

Universidad de Buenos Aires

Facultad De Ingeniería

Año 2020 - 1er Cuatrimestre

**ORGANIZACIÓN DE DATOS (75.06)**

TRABAJO PRÁCTICO Nº 1

TEMA: Análisis exploratorio de datos

GRUPO: 23 – Barbijo De Uranio

INTEGRANTES:



Repositorio en GitHub: <https://github.com/JulioCastillo1/TP1-Datos>

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

75.06 – Organización de datos Curso 1 – Luis Argerich

# INTRODUCCIÓN

El objetivo de este trabajo práctico es realizar un análisis exploratorio del set de datos de la competencia de Kaggle: https://www.kaggle.com/c/nlp-getting-started​ .

El archivo train.csv cuenta con los siguientes campos:

* Id: identificador unico para cada tweet
* text: el texto del tweet
* location: Ubicación desde donde fue enviado (podría no estar)
* keyword: Un keyword para el tweet (podría faltar)
* target: Indica si se trata de un desastre real (1) o no (0)

Se pretende descubrir elementos, patrones o atributos sobre los datos que puedan resultar interesantes.

El código fue escrito en lenguaje Python 3. Para la manipulación de los datos se hace uso de algunas bibliotecas como:

* Pandas
* Matplotlib
* Numpy
* Seaborn
* Geopy
* Wordcloud
* Nltk
* Collections
* Squarfy

**A saber**: Siempre que se refiera a algo como “On/Off Target” o similar, hacemos referencia a si el tweet habla verdaderamente o no de algún desastre. También cuando nos referimos a la “Veracidad” de un Tweet.

# ANÁLISIS GENERAL DE LOS DATOS

En primer lugar se leen los datos del archivo train.csv, y se pasan a un dataframe. Mediante el método info(), se observa la información referida al mismo.

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 7613 entries, 0 to 7612

Data columns (total 5 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 id 7613 non-null int64

1 keyword 7552 non-null object

2 location 5080 non-null object

3 text 7613 non-null object

4 target 7613 non-null int64

dtypes: int64(2), object(3)

memory usage: 297.5+ KB

Puede observarse que los campos id, text y target están completos, mientras que keyword y location contienen valores nulos.

Damos un vistazo a las frecuencia de **target**, donde el valor 0 se corresponde con falso, y el 1 con verdadero.

0 4342

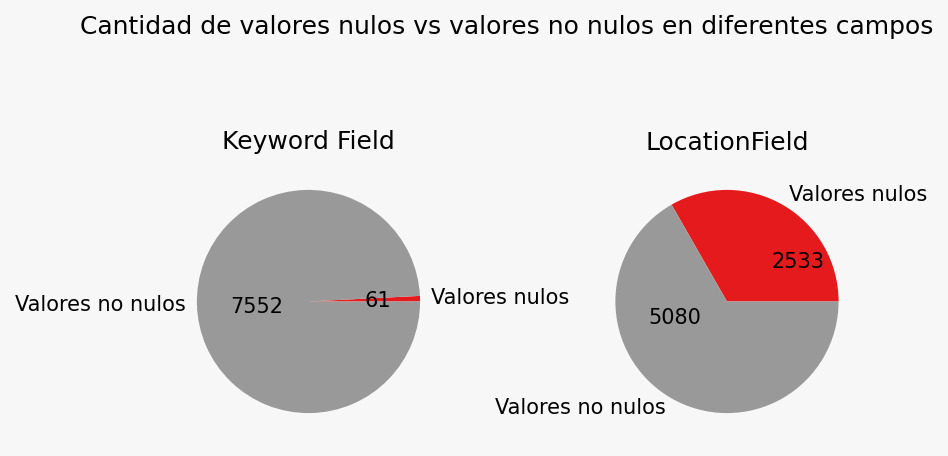
1 3271

Name: target, dtype: int64

## 

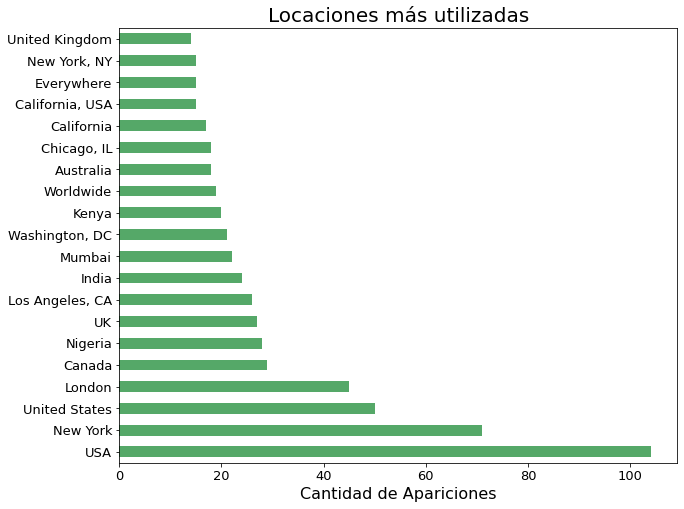
## Valores nulos

Se hecha un vistazo a la cantidad de valores nulos, los cuales se hallan en dos de los campos del dataset, “keyword” y “location”. En lugar de realizar una limpieza de los mismos al inicio, lo realizaremos cada caso de uso, dependiendo del empleo de cada Dataframe o serie durante el desarrollo del trabajo práctico.



## **Locaciones**

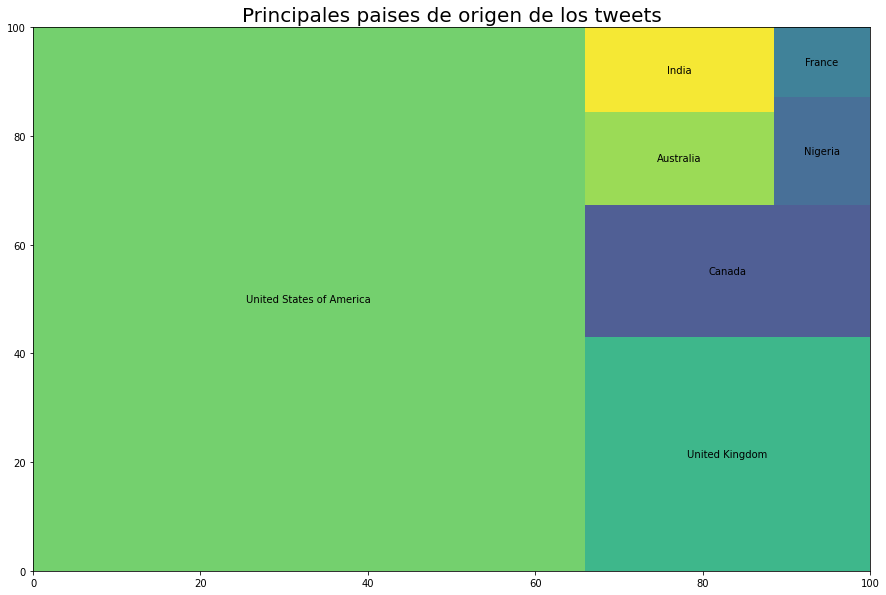
Como primera curiosidad, analizamos qué locaciones (columna ‘location’) son las que aportan más cantidad de tweets.



Y de éste gráfico, podemos ver que se repiten locaciones pero con distintas formas de referirse a ellos. Por ejemplo, ‘USA’ y ‘United States’ hacen referencia a la misma locación, sin embargo, no podemos procesar eso directamente, debemos interpretarlo. También destacamos que, por ejemplo, aparecen locaciones como ‘Los Angeles’, que se encuentra dentro de los Estados Unidos. Por lo que, si queremos contabilizar la cantidad de tweets que tienen origen en el país norteamericano, debemos incluir a estos.

Aquí es donde entra en juego la herramienta ‘Geopy’. Esta nos permitirá interpretar distintas locaciones y de ser posible asignarles propiedades como país, estado, latitud y longitud, entre otras.

A continuación se procedió a confeccionar un treemap. Para ello se realizó una limpieza de locaciones, se agruparon los diferentes tipos de ubicaciones en países, ya que además figuraban localidades, provincias, estados, etc, y se eliminaron repeticiones. Se seleccionaron los que poseían las mayores cantidades de ocurrencias, para que el gráfico no quede saturado.



Como era de esperar, por lejos el país de origen que agrupa con mayor frecuencia las locaciones del campo correspondiente en los tweets es Estados Unidos.

# 

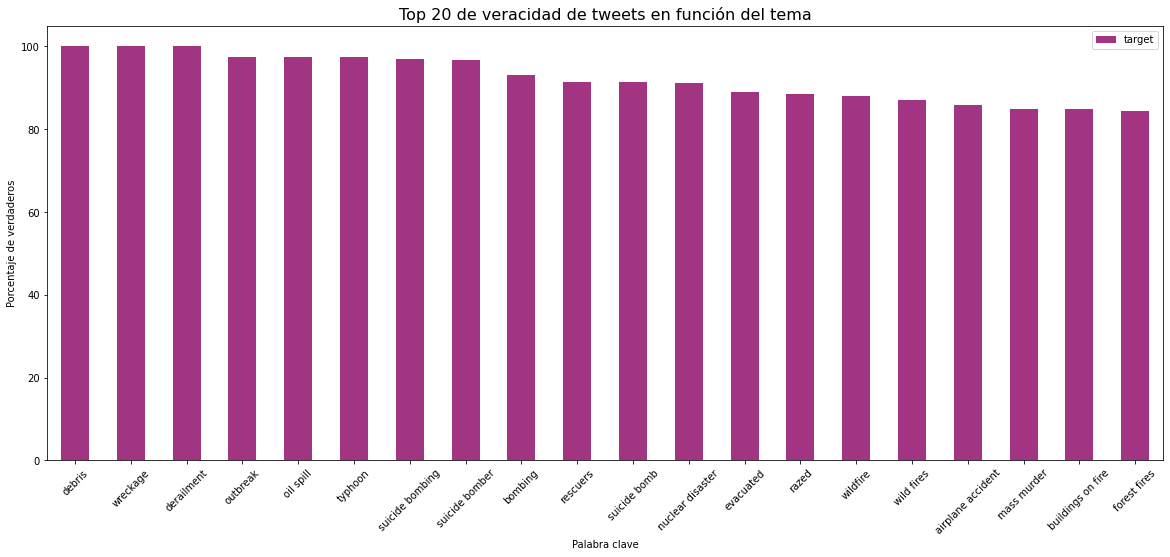
# VERACIDAD DE LOS TWEETS

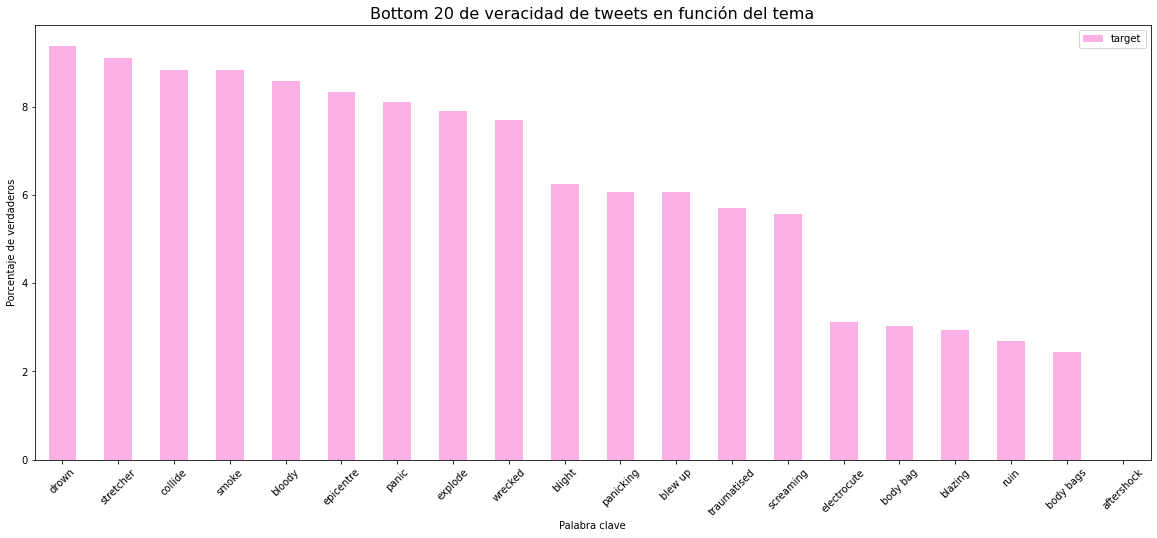
## Porcentaje de veracidad según keyword

En primer lugar se procede a asociar la frecuencia de las keywords con el target.

**¿Qué asuntos están más asociados a desastres reales y cuáles a metafóricos?**

Como las categorías son más de 200, hacemos un gráfico con el top 20 y el bottom 20.





Observando los dos gráficos previos, se puede notar qué tipos de palabras clave suelen ser sobre desastres reales y cuáles se prestan más a la broma, metáfora, chiste u otro hecho artístico.

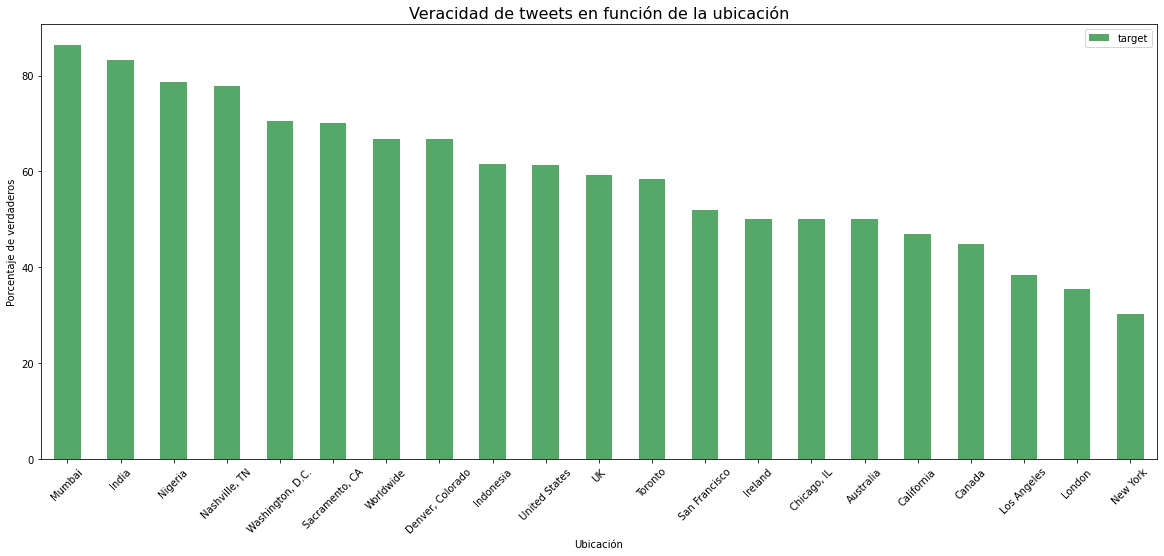
La lista de veracidad la encabezan palabras clave como debris (escombros), wreckage (destrucción), derailment (descarrilamiento), outbreak (brote), oil spill (derrame de petróleo), razed (arrasado), typhoon (tifón), wildfire (incendio forestal), airplane accidents (no hace falta traducción :)) y buildings on fire (incendios en edificios).

La de tweets sobre desastres no reales tiene temas como drown (ahogar o empapar), stretcher (camilla), collide (colisionar), bloody (sangrante), panic (pánico), explode (explotar), wrecked (borracho), screaming (gritando), electrocute (electrocución), body bag (bolsa de cadáver), aftershock (que no es usado en el sentido del terremoto físico sino metafórico), blight (plaga de plantas, literal o metafórica).

Entonces lo que se nota es que cuando las palabras claves están asociadas a desastres en gran escala, donde no se involucra a una persona sino a multitudes, cuando refiere a estructuras complejas, grandes o a un accidente en zona geográfica, estos hablan probablemente de desastres reales. En cambio los desastres no reales tienen como palabras claves las de asuntos más personales, individuales, emociones, o sea asuntos de menor escala. Esa es la tendencia que se divisa.

## Veracidad según ubicación

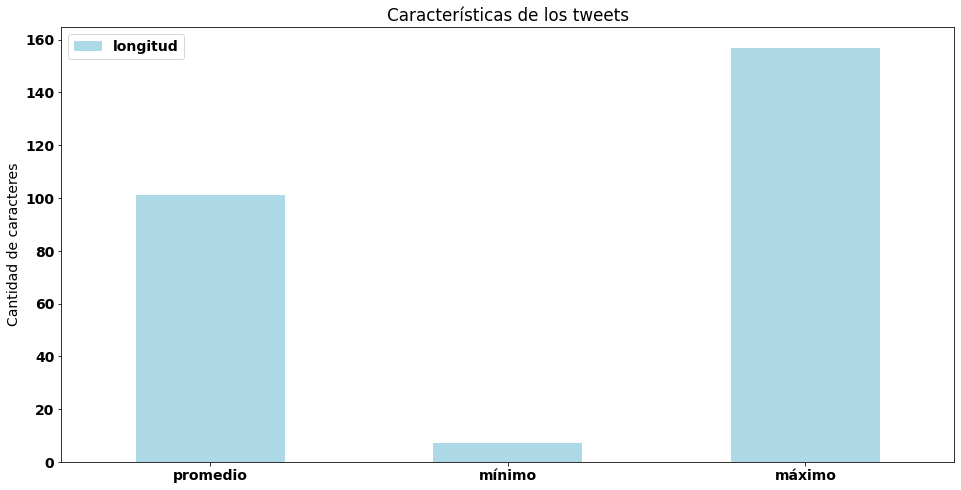
Análogamente a lo que se hizo respecto al porcentaje de veracidad de tweets según keyword, ahora lo hacemos respecto a ubicaciones (location). **¿Qué ubicaciones están más asociadas a desastres reales y cuáles a metafóricos?**  
 Hay locaciones muy pocos tweets así que vamos a quitar las locations con menos del 5% de ocurrencia respecto a la que está al tope con 104, o sea borro todos los tweets con las locations usadas 5 veces o menos.



No hay un gran descubrimiento aquí, sólo que las dos primeras locaciones son de India, si provienen de allí es muy probable que sean desastres reales. En tercer lugar está Nigeria. Tal vez se puede pensar una relación entre presencia de clases sociales bajas económicamente y la veracidad del tweet.

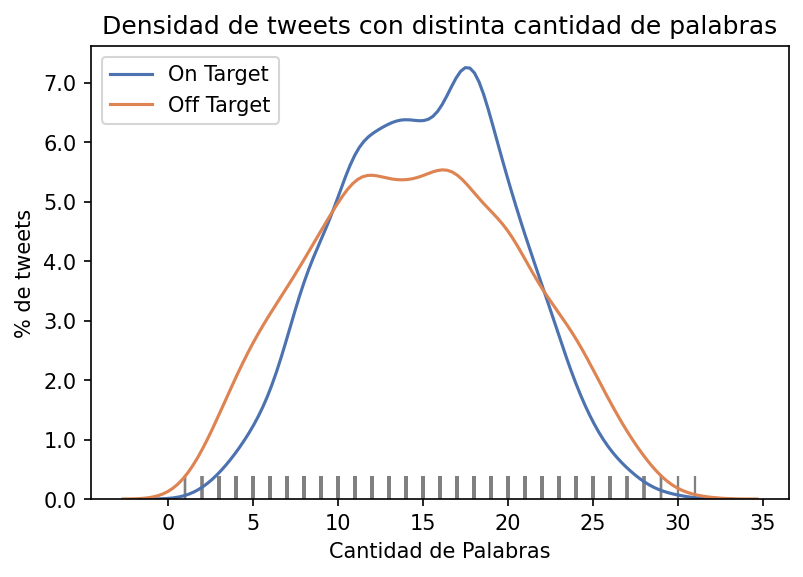
## Veracidad según longitud de los tweets

* Una de las primeras propuestas que se realizó, es sobre si los tweets más largos (en caracteres o palabras) tienen tendencias a hablar o no de desastres.

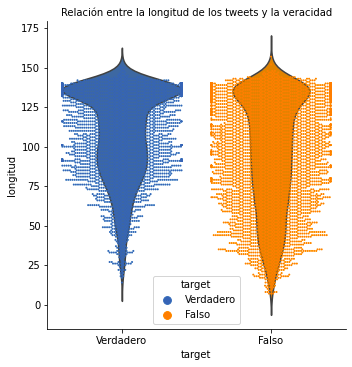


## 

* Es evidente que a mayor cantidad de caracteres, más probabilidad de hablar de un desastre real.

****

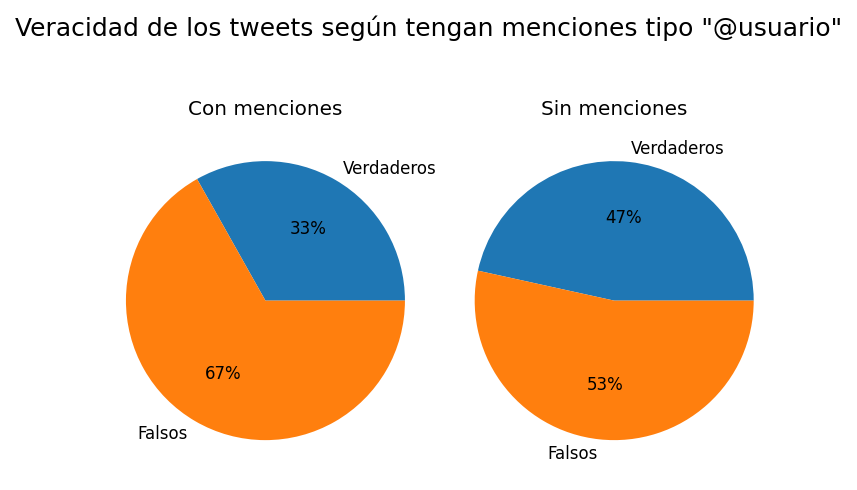
Nótese como los tweets sobre desastres reales, se acumulan más en la franja entre 10 y 23 palabras. Mientras que los tweets falsos, se distribuyen con un pico menos evidente.

****

En la visualización previa se combinó un violín plot con un swarm plot, observándose la concentración tweets verdadero y falsos para la escala de longitudes.

## Veracidad según se mencione a otros usuarios

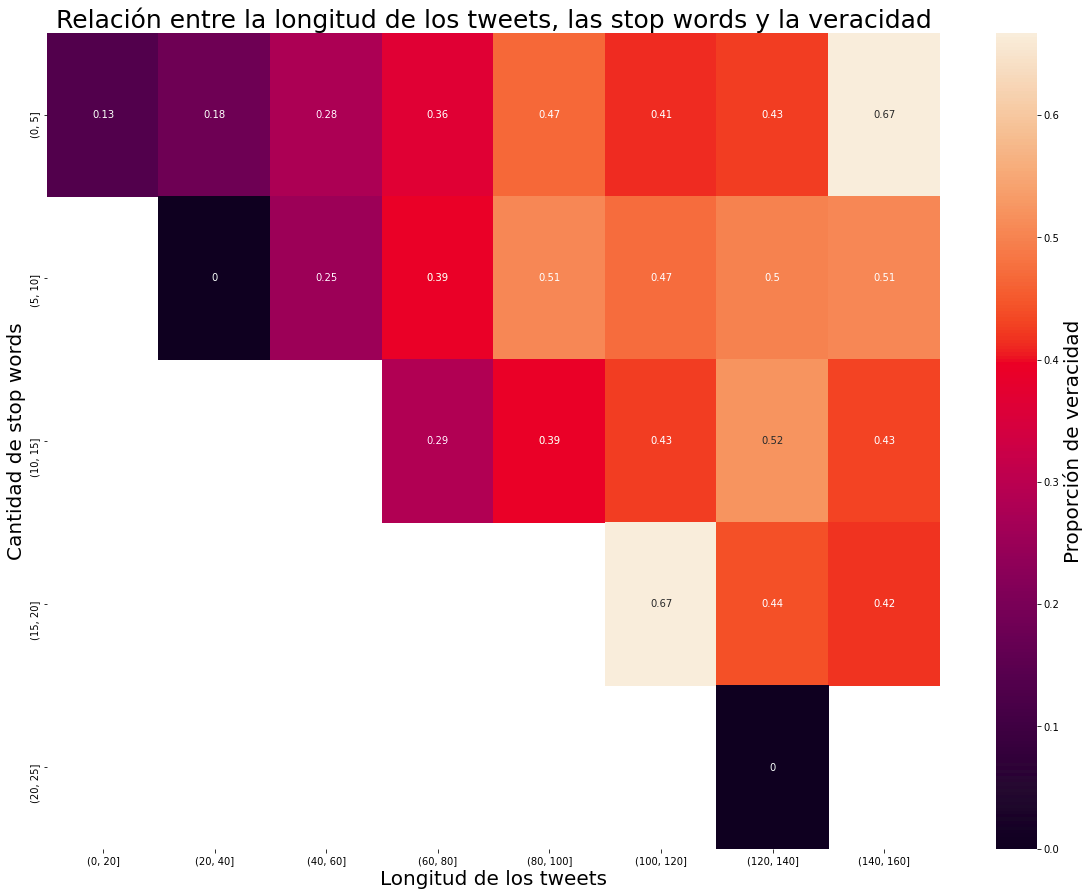
**¿Hay una relación entre la veracidad del desastre referido y la presencia o no de una mención con arroba, a una cuenta de Twitter, en el texto?**



Se evidencia que si se menciona otra cuentas en el texto, es más probable que refiera a desastre no real.

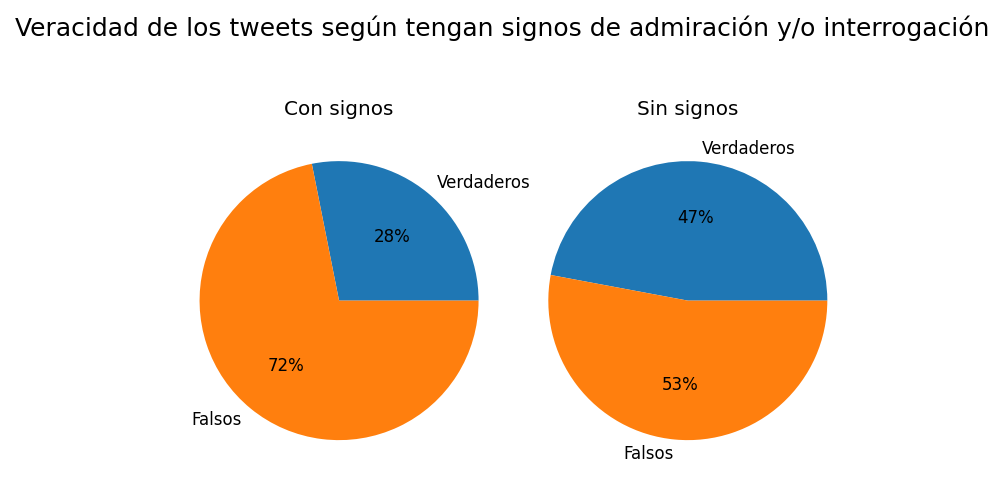
## Veracidad, longitud de tweets y cantidad de stop words

Para poder confeccionar un gráfico tipo heatmap se procedió a agrupar los valores de las columnas que contienen la longitud de los tweets y la cantidad de stop words en categorías, de manera de obtener valores de proporción de veracidad en cada una.



## Veracidad según cantidad de signos de interrogación y/o exclamación

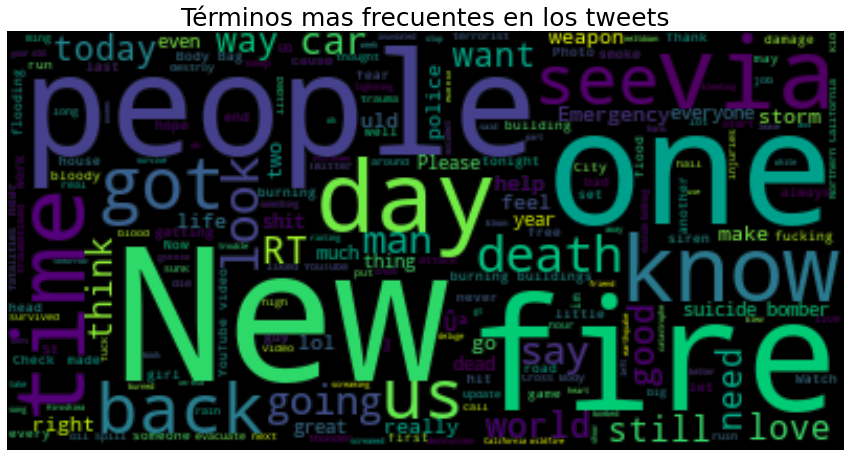
Una idea que se nos ocurrió, tiene que ver con que los Tweets sobre desastres naturales tienden a ser mensajes serios y de enunciación, no consulta, duda o exclamación. Por lo que podrían tener menos signos de exclamación y/o pregunta. A continuación, una visualización que muestra nuestro análisis en éste aspecto.



Claramente la presencia de signos de pregunta o admiración son un indicador de que el desastre referido tiene más chances de no ser real.

## Frecuencia de aparición de los términos

En una primera vista al contenido del campo text, puede observarse que muchos de los términos que se muestran como los más frecuentes son irrelevantes, al tratarse stop words, protocolos y dominios de internet, etc. Se procede a realizar una limpieza de dichas palabras.



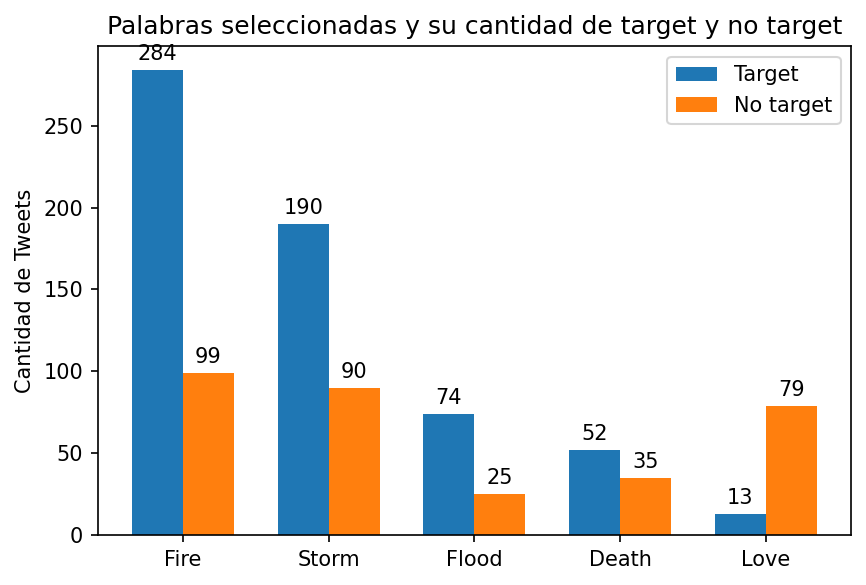
En el gráfico anterior pueden observarse los términos generales relevantes pero también las stop words como RT, us, one, etc.

Se procede a realizar el mismo tipo de gráfico, esta vez analizando cuales son las keywords más prevalentes.



De esta manera pueden visualizarse los términos que son más frecuentes, y que a la vez son más relevantes.

También seleccionamos algunos términos que nos parecían relevantes, viendo los Word Cloud creados y también proponiendo entre nosotros. Con ellos, generamos una visualización con resultados interesantes, donde se muestra la diferencia de cantidad entre tweets verdaderos y falsos, cuando aparecen estos términos.



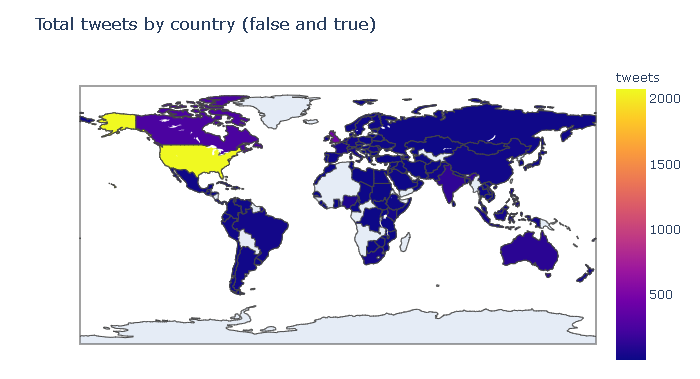
## Plots a nivel mundial de cantidad de tweets por país

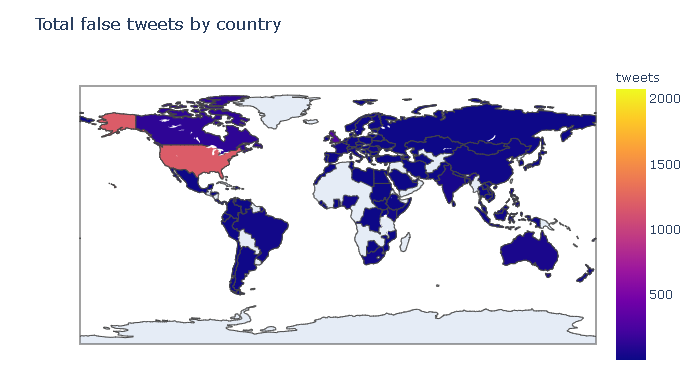
Primero en “tweets\_by\_country.ipynb”, se toman todos los tweets cuya location no sea nula, independientemente de si son reales o no, que produce un CSV con todos los datos de geolocalización que provee dicha API, de todos los tweets cuyo valor del campo “location” pudieron ser resueltos satisfactoriamente. Se usó la API de servicio de geolocalización Nominativ, que no requiere keys de conexión, pero se encontró que no resuelve los valores de “location” a nombres únicos de país para el mismo país (ejemplo: algunos valores de “location” resolvieron a “United States” y otros a “United States of America”).

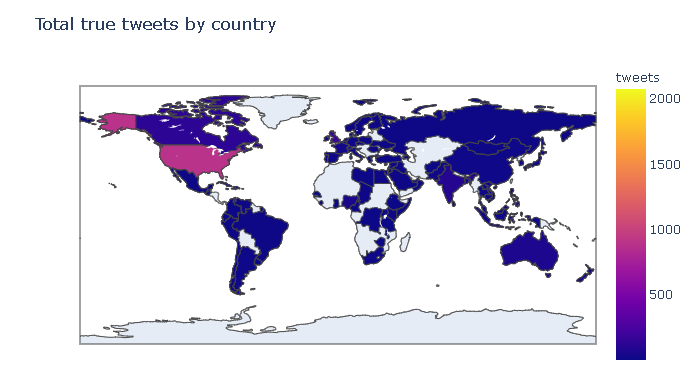
Para obtener un nombre de país único, se nota que convirtiendo la "location" a coordenadas primero, y luego haciendo un reverse lookup de esas coordenadas de la cual se extrae la "address" resultante, el país ahora sí tiene nombre único. Obviamente esto toma el doble de tiempo, ya que se hacen dos consultas para procesar cada “location” correctamente. Además, teniendo en cuenta la protección del servicio de geolocalización contra pedidos masivos (bulk requests), se usa un Rate limiter, que hace que en promedio el tiempo de procesamiento de cada transacción resulte ser de 1 segundo (ver detalles de su ejecución en “tweets\_by\_country-20200519.pdf”). Esto produce una salida intermedia, que luego se limpia eliminando valores nulos de países (ya que algunas “location” no resuelven a nada).

Luego, para hacer el gráfico de choropleth tipo mapamundi, se utiliza la biblioteca “plotly” que tuvo que ser instalada con pip. Pero la función utilizada para graficar dicho choropleth requiere que los países estén dados como código de tres caracteres alfabéticos (ISO-3166), por eso se requirió la instalación de la biblioteca “iso3166”, también utilizando pip.

Resueltas esas dependencias, se generaron tres archivos CSV. Uno con la cuenta de todos los tweets por país independientemente de su valor de verdad, y otros dos con la cuenta de todos los los tweets por país, en uno verdaderos, y en el otro falsos, respectivamente.

Utilizando dichos archivos, se ejecutó el notebook “tweets\_by\_country\_plots”, que genera los tres gráficos mostrados a continuación. **Cabe notar que dichos gráficos son interactivos en el notebook, y permiten hacer zoom, desplazamientos, y hover sobre los países para ver sus datos individuales.  
  
**

****

****

# 

# INSIGHTS OBTENIDOS

El primer dato como conclusión general es que sin necesidad de entender la sintáctica y semántica de los tweets, ni de estar en tema con las problemáticas locales o mundiales, hay muchos criterios fáciles de trabajar computacionalmente que sirven de indicio para determinar si los tweets refieren a un desastre real o no.

La lista siguiente corresponde a **indicios de veracidad**, dentro de este set de datos:

* Longitud extensa del tweet.
* Ausencia de menciones tipo @usuario.
* Ausencia de signos de interrogación y exclamación.
* Si provienen de India.
* Poca cantidad de stop words en relación a la longitud del tweet.
* Palabras claves asociadas a desastres en gran escala, donde no se involucra a una persona sino a multitudes, cuando refiere a estructuras complejas, grandes o a un accidente en zona geográfica (en contraste con palabras claves de asuntos más personales, individuales y emociones).

Con respecto a los términos más frecuentes:

* Dentro de los términos generales, no predominan temáticas o referencias específicas.
* Cuando se evalúan las keywords más frecuentes, se nota la prevalencia de las que refieren a diferentes tipos de desastres.

Algunas observaciones poco relevantes sobre el dataset.

* Predominan los desastres no reales (57%)
* El tweet promedio tiene 101 caracteres.
* Twitter es mayormente usado en U.S.A.