambiar el estatus quo de las víctimas; o en su caso, si no existe esa actitud o incluso si ésta puede ser negativa.

El feminismo es un fenómeno omnipresente en todo el mundo, y quizás donde más se debata, sea en Twitter y otras redes sociales. En la actualidad la influencia de las redes sociales es constante en la población de todo el mundo y quizás el hashtag #Feminismo haya creado tendencia social al ser capaz de expandir mensajes de toda índole.

Es por ello que, a través del hashtag **#Feminismo**, que es una forma de relacionar directamente los mensajes de los usuarios en Twitter, se podrán observar las opiniones y sentimientos que este tema provoca en la sociedad y podremos responder, tras analizar datos de twitter, qué y cómo se opina del feminismo.

## **B) Descripción del dataset**

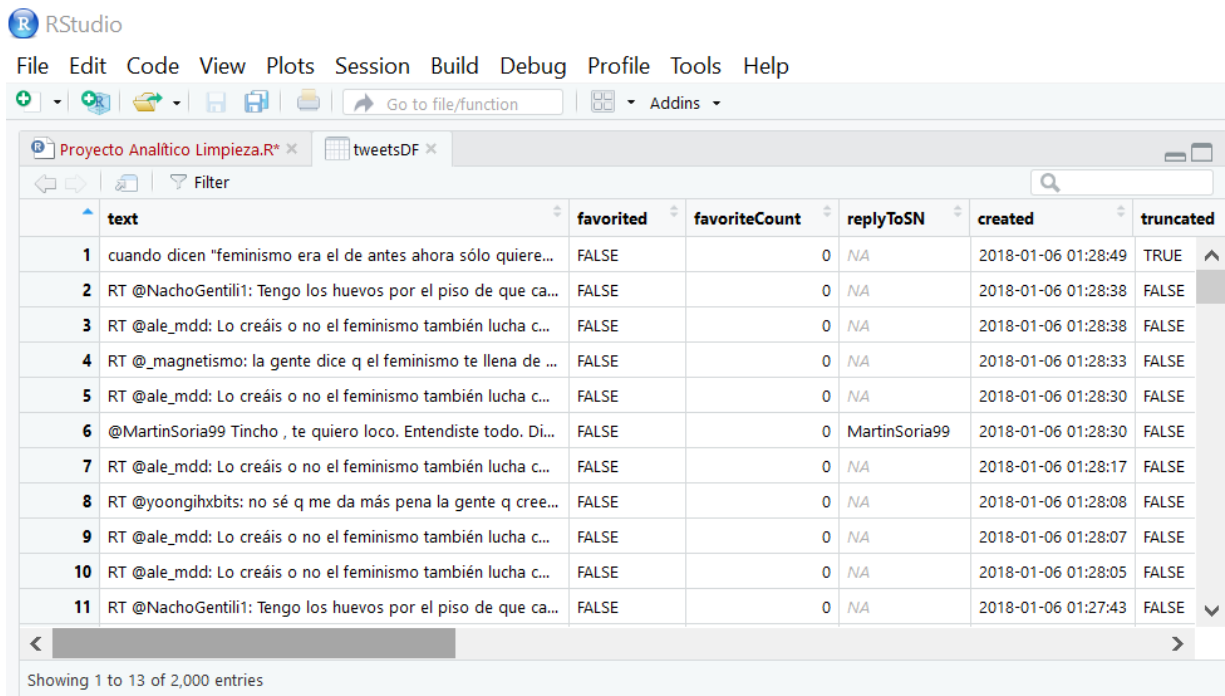
Las operaciones a realizar a través de **RStudio**, son las siguientes:

1. Acceder a la API pública de Twitter. 
2. Descargar el conjunto de datos en un dataset.
3. Limpieza de los datos y crear el dataset para el análisis.
4. Análisis de los datos:
  - 4.1. Análisis de Sentimientos.
  - 4.2. Análisis estadísticos de los resultados
5. Representación de los resultados:
  - 5.1. Representación de las valoraciones por sectores
  - 5.2. Gráfico dinámico: 'highchart'
  - 5.3. Otros gráficos: histograma y boxplot.
6. Modelización y Predicción.
  - 6.1 Machine Learning: 'RTextTools'
  - 6.2 Recta de regresión y pruebas de normalidad

El lenguaje de programación R permite a los *'data scientists'* gestionar grandes volúmenes de datos estadísticos, pero necesita en ocasiones instalar y cargar determinadas librerías -paquetes adicionales que le dan una capacidad de gestión de datos-, por lo que en este caso será necesario cargar: `library(twitteR)`, `library(plyr)`, `library(stringr)`, `library(ggplot2)`, `library(highcharter)`, `library(RTextTools)`, `library(e1071)`, `library(SparseM)` y `library(normtest)`.

Para acceder a la API pública de Twitter, será necesario previamente gestionar los permisos correspondientes en su aplicación y obtener los parámetros de acceso: `consumer_key`, `consumer_secret`, `access_token` y `access_secret`.

Para obtener unos datos suficientemente representativos, se ha optado por descargar 2000 tweets, relacionados con el hashtag #Feminismo, con 16 variables por tweet. Posteriormente, los datos serán incorporados a un *data frame* y exportados a un archivo con formato csv, con el nombre: "*1-tweetsDF-ORIGINAL.csv*".



|    | text   | favorited | favoriteCount | replyToSN     | created             | truncated |
|----|--|-----------|---------------|---------------|---------------------|-----------|
| 1  | cuando dicen "feminismo era el de antes ahora sólo quiere...   | FALSE     | 0             | NA            | 2018-01-06 01:28:49 | TRUE      |
| 2  | RT @NachoGentili1: Tengo los huevos por el piso de que ca...   | FALSE     | 0             | NA            | 2018-01-06 01:28:38 | FALSE     |
| 3  | RT @ale_mdd: Lo creáis o no el feminismo también lucha c...    | FALSE     | 0             | NA            | 2018-01-06 01:28:38 | FALSE     |
| 4  | RT @magnetismo: la gente dice q el feminismo te llena de ...   | FALSE     | 0             | NA            | 2018-01-06 01:28:33 | FALSE     |
| 5  | RT @ale_mdd: Lo creáis o no el feminismo también lucha c...    | FALSE     | 0             | NA            | 2018-01-06 01:28:30 | FALSE     |
| 6  | @MartinSoria99 Tincho , te quiero loco. Entendiste todo. Di... | FALSE     | 0             | MartinSoria99 | 2018-01-06 01:28:30 | FALSE     |
| 7  | RT @ale_mdd: Lo creáis o no el feminismo también lucha c...    | FALSE     | 0             | NA            | 2018-01-06 01:28:17 | FALSE     |
| 8  | RT @yoongihxbits: no sé q me da más pena la gente q cree...    | FALSE     | 0             | NA            | 2018-01-06 01:28:08 | FALSE     |
| 9  | RT @ale_mdd: Lo creáis o no el feminismo también lucha c...    | FALSE     | 0             | NA            | 2018-01-06 01:28:07 | FALSE     |
| 10 | RT @ale_mdd: Lo creáis o no el feminismo también lucha c...    | FALSE     | 0             | NA            | 2018-01-06 01:28:05 | FALSE     |
| 11 | RT @NachoGentili1: Tengo los huevos por el piso de que ca...   | FALSE     | 0             | NA            | 2018-01-06 01:27:43 | FALSE     |

Muestra de 1 a 13 tweets, de los 2000 que contiene el archivo *1-tweetsDF-ORIGINAL.csv*

## 2. Limpieza de los datos.

### 2.1. Selección de los datos de interés a analizar. ¿Cuáles son los campos más relevantes para responder al problema?

Es a través del hashtag #Feminismo, como los usuarios en Twitter relacionan directamente sus mensajes con cadena de caracteres formada por la palabra feminismo precedida por una almohadilla.

Dado que se necesita conocer, si en Twitter existe una actitud proactiva o no a favor del feminismo y la tendencia social que se expande tras el hashtag #Feminismo, es en el campo "text", donde se podrán observar las opiniones y sentimientos que este tema provoca en la sociedad. Si los datos de interés a analizar están en el campo "text" de cada tweet, se pueden descartar el resto de variables del tweet al no ser de interés para esta práctica.

Posteriormente con los datos del campo "text", realizaremos las siguientes operaciones:

- I. Limpieza del texto, mediante la aplicación de una función diseñada *ad hoc*.
- II. Realizar un análisis de sentimientos, mediante la aplicación de la correspondiente función.

## 2.2. ¿Los datos contienen ceros o elementos vacíos? ¿Y valores extremos? ¿Cómo gestionarlos cada uno de estos casos?

Un **Tweet** es un mensaje que puede alcanzar actualmente hasta 280 caracteres: letras, números, signos y enlaces. Si en un tweet existen diferentes caracteres que no tienen significación sentimental alguna, en 2000 tweets la cifra es muy significativa, por ello, eliminarlos supondría una laboriosa transformación sino se utilizase un procesado de texto. En R existen varias funciones, si bien en este caso se ha utilizado *gsub ()*, la cual reemplaza la aparición de una subcadena con otra subcadena dentro de un vector.

La función de limpieza del texto diseñada, que se puede consultar en el archivo del código generado en R, permitió eliminar en cada uno de los tweets: números, RT, espacios vacíos, caracteres propios de Twitter sin significación de sentimientos y los links. Como resultado de la limpieza de datos, el campo "text" estará en condiciones de serle aplicado el análisis de sentimientos en cada uno de los 2000 tweets.

## 3. Análisis de los datos.

### 3.1. Selección de los grupos de datos que se quieren analizar/comparar.

Puesto que el objetivo es analizar las palabras contenidas en el texto limpio de los tweets, será necesario disponer de un **diccionario de palabras ponderadas según sentimientos** relacionados con la palabra "Feminismo", para comparar cada palabra de los tweets con las del diccionario y si coinciden asignarle el valor correspondiente.

Señalar tres cuestiones: la primera, que el diccionario de sentimientos en origen corresponde a valoración de películas descargado de internet con muchos errores y palabras repetidas, por lo que se tuvo que limpiar previamente a su uso; la segunda, que las palabras incluidas en los tweets relacionados con el hashtag #Feminismo no resultaban significativas en los primeros análisis de sentimientos, ofreciendo casi exclusivamente el resultado final de "neutro" a pesar de observar que no se correspondía con las opiniones de algunos tweets; y tercero, una vez analizadas las palabras de los tweets, fueron incorporadas al diccionario 200 de ellas cargadas de significación con el hashtag #Feminismo y otras 200 palabras incluidas en el diccionario fueron cambiadas del masculino al femenino (ejemplo: torturada por torturado, agredida por agredido, humillada por humillado, etc.), e incluso en determinados casos se han mantenido las dos formas. El nombre del archivo del diccionario de sentimientos es: "*LISTA-Palabras-Sentimientos.txt*"

Una vez que se tiene el texto de los tweets limpio y el diccionario de sentimientos confeccionado, se crea una FUNCIÓN para clasificar palabras del texto de los tweet, en 4 categorías: muy Negativo, negativo, positivo y muy Positivo.

Seguidamente se creará una matriz "*scores\_final*", la cual incluirá el campo "text" y los cuatro campos correspondientes a la clasificación de las palabras. Se calcula la puntuación de cada tweet y se añaden dos nuevas columnas, la del valor alcanzado por cada tweet y el del sentimiento asignado.

El *data frame* obtenido se exporta a un archivo csv: "**2-tweet.Resultado.csv**".

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Proyecto Analítico Limpieza.R\* tweetResultado

Filter

|    | texto   | muyNeg | Neg | Pos | muyPos | Valor      | Sentimiento |
|----|---|--------|-----|-----|--------|------------|-------------|
| 1  | cuando dicen feminismo era el de antes ahora sólo quieren...      | 0      | 0   | 0   | 0      | 0.000000   | Neutro      |
| 2  | Tengo los huevos por el piso de que cada tweets uno sea d...      | 0      | 0   | 0   | 0      | 0.000000   | Neutro      |
| 3  | Lo creáis o no el feminismo también lucha contra esto Si lo...    | 0      | 1   | 0   | 0      | -0.500000  | Negativo    |
| 4  | la gente dice q el feminismo te llena de odio y re q te llena ... | 0      | 1   | 2   | 0      | 0.3571429  | Positivo    |
| 5  | Lo creáis o no el feminismo también lucha contra esto Si lo...    | 0      | 1   | 0   | 0      | -0.500000  | Negativo    |
| 6  | Tincho te quiero loco Entendiste todo Dio un discurso sobr...     | 0      | 2   | 1   | 0      | -0.3571429 | Negativo    |
| 7  | Lo creáis o no el feminismo también lucha contra esto Si lo...    | 0      | 1   | 0   | 0      | -0.500000  | Negativo    |
| 8  | no sé q me da más pena la gente q cree q el feto siente a l...    | 0      | 1   | 0   | 0      | -0.500000  | Negativo    |
| 9  | Lo creáis o no el feminismo también lucha contra esto Si lo...    | 0      | 1   | 0   | 0      | -0.500000  | Negativo    |
| 10 | Lo creáis o no el feminismo también lucha contra esto Si lo...    | 0      | 1   | 0   | 0      | -0.500000  | Negativo    |
| 11 | Tengo los huevos por el piso de que cada tweets uno sea d...      | 0      | 0   | 0   | 0      | 0.000000   | Neutro      |
| 12 | Lo creáis o no el feminismo también lucha contra esto Si lo...    | 0      | 1   | 0   | 0      | -0.500000  | Negativo    |

Showing 1 to 13 of 2,000 entries

Muestra de 1 a 13 tweets, de los 2000 que contiene el archivo **2-tweet.Resultado.csv**

Finalmente, se crea una tabla de sentimientos 5x2, correspondiendo a la primera columna sentimientos (cuatro categorías de sentimientos más el neutro -ausencia de sentimientos-) y a la segunda columnas la frecuencia observada para cada sentimiento. La tabla de sentimientos es exportada a un archivo csv: "**3-Tabla-Conteo-Sentimientos.csv**".

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Proyecto Analítico Limpieza.R\* conteo

Filter

|   | Var1         | Freq |
|---|--------------|------|
| 1 | Muy Negativo | 84   |
| 2 | Muy Positivo | 17   |
| 3 | Negativo     | 944  |
| 4 | Neutro       | 588  |
| 5 | Positivo     | 367  |

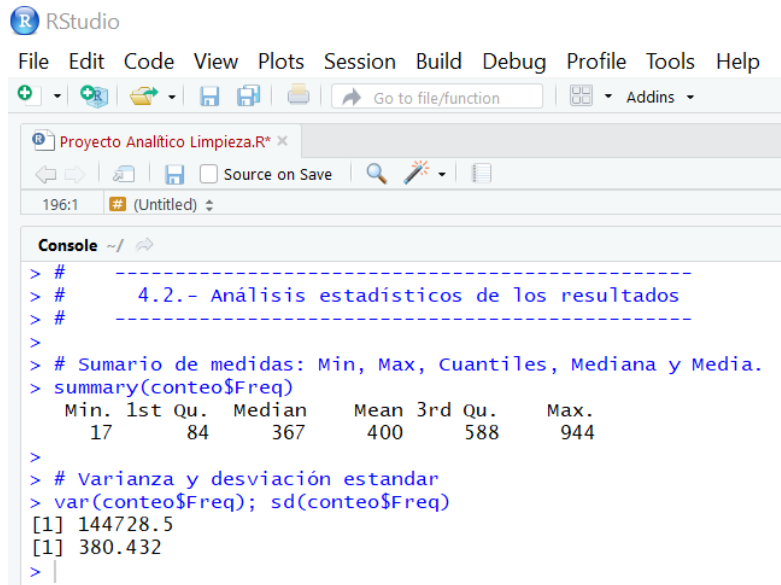
Showing 1 to 5 of 5 entries

**3-Tabla-Conteo-Sentimientos.csv**".

### 3.2. Análisis estadísticos de los resultados.

#### A) 3-Tabla-Conteo-Sentimientos

Dado que inicialmente, lo que tenemos es una muestra de 2000 observaciones, de la cual hemos obtenido una tabla de sentimientos desde muy negativo a muy positivo, a través de la función *summary* comprobaremos el sumario de medidas que figuran en el gráfico que figura más abajo. También calcularemos su varianza y desviación estándar.



```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Proyecto Analítico Limpieza.R*
Source on Save
196:1 (Untitled)
Console
> # -----
> # 4.2.- Análisis estadísticos de los resultados
> # -----
> # Sumario de medidas: Min, Max, Cuantiles, Mediana y Media.
> summary(conteo$Freq)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    17     84     367    400    588    944
>
> # Varianza y desviación estandar
> var(conteo$Freq); sd(conteo$Freq)
[1] 144728.5
[1] 380.432
>

```

Seguidamente se realizarán dos pruebas, una de *machine learning* utilizando la librería *RtexTools*, para crear una tabla de predicciones y calcular la precisión de la recuperación de los datos y otra segunda prueba, de comparación de dos muestras (antes-después), donde se estudiará el ajuste de datos reales entre ambas.

#### B) Predicción de los resultados - Machine Learning: 'RtexTools'

En esta prueba de clasificación automática de texto mediante aprendizaje supervisado, se realizarán las siguientes operaciones:

- 1.- Del *tweetResultado*, con 2000 observaciones y 7 variables, se obtendrá una tabla del conjunto total de filas pero con solo dos campos, el de texto y el de sentimientos.
- 2.- De la tabla se obtendrá una matriz, las filas serán los "Docs" y las columnas los "Terms". El tamaño de los terms se ha configurado entre 4 y 15 letras, al entender que son los terms de mayor significado sentimental, además reduce el tamaño de la matriz significativamente.
- 3.- Asignamos al primer argumento el 80% de los resultados (1600 tweets) y al segundo el 20% restante.
- 4.- Una vez creada la tabla de predicciones, calculamos la precisión de recuperación de los datos.

Como se puede observar en la tabla que figura más abajo, los resultados muestran que la totalidad de las predicciones son negativas, calculando una precisión de la recuperación de los datos del 46,63%.

```

Proyecto Analítico Limpieza.R* x
# (Clasificación automática de texto mediante aprendizaje supervisado)
library(RTextTools)
library(e1071)
library(SparseM)

# Modelo de Predicciones
tweets_ponderados <- tweetResultado[,c(1,7)]

matrix <- create_matrix(tweets_ponderados[,1], minWordLength=4, language="spanish",
                        maxWordLength=15, removeStopwords = F, removeNumbers = T,
                        removePunctuation=F, toLower=T, stemWords = F)

mat <- as.matrix(matrix)
classifier <- naiveBayes(mat[1:(num.tweets*0.8)], as.factor(tweets_ponderados[1:(num.tweets*0.8),2]))

# Función genérica para predicciones a partir de los resultados de ajuste del primer argumento.
predicted <- predict(classifier, mat[(num.tweets*0.8):num.tweets]); predicted

# Crea la tabla de predicción
table(tweets_ponderados[(num.tweets*0.8):num.tweets,2], predicted)

# Calcula la precisión de recuperación de los datos clasificados
recall_accuracy(tweets_ponderados[(num.tweets*0.8):num.tweets,2], predicted)

```

```

Console ~/
[386] Negativo Negativo Negativo Negativo Negativo Negativo Negativo Negativo Negativo Negativo Negativo
[397] Negativo Negativo Negativo Negativo Negativo
Levels: Muy Negativo Muy Positivo Negativo Neutro Positivo
>
> # Crea la tabla de predicción
> table(tweets_ponderados[(num.tweets*0.8):num.tweets,2], predicted)
predicted
Muy Negativo Muy Positivo Negativo Neutro Positivo
Muy Negativo      0          0      13      0      0
Muy Positivo      0          0       7      0      0
Negativo          0          0    187      0      0
Neutro            0          0    130      0      0
Positivo          0          0     64      0      0
>
> # Calcula la precisión de recuperación de los datos clasificados
> recall_accuracy(tweets_ponderados[(num.tweets*0.8):num.tweets,2], predicted)
[1] 0.4663342

```

### 3.3. Aplicación de pruebas estadísticas (tantas como sea posible) para comparar los grupos de datos.

Como se indicaba anteriormente la segunda prueba, será de comparación de dos muestras (antes-después), donde se estudiará el ajuste de datos reales entre ambas.

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Proyecto Analítico Limpieza.R\* x conteo x código 2.R x conteo\_2 x conteo

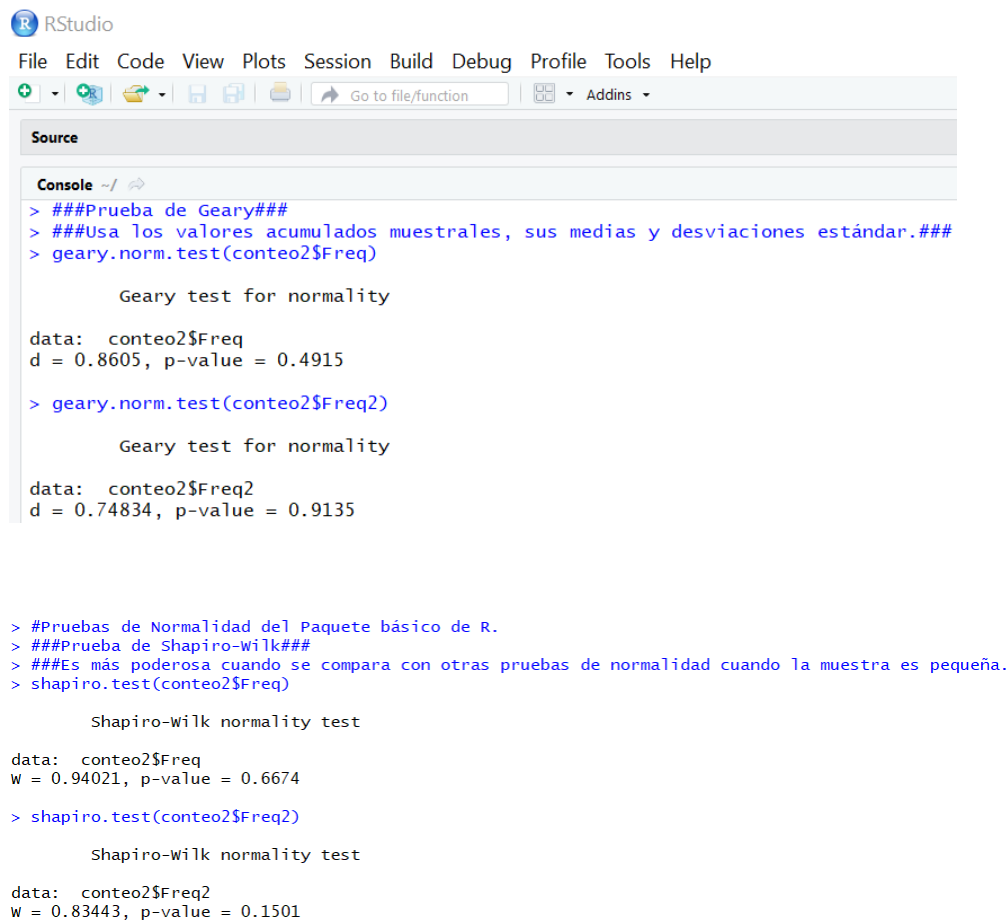
| Var1           | Freq | Freq2 |
|----------------|------|-------|
| 1 Muy Negativo | 84   | 108   |
| 2 Muy Positivo | 17   | 40    |
| 3 Negativo     | 944  | 366   |
| 4 Neutro       | 588  | 1111  |
| 5 Positivo     | 367  | 375   |

Previamente se han realizado Pruebas de Normalidad:

Hipótesis:                      H0: La muestra proviene de una distribución normal.  
                                      H1: La muestra no proviene de una distribución normal.

Nivel de Significancia: Alfa=0.05

Criterio de Decisión:    Si  $p < \text{Alfa}$  Se rechaza  $H_0$   
                                     Si  $p \geq \text{Alfa}$  No se rechaza  $H_0$



```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
+ [Icons] Go to file/function [Icons] Addins

Source

Console ~/
> ###Prueba de Geary###
> ###Usa los valores acumulados muestrales, sus medias y desviaciones estándar.###
> geary.norm.test(conteo2$Freq)

      Geary test for normality

data:  conteo2$Freq
d = 0.8605, p-value = 0.4915

> geary.norm.test(conteo2$Freq2)

      Geary test for normality

data:  conteo2$Freq2
d = 0.74834, p-value = 0.9135

> #Pruebas de Normalidad del Paquete básico de R.
> ###Prueba de Shapiro-Wilk###
> ###Es más poderosa cuando se compara con otras pruebas de normalidad cuando la muestra es pequeña.
> shapiro.test(conteo2$Freq)

      Shapiro-Wilk normality test

data:  conteo2$Freq
W = 0.94021, p-value = 0.6674

> shapiro.test(conteo2$Freq2)

      Shapiro-Wilk normality test

data:  conteo2$Freq2
W = 0.83443, p-value = 0.1501

```

Conclusión: No existe evidencia estadística para rechazar  $H_0$ . Es decir, podemos afirmar que **las variables tienen una distribución normal**. Esto se cumple para todas las pruebas.

Reseñar, que a la vista de los resultados de las pruebas realizadas, se utilizarán pruebas paramétricas al asumirse que las distribuciones estadísticas subyacentes a los datos cumplen condiciones de validez y fiabilidad; además, consideradas las dos muestras independientes, éstas se ajuntan a una distribución normal y las varianzas son homogéneas.

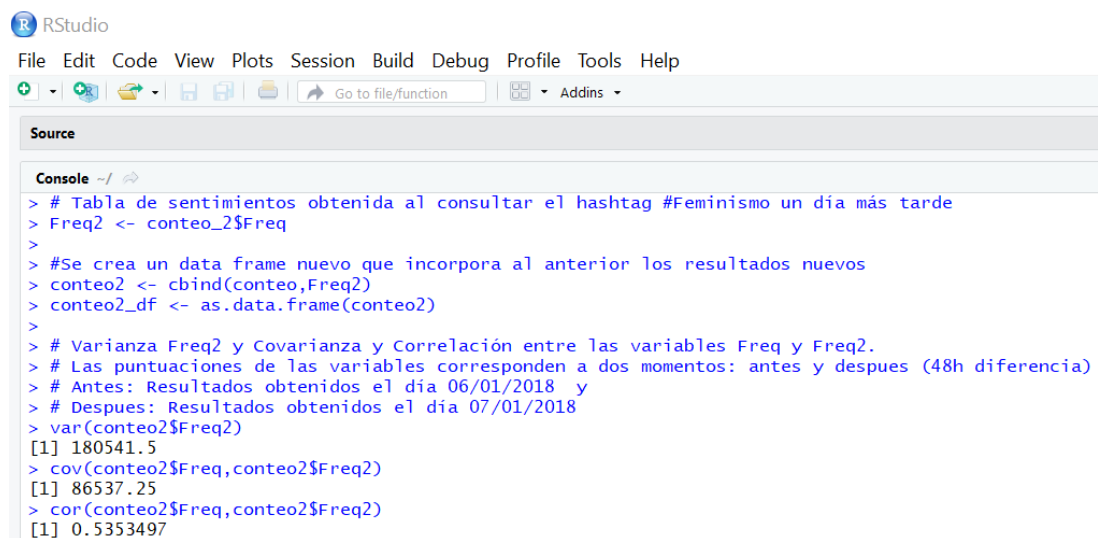
A diferencia de la prueba de *machine learning*, ahora los datos son muestras reales y por lo tanto, nos indicarán con mayor fiabilidad, no solo como funciona el ajuste del modelo establecido para el análisis de sentimientos sino también la variabilidad de los sentimientos recogidos en los tweets en tan solo 48 horas de diferencia. Para hacer un análisis de series temporales necesitaríamos una continuidad de datos a lo largo de un tiempo amplio y constante.



El proceso de descarga de los datos y limpieza es el mismo, lo que cambiarán serán los contenidos de los mensajes de texto de los tweets y por lo tanto también el resultado del análisis de sentimientos.

Ahora con dos muestras, y por lo tanto con dos columnas de resultados de frecuencia para la valoración de sentimientos, se podrán realizar análisis de covarianza para analizar la variación común a dos variables y, por lo tanto, una medida del grado y tipo de su relación.

Lo que resulta de gran interés es conocer el valor de la correlación, es decir la fuerza y la relación lineal entre dos variables, que en este caso es de 0,535, lo que nos indica que carece de significación suficiente aunque la relación sea directa.



```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Source
Console ~/
> # Tabla de sentimientos obtenida al consultar el hashtag #Feminismo un día más tarde
> Freq2 <- conteo_2$Freq
>
> #Se crea un data frame nuevo que incorpora al anterior los resultados nuevos
> conteo2 <- cbind(conteo,Freq2)
> conteo2_df <- as.data.frame(conteo2)
>
> # Varianza Freq2 y Covarianza y Correlación entre las variables Freq y Freq2.
> # Las puntuaciones de las variables corresponden a dos momentos: antes y despues (48h diferencia)
> # Antes: Resultados obtenidos el día 06/01/2018 y
> # Despues: Resultados obtenidos el día 07/01/2018
> var(conteo2$Freq2)
[1] 180541.5
> cov(conteo2$Freq,conteo2$Freq2)
[1] 86537.25
> cor(conteo2$Freq,conteo2$Freq2)
[1] 0.5353497

```

En relación a la recta de regresión entre los resultados obtenidos antes y después, señalar que nos ha dado una pendiente  $b=0,4793$  y una ordenada en el origen  $a=208,271$ .

Un valor muy importante es el **coeficiente de determinación  $R^2=0,2826$** , que es un valor bajo y por lo tanto nos está indicando que no hay una relación significativa entre los resultados obtenidos entre las dos muestras.

Los resultados obtenidos pueden ser consecuencia del ajuste del método de análisis empleado que debería afinarse más, tanto el diccionario semántico que ya desde un principio se ha revelado como muy determinante sobre los resultados como las ponderaciones asignadas a las mediciones.

Tampoco podríamos descartar que eventos puntuales hayan pesado significativamente sobre tema de opinión (ejemplo: el hallazgo de Diana Quer y el encarcelamiento de su asesino).

También cabe indicar, que las descargas de tweets, tanto antes como después, han coincidido en días festivos dentro del periodo de Navidad.

Seguidamente, calculamos la recta de regresión entre los resultados obtenidos antes y después, así como, la línea de regresión:

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
+ - [R] [Save] [Run] [Go to file/function] [Addins]

Source

Console ~/
> # Hallamos la recta de regresión entre ambos resultados.
> # Obtenemos: pendiente de la recta, ordenada en el origen y coeficiente de determinación R2
> RegModel.1 <- lm(formula=Freq~Freq2, data=conteo2)
> summary(RegModel.1)

Call:
lm(formula = Freq ~ Freq2, data = conteo2)

Residuals:
    1      2      3      4      5 
-176.04 -210.44  560.30 -152.80  -21.02 

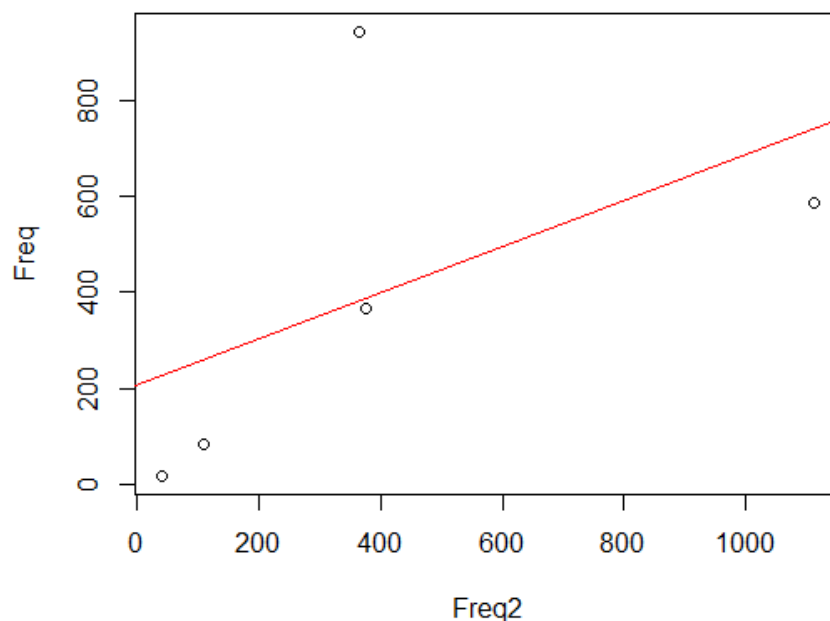
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  208.2718   240.9021    0.865   0.451
Freq2         0.4793     0.4366    1.098   0.352

Residual standard error: 371 on 3 degrees of freedom
Multiple R-squared:  0.2866,    Adjusted R-squared:  0.0488 
F-statistic: 1.205 on 1 and 3 DF,  p-value: 0.3525

> 
> # Realizamos el gráfico de la nube de puntos de las dos variables con la recta de regresión
> plot(Freq~Freq2, data=conteo2)
> abline(lm(formula=Freq~Freq2, data=conteo2), col="red")

```

Como puede observarse en la línea de regresión, existen puntuaciones que quedan muy alejadas de la línea. Para estos puntos el error o residuo es considerable, por lo tanto, el ajuste de las puntuaciones a la recta no es bueno.



#### 4. Representación de los resultados a partir de tablas y gráficas.

##### A) PRIMER ANÁLISIS DE SENTIMIENTOS

Tras el hashtag #Feminismo, los usuarios en Twitter en sus mensajes manifiestan opiniones y sentimientos mayoritariamente negativos o neutros.

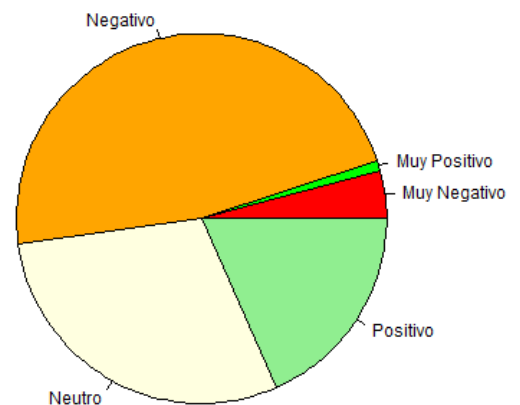
##### 3.2-Rplot.Sentimientos.Feminismo.png

Gráfica de sectores obtenida con **Rstudio** y corresponde al **primer análisis de sentimientos**.

Como puede verse en la tabla, los sentimientos negativos superan al resto de valores.

| Var1         | Freq |
|--------------|------|
| Muy Negativo | 84   |
| Muy Positivo | 17   |
| Negativo     | 944  |
| Neutro       | 588  |
| Positivo     | 367  |

Distribución de Sentimientos - Feminismo (tweets)



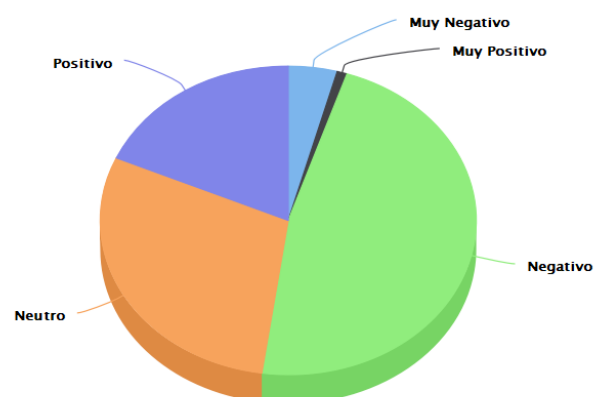
##### 3-Rplot.Sentimientos.Feminismo

Gráfica de sectores obtenida con **highcharter** y corresponde al **primer análisis de sentimientos**.

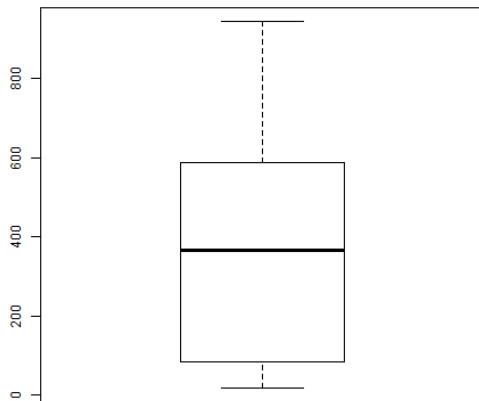
Como puede verse en la tabla, los sentimientos negativos superan al resto de valores.

| Var1         | Freq |
|--------------|------|
| Muy Negativo | 84   |
| Muy Positivo | 17   |
| Negativo     | 944  |
| Neutro       | 588  |
| Positivo     | 367  |

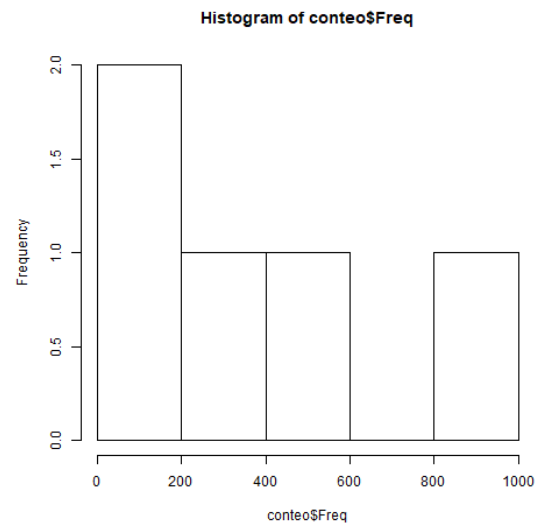
Distribución de Sentimientos – Feminismo (tweets)



5.3.2-Boxplot.Feminismo.png



5.3.1-Hist.Feminismo.png



En el boxplot, podemos observar que las frecuencias están acumuladas mayoritariamente en las tres primeras zonas; así como, que la mediana se encuentra en la segunda zona y que el máximo correspondiente al valor muy positivo con una puntuación de 17 se encuentra alejado. No existe simetría respecto a la mediana, ya que hay una cola hacia la derecha, estando los datos más apretados en el intervalo de Q1 a Q3 y más dispersos del intervalo Q3 al máximo.

## B) COMPARATIVA DE ANÁLISIS DE SENTIMIENTOS ANTES-DESPUÉS

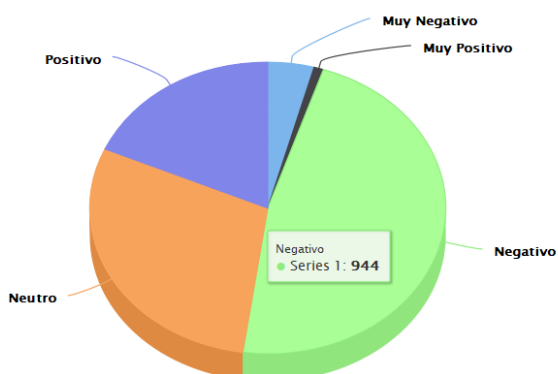
RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

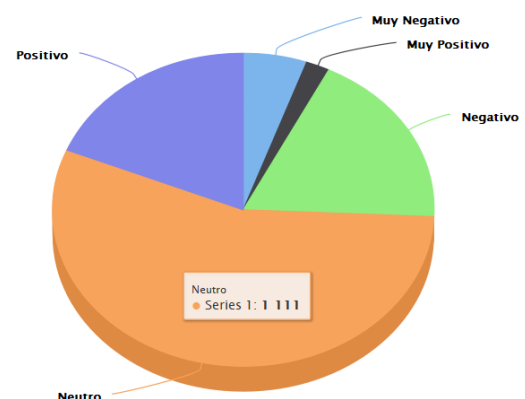
Proyecto Analítico Limpieza.R\* x conteo x codigo 2.R x conteo\_2 x conteo\_3 x

|   | Var1         | Freq | Freq2 |
|---|--------------|------|-------|
| 1 | Muy Negativo | 84   | 108   |
| 2 | Muy Positivo | 17   | 40    |
| 3 | Negativo     | 944  | 366   |
| 4 | Neutro       | 588  | 1111  |
| 5 | Positivo     | 367  | 375   |

Distribución de Sentimientos – Feminismo (tweets)



Distribución de Sentimientos – Feminismo (tweets)



5. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Tras el hashtag #Feminismo, los usuarios en Twitter en sus mensajes manifiestan opiniones y sentimientos mayoritariamente negativos o neutros, no siendo despreciable tampoco la cantidad de mensajes muy negativos.

Si bien el diccionario de sentimientos es muy sensible, éste mejoraría incorporando palabras más representativas del feminismo. También son sensibles las ponderaciones, las cuales deberán ser ajustadas en sus precisiones al variarse el diccionario de sentimientos.

Para ajustar el proceso de análisis de sentimientos es necesario estudiar el fenómeno con más muestras a lo largo de diferentes días de la semana y controlando los efectos que los eventos negativos puedan tener como impacto sobre las emociones que se escriban en los tweets.

Se ha podido comprobar que las opiniones que se vierten contra el feminismo son negativas o muy negativas en un número considerable.

El feminismo es un fenómeno omnipresente en Twitter por en gran número de tweets que genera cada día y las opiniones de muchos usuarios son negativas al mismo y no dudan en manifestarlo.

El camino a recorrer hacia la plena igualdad entre hombres y mujeres será muy largo y pasa por educar en principios de igualdad de derechos de la mujer y el hombre y cuestionar la dominación y violencia de los varones sobre las mujeres.

Los resultados si permiten responder al problema, resaltando que existe en la sociedad una actitud en contra del feminismo, por lo que no será fácil el destierro del machismo de la sociedad.

6. Código: Hay que adjuntar el código en R, con el que se ha realizado la limpieza, análisis y representación de los datos.

Ha sido subido a GitHub y se encuentra en el siguiente enlace:

[https://github.com/salgadogb/Tipologia\\_Practica\\_2](https://github.com/salgadogb/Tipologia_Practica_2)

El código está completamente comentado.

## LIMPIEZA Y VALIDACIÓN DE LOS DATOS

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico, donde se aplicarán las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis. Se entregará un solo archivo con el enlace Github (<https://github.com>). Se utilizará la Wiki de Github para describir el proyecto.

El objetivo de esta actividad será el tratamiento de un dataset. Las diferentes tareas a realizar (y **justificar**) son las siguientes:

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?
2. Limpieza de los datos.
  - 2.1. Selección de los datos de interés a analizar. ¿Cuáles son los campos más relevantes para responder al problema?
  - 2.2. ¿Los datos contienen ceros o elementos vacíos? ¿Y valores extremos? ¿Cómo gestionarías cada uno de estos casos?
3. Análisis de los datos.
  - 3.1. Selección de los grupos de datos que se quieren analizar/comparar.
  - 3.2. Comprobación de la normalidad y homogeneidad de la varianza. Si es necesario (y posible), aplicar transformaciones que normalicen los datos.
  - 3.3. Aplicación de pruebas estadísticas (tantas como sea posible) para comparar los grupos de datos.
4. Representación de los resultados a partir de tablas y gráficas.
5. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?
6. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos.

Hay que entregar un único fichero que contenga el enlace Github donde haya:

1. Una Wiki con el nombre del componente y una descripción de los ficheros.
2. Un documento Word, Open Office o PDF con las respuestas a las preguntas y con el nombre del componente.
3. Una carpeta con el código generado para analizar los datos.
4. El fichero CSV con los datos originales.
5. El fichero CSV con los datos finales analizados.

Este documento de entrega final de la Práctica 2 se tiene que entregar en el espacio de Entrega y Registro de AC del aula virtual.