

Universidad de Buenos Aires

Facultad De Ingeniería

Año 2020 - 1er Cuatrimestre

**ORGANIZACIóN DE DATOS (75.06)**

TRABAJO PRÁCTICO N.º 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 01 – Luis Argerich

INTRODUCCION

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: En train.csv, indica si se trata de un desastre real (1) o no (0)

Se pretende descubrir cosas sobre los datos que puedan resultar interesantes.

Para la manipulación de los datos se hace uso de algunas bibliotecas como:

- Pandas

- Matplotlib

- Numpy

- Seaborn

- Geopy

- Wordcloud

- Nltk

- Collections

- Squarfy

VISTA 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 tienen valores nulos.

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

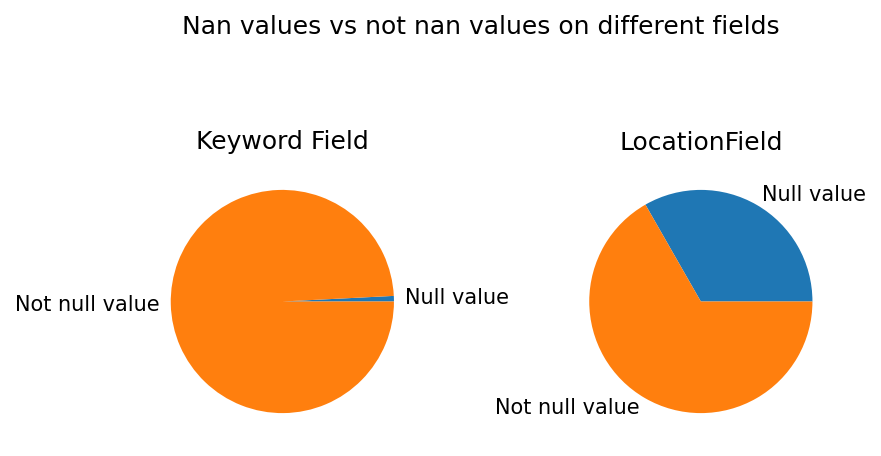
0 4342

1 3271

Name: target, dtype: int64

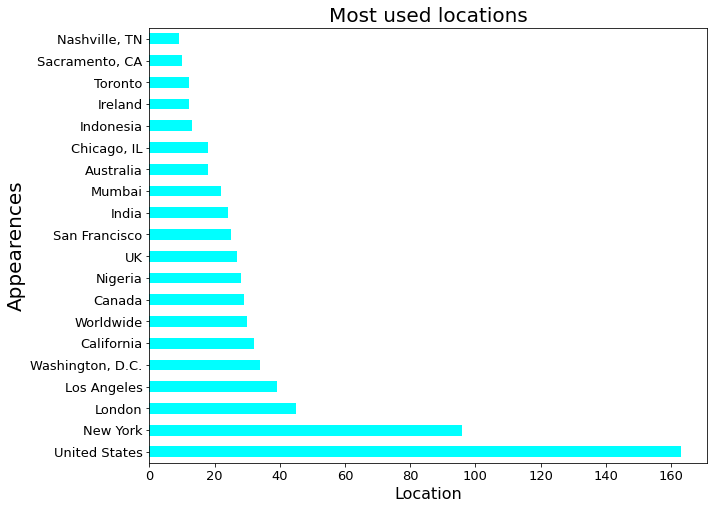
* Valores nulos.

Se hecha un vistazo al predominio de valores nulos, los cuales se hallan en dos de los campos del dataset.



* Locaciones

De la información del dataset se observa que es imprecisa la clasificación de ubicaciones, ya que contiene distintas categorías: hay países, estados, ciudades e incluso hay uno que es el mundo entero (worldwide). Se procede a realizar una limpieza



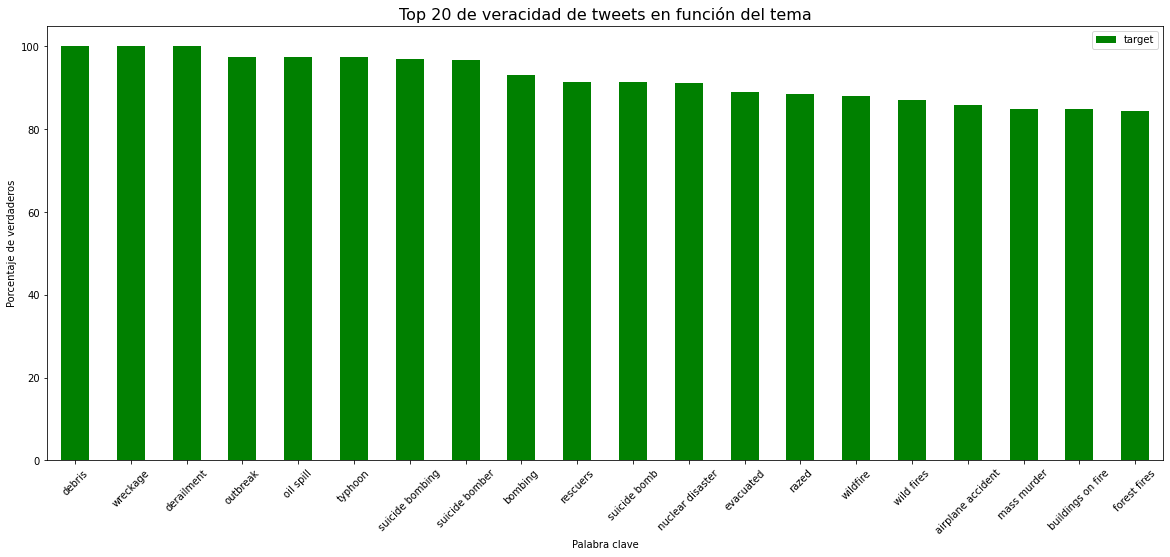
ANALISIS DE VERACIDAD DE LOS TWEETS

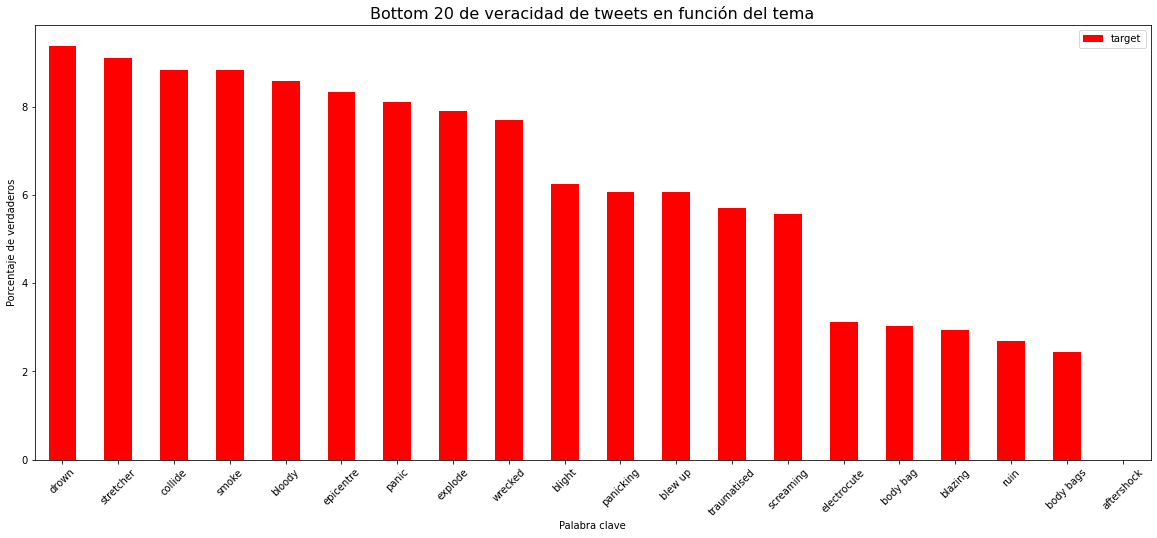
* Porcentaje de veracidad de tweets según keyword.

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, tengo que hacer un gráfico con el top 20 y el bottom 20, algo así, sino sería mala la visualización



****

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).

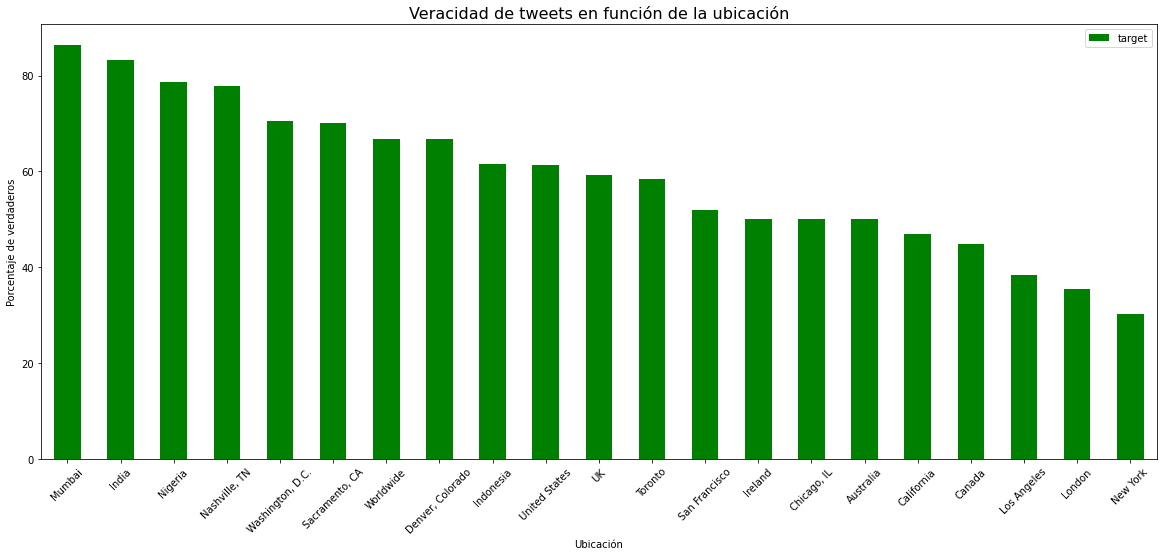
La de tweets sobre desastres no reales tiene temas como drown (ahogar o empapar), stretcher (camilla), collide (colisionar), bloody (sangrante), panic (pánico), explode (explotar), wrecked (borracho), screaming (gritando), electrocute (electrocución), body bag (bolsa de cadáver), aftershock (que no es 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 casi seguro 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.

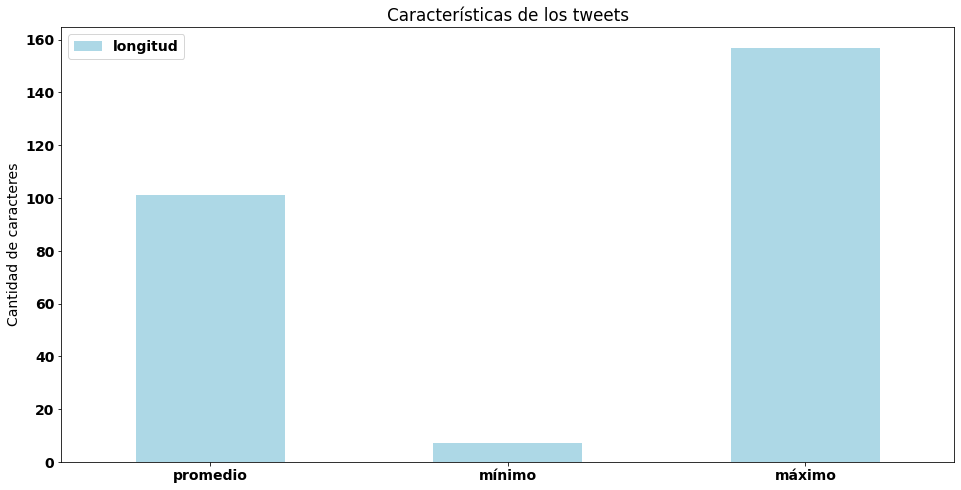
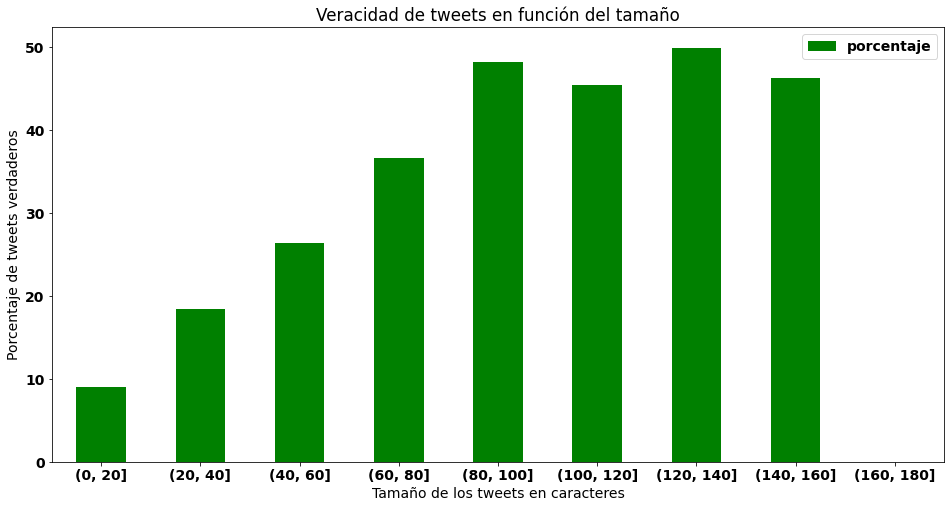
Analogamente a lo que se hizo respecto al porcentaje de veracidad de tweets según keyword, ahora lo hacemos respecto a ubicaciones (location). **¿Qué ubicaciones están más asociados a desastres reales y cuáles a metafóricos?**

Hay locaciones con un sólo tweet 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 todas los tweets con las locations usadas 5 veces o menos.



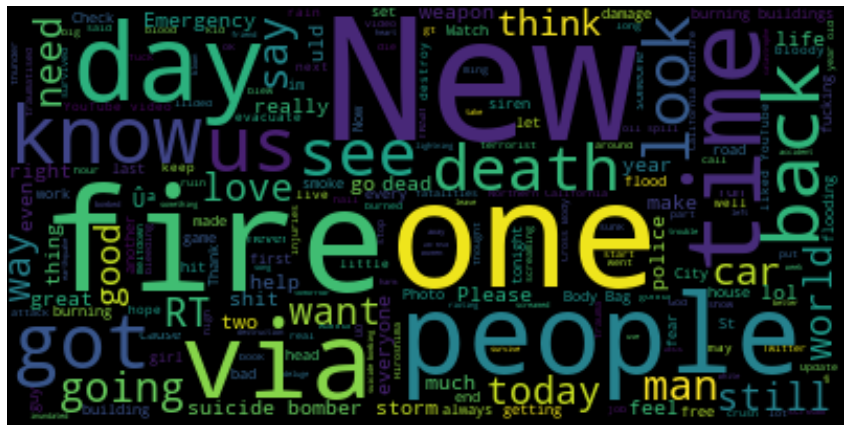
Puede observarse que los tweets que provienen de India tienen un porcentaje de veracidad mayor que los de Estados Unidos, los cuales son los que predominan.

* Veracidad según longitud de los tweets.



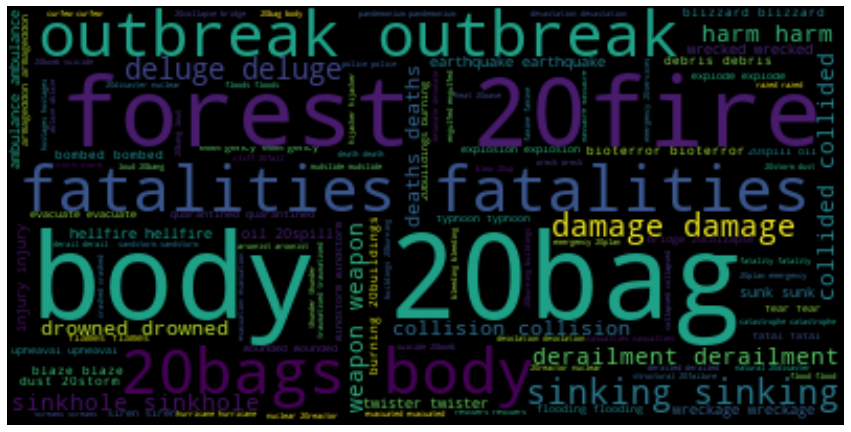
FRECUENCIA DE APARICIÓN DE LOS TÉRMINOS

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.

A continuación se procede a analizar las keywords mas prevalentes.



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

INSIGHTS