FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico

N°1

"Análisis exploratorio de un set de datos"

- **Carrera:** Ingeniería en Informática.
- **Materia:** Organización de Datos.
- **Profesor:** Argerich, Luis
- **JTP**: Golmar, Natalia.
- **Cuatrimestre**: 1er Cuatrimestre 2020.
- **Fecha de entrega:** 21/05/2020
- **Nombre del grupo:** *Team_Undav*
- **Integrantes:**
    - Calonge, Federico Matías.
    - Ceballos Pardo, Sarah.
    - Flores, Matías.
    - Loiseau, Matías.
- **Repositorio - Colab:**

  https://colab.research.google.com/drive/1AvotyxAm398jwVb6nPAUyh6v2SYBqRkS?usp=sharing
- **Repositorio - Github:**

  https://github.com/MatiasLoiseau/TP1-OrganizacionDeDatos

Índice

Índice **1**

1-Objetivo **2**

2-Introducción y conceptos previos **3**

2.1-Conceptos de estadística 3

2.2-Librerías de Python utilizadas 5

2.3-Importancia del análisis exploratorio 6

2.4-Dataset 7

3-Desarrollo **8**

3.1-Formatos y tipos de datos 8

3.2-Primeros resúmenes estadísticos 8

3.3-Preguntas y visualizaciones 10

4-Conclusión **19**

5-Bibliografía **20**

1-Objetivo

El objetivo de este Trabajo Práctico N°1 consiste en **realizar un análisis exploratorio de un set de datos**(*dataset*). Dicho dataset lo obtenemos de la siguiente competencia: https://www.kaggle.com/c/nlp-getting-started; está descripto más detalladamente en la sección ***datasets e importancia del análisis de datos.***

En este trabajo buscamos comprender y entender los datos mediante distintos análisis estadísticos que aplicaremos sobre los datos de nuestro dataset. Esto lo haremos para encontrar relaciones entre nuestras columnas / atributos.

Dependiendo de cada análisis que se haga, se buscará de mostrar los resultados en base a **visualizaciones**, las cuales son de gran importancia, ya que nos permitirán transmitir información que no es fácilmente apreciada por las personas cuando se observan los datos de una manera más "cruda".

2-Introducción y conceptos previos

En esta Sección describiremos los temas teóricos y matemáticos para tratarlos a lo largo del Informe y llevarlos a cabo mediante algoritmos de Python.

2.1-Conceptos de estadística

-Muestras y Población:

Se los denomina **muestras** o **valores** a todos los datos obtenidos, mientras que, al conjunto de estos se los denomina **población**. Esta población estará conformada por las N muestras o valores. Para nuestro dataset estas muestras son cada uno de los tweets.

-Media:

Es una herramienta que permite describir la **tendencia de un conjunto de N valores o muestras**, que se encuentran dentro de una **población**. Matemáticamente, la media es la suma de cada uno de los valores, dividido por el total N. La *Fórmula 1* describe el cálculo de la media para una distribución NORMAL (generalmente los datos siguen esta distribución).

$$\mu = \frac{1}{N} \cdot \sum_{i}^{N} X_i$$

Formula 1

-Donde:

μ = Media de la muestra.

N = Cantidad de valores de la muestra ó población obtenida.

X_i = Muestra i-ésima.

Desvío:

Herramienta que muchas veces no resulta suficiente para describir a un conjunto de datos. Por esta razón, se puede utilizar el desvío estándar, el cual describe cuán dispersos se encuentran los valores respecto de la media. Cuánto más grande sea el desvío estándar, más grande será la dispersión entre los valores de la población. Matemáticamente se la describe en la *Fórmula 2*.

$$\sigma = \sqrt{\frac{1}{n} \cdot \sum (x_i - u)^2}$$

Fórmula 2

-Donde:
σ = Desvío estándar μ = Media de la muestra.
N = Cantidad de valores de la muestra o población obtenida.
X_i = Muestra i-ésima.

Varianza:

Similar a la Varianza, ya que también mide la dispersión de los valores, pero matemáticamente se la obtiene de distinta forma (observar *Fórmula 3*).

$$\sigma^2 = \frac{1}{n} \cdot \sum (x_i - u)^2$$

Fórmula 3

Donde:
σ^2 = Varianza.

Es decir, que si bien ambos conceptos miden la dispersión de las muestras de la población, el desvío estándar tiene las mismas unidades que la media, por eso es que muchas veces se suele utilizar el desvío a la varianza. En conclusión la varianza mide algo similar al desvío pero en otras unidades.

2.2-Librerías de Python utilizadas

Utilizaremos Las siguiente bibliotecas en Python para realizar nuestro análisis exploratorio de los datos:

Numpy: Paquete fundamental para la computación científica con Python. Consiste en una extensión de Python, que le agrega mayor soporte para vectores y matrices de n-dimensiones, constituyendo una biblioteca de funciones matemáticas de alto nivel para operar con esos vectores o matrices.

Pandas: Librería de Python que depende de Numpy. Proporciona estructuras de datos llamadas dataframes a Python, para facilitarnos la manipulación de los datasets. Es una librería muy sencilla de usar, versátil y flexible: permitiéndonos cambiar la forma a los datasets si así lo necesitamos, escribir datos en diferentes formatos (CSV, XLS, SQL, etc.), filtrar información, fusionar y unir datos, realizar operaciones matemáticas con los campos , etc.

Los principales tipos de datos que pueden representarse con Pandas son:

- Datos tabulares con columnas de tipo heterogéneo con etiquetas en columnas y filas. Se los denomina Dataframes.
- Series temporales.
- Panels (tablas 3D).

Matplotlib: Librería de Python para la generación de gráficos a partir de datos contenidos en listas, arrays, diccionarios, Data Frames o Series. Permite analizar el comportamiento de un conjunto de datos / dataset y cómo éste evoluciona en una medida de tiempo o en función de una o más variables. Además, permite poder determinar la distribución que se asemeje al conjunto de datos, es decir, si sigue un comportamiento lineal, exponencial, distribución normal, entre otras.

Seaborn: Librería de Python basada en Matplotlib utilizada para la visualización de datos. Provee una interfaces de alto nivel para dibujar atractivos e informativos gráficos estadísticos.

WordCloud: Librería de Python para generar una nube de palabras.

2.3-Importancia del análisis exploratorio

El objetivo del análisis exploratorio es entender los datos con los que se va a trabajar para llegar a un objetivo determinado. Sin embargo, en este Trabajo solo nos encargaremos de realizar únicamente el análisis sin llegar a elaborar ningún tipo de herramienta de predicción. El típico procedimiento del análisis exploratorio de datos consiste en formular preguntas interesantes y mediante análisis estadísticos, algoritmos y visualizaciones poder responderlas. De esa manera, se pueden detectar características y relaciones entre nuestros datos, e irregularidades en los mismos.

2.4-Dataset

El dataset que analizaremos en este Trabajo Práctico es un archivo CSV. Será el archivo 'train.csv' de https://www.kaggle.com/c/nlp-getting-started/data?select=train.csv, el cual cuenta con datos de tweets de distintas personas hechas en distintas ubicaciones('location') y que fueron o no tweets que tratan sobre desastres; esto es dependiendo el valor de la columna 'target': '1' si fueron realmente sobre desastres, o '0' si no lo fueron.

Este Dataset contiene 7613 filas y 5 columnas (Cada columna representa un feature de nuestros tweets); podemos observar como ejemplo un par de registros de este dataset en la Imagen 1.

id	keyword	location	text	target
32			London is cool ;)	0
33			Love skiing	0
34			What a wonderful day!	0
36			LOOOOOOL	0
37			No way...I can't eat that shit	0
38			Was in NYC last week!	0
39			Love my girlfriend	0
40			Cooool :)	0
41			Do you like pasta?	0
44			The end!	0
48	ablaze	Birmingham	@bbcmtd Wholesale Markets ablaze http:/	1
49	ablaze	Est. September 2(	We always try to bring the heavy. #metal #	0
50	ablaze	AFRICA	#AFRICANBAZE: Breaking news:Nigeria flag	1
52	ablaze	Philadelphia, PA	Crying out for more! Set me ablaze	0
53	ablaze	London, UK	On plus side LOOK AT THE SKY LAST NIGHT	0
54	ablaze	Pretoria	@PhDSquares #mufc they've built so much	0
55	ablaze	World Wide!!	INEC Office in Abia Set Ablaze - http://t.co	1
56	ablaze		Barbados #Bridgetown JAMAICA Â‰Ã›Ã' T	1
57	ablaze	Paranaque City	Ablaze for you Lord :D	0
59	ablaze	Live On Webcam	Check these out: http://t.co/rOI2NSmEJJ h	0
61	ablaze		on the outside you're ablaze and alive	0
62	ablaze	milky way	Had an awesome time visiting the CFC hea	0
63	ablaze		SOOOO PUMPED FOR ABLAZE ???? @south	0

***Imagen 1**. Data Set*

Columnas de nuestro Data Set:

- id - identificador único para cada tweet.
- keyword - una palabra clave del tweet (puede no tener valor: Nan).
- location - la ubicación de donde el tweet fue enviado (puede no tener valor: Nan).
- text - texto del tweet.
- target - permite saber si el tweet trata sobre desastres (1) o no (0).

3-Desarrollo

3.1-Formatos y tipos de datos

Antes de empezar con el análisis exploratorio de datos, hay que saber en qué formato y con qué tipos de datos vamos a trabajar. Como se dijo anteriormente, usaremos la biblioteca Pandas que sirve para trabajar con *Data Frames.* Básicamente un Data Frame es el equivalente a una tabla donde tenemos un conjunto de datos separados por filas y que cada columna es un atributo que define cada dato. Por esa razón cada dato de una misma columna tiene que ser de un mismo tipo de dato. Sumando a estas características, los Data Frames manejan índices y subíndices, creando así distintas formas de flexibilizar la manipulación de datos.

Cargando los datos del csv y creando un Data Frame con los valores estándar, los índices, columnas y datos tienen el siguiente tipo de dato:

- Índice: Un rango que empieza desde 0 hasta 7613 con pasos que suman de a 1.
- Columnas: Las columnas son de tipo de datos objetos que son juntos son una lista de índices llamados: ['id', 'keyword', 'location', 'text', 'target'].
- Datos: Son un conjunto de filas donde los tipos de datos dependiendo de cada columna son:
    - ‘id’: Números enteros.
    - ‘keyword’: Objeto.
    - ‘location’: Objeto.
    - ‘text’: Objeto.
    - ‘target’: Números enteros.

3.2-Primeros resúmenes estadísticos

Antes de profundizar en el análisis, haremos un primer acercamiento a los datos para ver con cuántos valores maneja el dataset, y si hay valores nulos.

Gracias a los análisis de los tipos de datos y más específicamente el índice, podemos saber que hay 7613 datos en el dataframe. Pero esto no garantiza que haya valores nulos. Entonces se debe contar la cantidad de datos por columna.

Columna	Cantidad de valores
id	7613
keyword	7552
location	5080
text	7613
target	7613

Tabla 1

Según la información de la tabla 1, se puede observar que no todos los datos están completos y hay valores nulos, donde la columna 'location' es la más afectada.

Continuando con los primeros análisis, podemos ver cuántos tweets hay que hablen de desastres y cuántos no.

Target	Cantidad de valores
0	4342
1	3271

Tabla 2

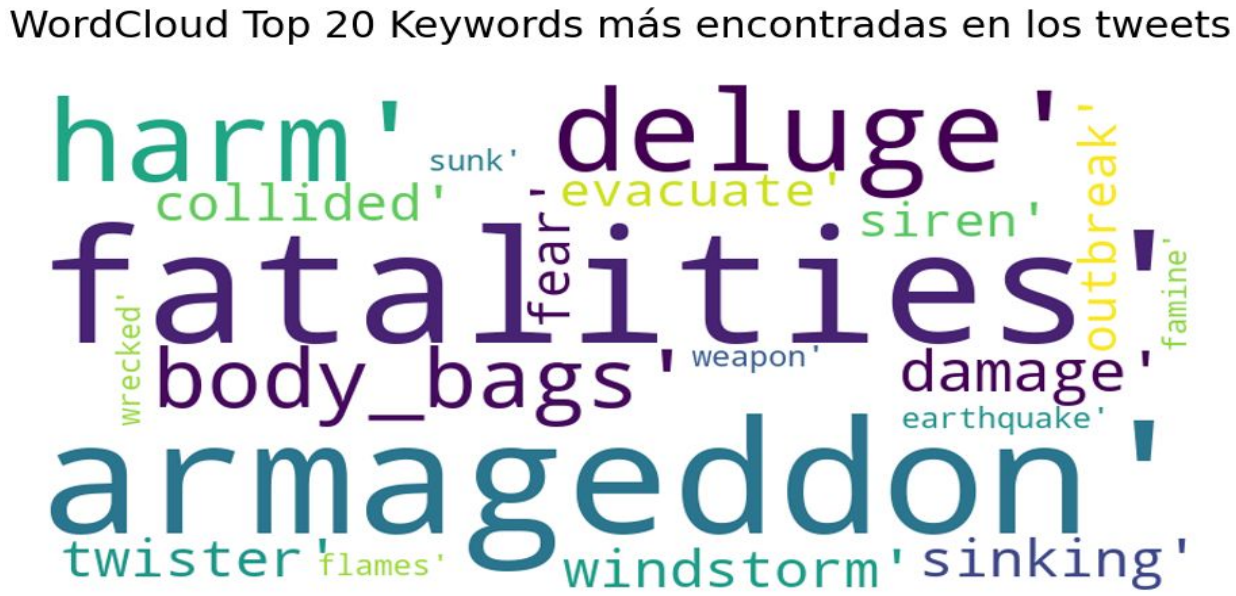
En la tabla 2 se observa que hay un poco más de mil datos de diferencia, teniendo más tweets que no tratan sobre desastres. Este análisis es importante ya que en los algoritmos de Machine Learning, en la parte de entrenamiento, el sistema va a tener más recursos de aprendizaje del target 0 que del target 1. Esto podría generar un significativo desbalance que los científicos de datos deben tener en cuenta.

3.3-Preguntas y visualizaciones

Luego de realizar nuestros primeros análisis estadísticos de los datos, nos planteamos distintas preguntas interesantes para responderlas mediante visualizaciones.

1.¿Cuáles son las keywords (palabras claves) con mayor aparición en los tweets?

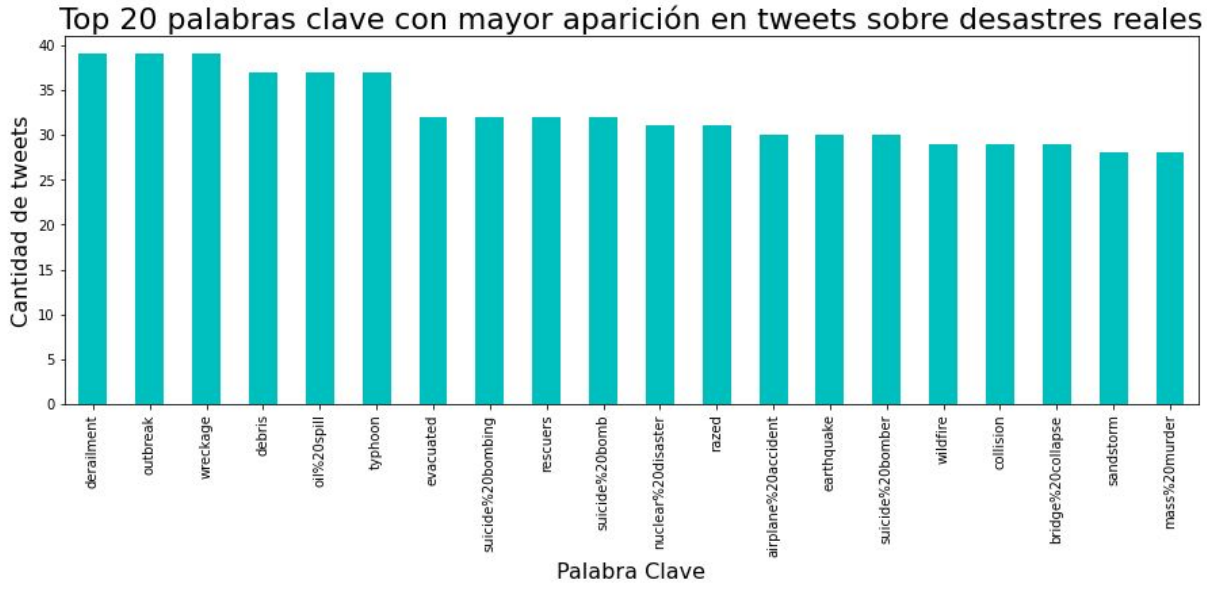
Decidimos realizar una nube de palabras para la visualización de este set de datos, debido a que nos pareció una mejor manera de demostrar la aparición de las palabras claves. De esta forma se demuestra en la nube que las palabras más grandes son las que tienen una mayor aparición y en medida que van disminuyendo tienen una menor aparición.

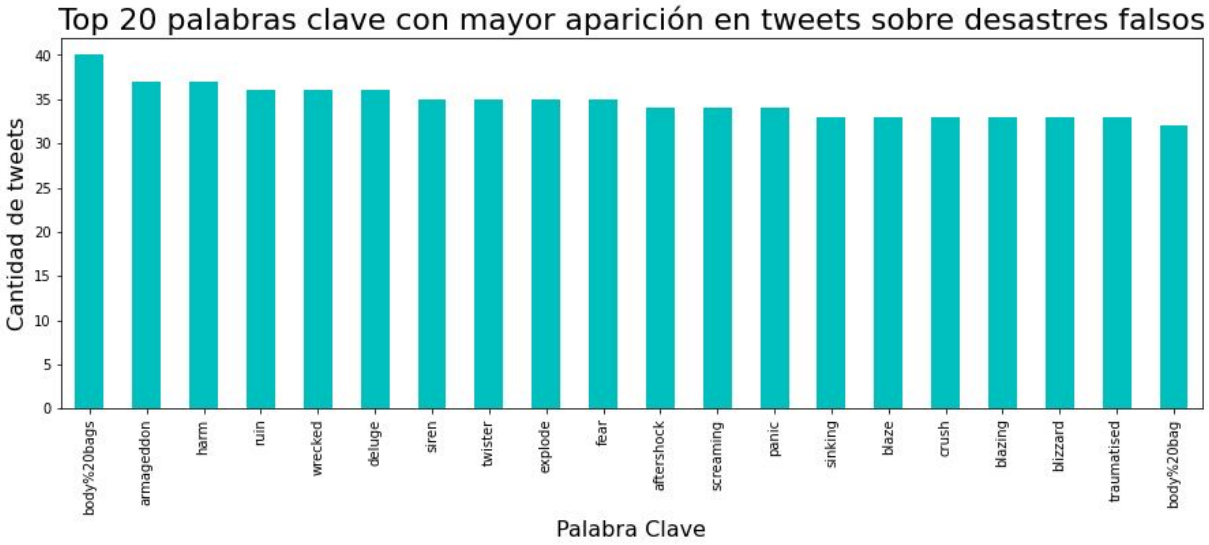


Además de la nube de palabras, realizamos un gráfico de barras en el cual se pueden apreciar las 20 palabras con mayor aparición. También, se puede notar que no es tanta la diferencia de cantidad entre cada una.

2.¿Varía la aparición de las keywords según el tipo de noticia (Target)?

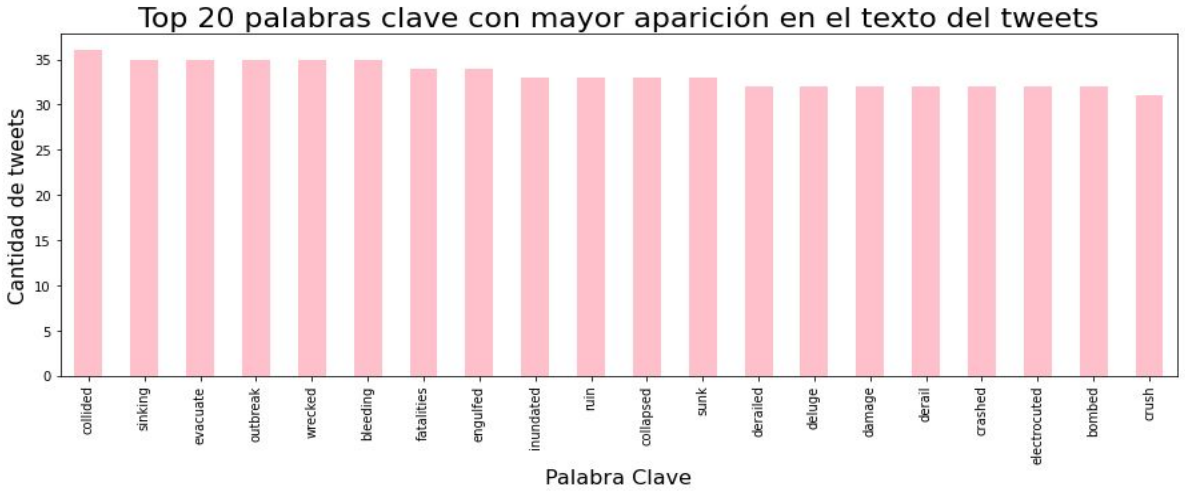
Podemos apreciar que el top de palabras cambian drásticamente, siendo que la palabra clave que aparece en mayor cantidad de tweets sobre desastres reales(derailment) es la número 20 entre todas las palabras. Mientras que la palabra con más aparición en los tweets en general (fatalities) no aparece en ninguno de los siguientes gráficos. Esto es debido a que esa palabra aparece con mayor distribución en ambos tipos de tweets.





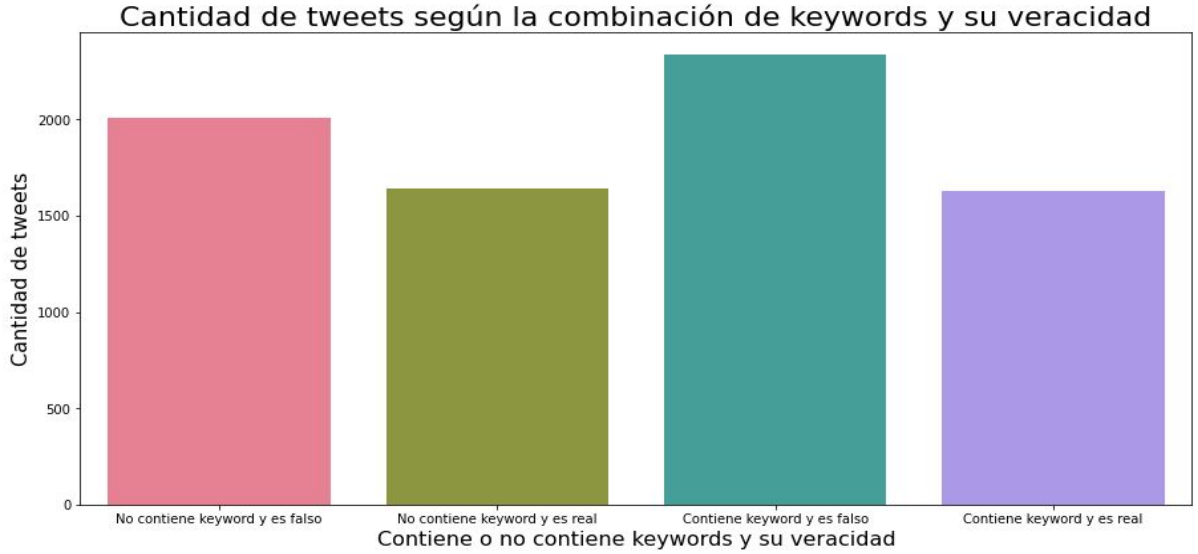
3.¿Cómo influye si la keyword aparece en el texto del tweet?

Vemos que la aparición de las keywords en el texto influye demasiado en el gráfico. En este tenemos como palabra con más aparicion a “collided”, la cual anteriormente aparecía número 12. Además, la aparición de las palabras varía drásticamente a comparación del gráfico de todas las palabras.



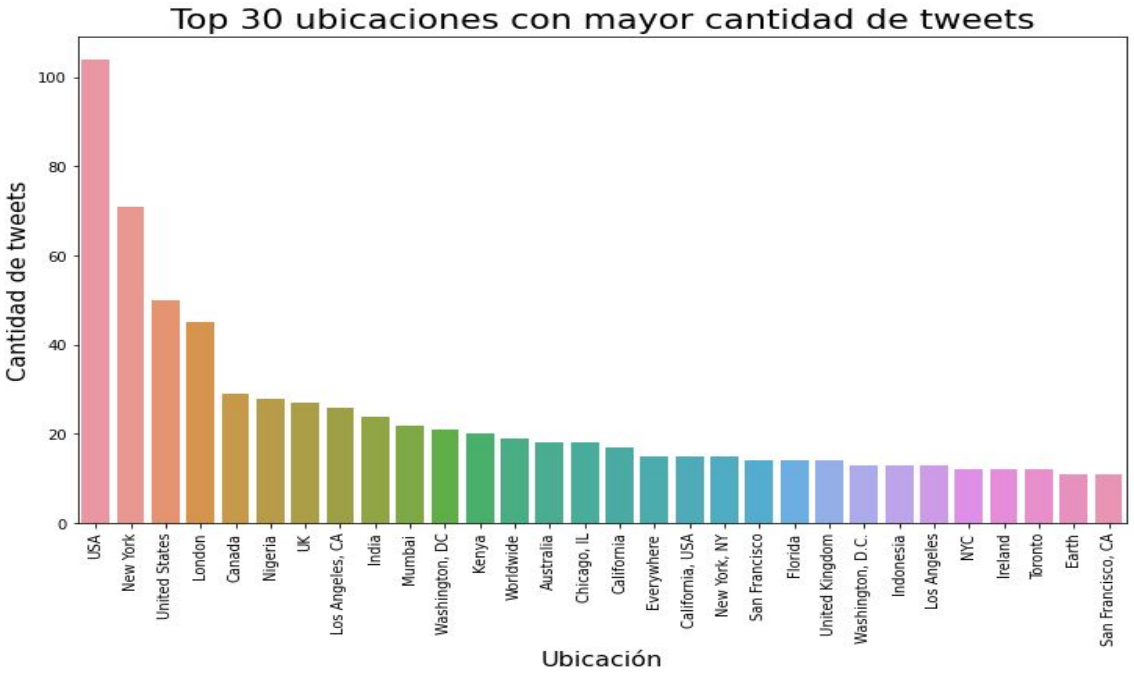
4.¿Cómo están distribuidos los tweets según su keyword y su valor de target / veracidad?

En la siguiente visualización analizamos la cantidad de tweets según su combinación de keyword (si posee o no keyword) y su valor de target (si se trata de un tweet de desastre real o no). Este gráfico lo consideramos importante, ya que se observa que los valores de la cantidad de tweets según dicha clasificación, tienen una distribución parecida. Esto podría ayudarnos mucho en la segunda etapa del trabajo, en el proceso de entrenamiento del algoritmo de Machine Learning para predecir su veracidad.



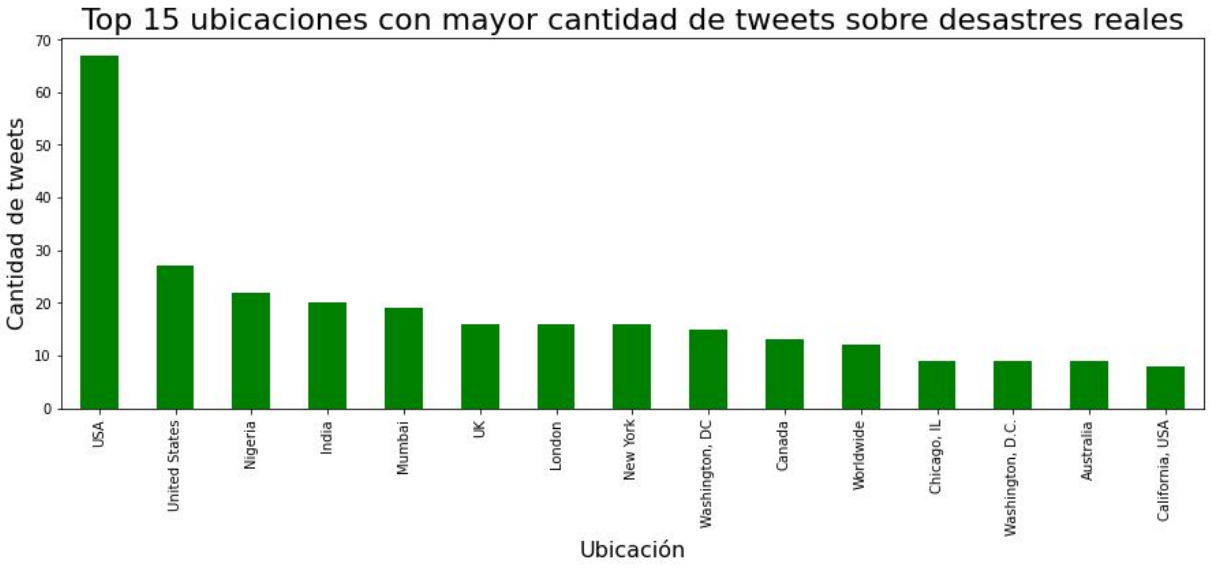
5.¿Cuáles son las Top 30 ubicaciones con mayor cantidad de tweets?

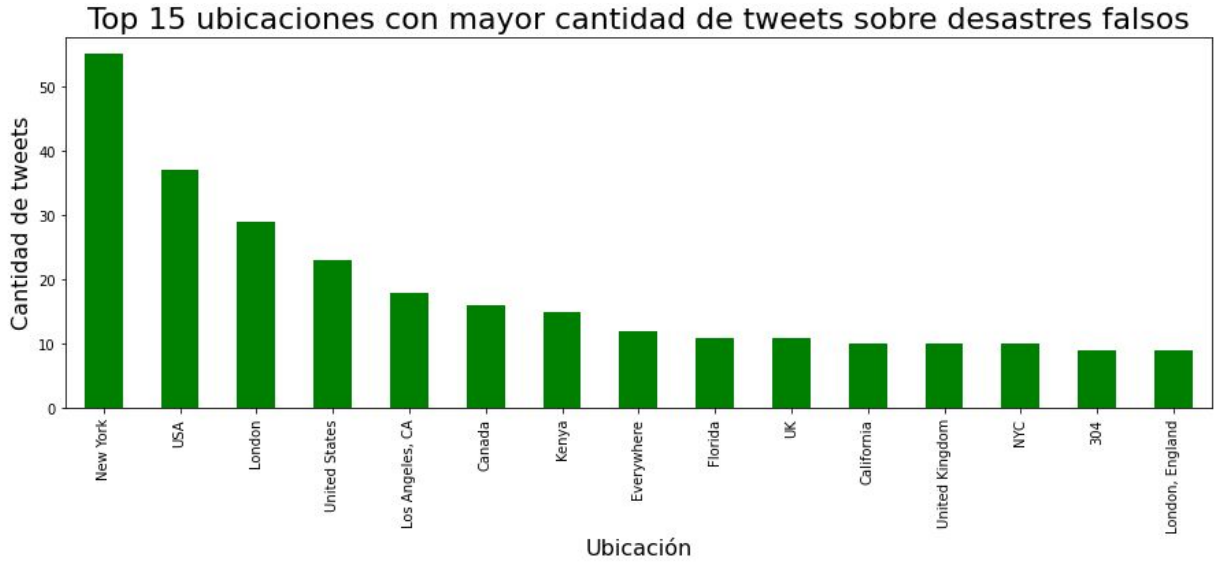
En el siguiente gráfico de barras podemos apreciar las 30 ubicaciones con mayor aparición en los tweets. En primer lugar tenemos a USA con 104 apariciones y podemos notar que va descendiendo exponencialmente, siendo que a partir del puesto 20 aparecen en menos de 20 tweets.



6. ¿Varía la aparición de las keywords según el tipo de noticia (Target)?

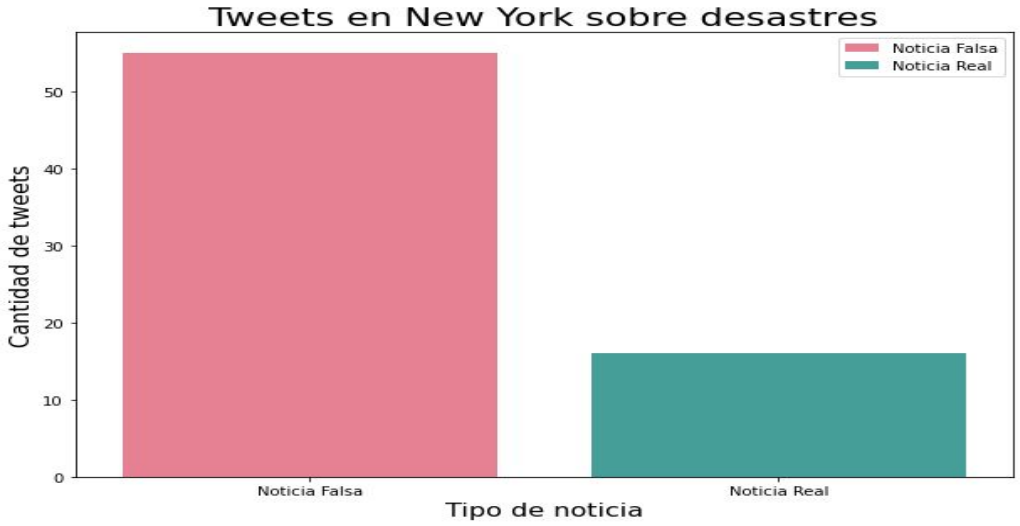
Vemos que con la división entre reales y falsos afectó a la lista de elementos. En el caso de los tweets reales, USA sigue siendo el que aparece en mayor cantidad de tweets con casi 70 tweets, mientras que en el de los falsos, el que tiene más apariciones es New York con más de 50 tweets.





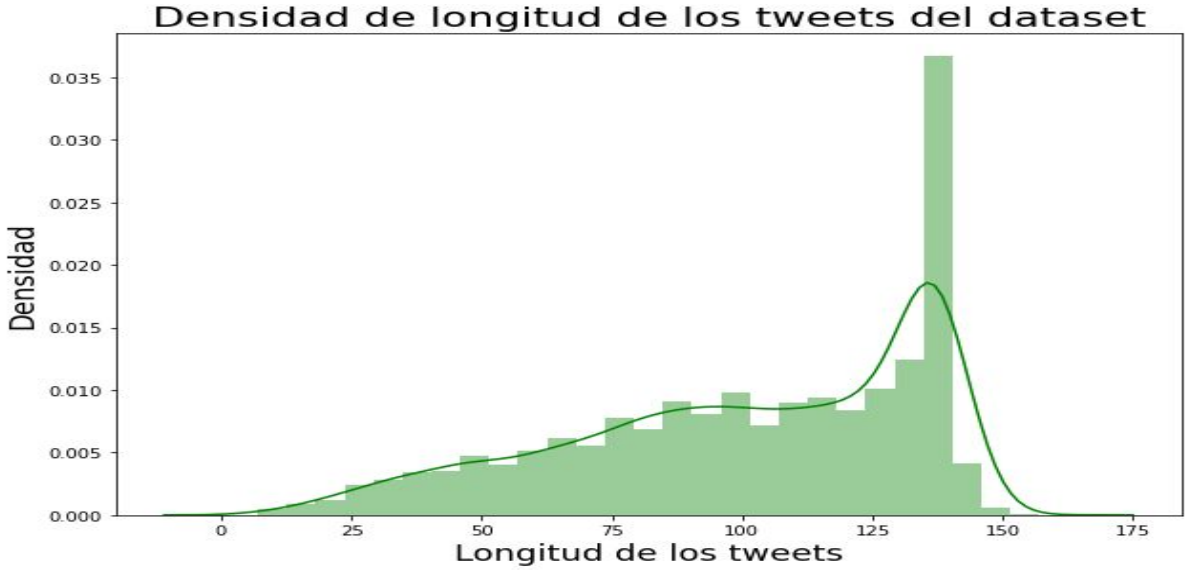
7.¿Cuál es la veracidad de los tweets con ubicación New York?

La siguiente visualización es un barplot en la cual comparamos la cantidad de tweets según el tipo de noticia (verídica = azul / falsa = rosa). Se puede observar que más del 70% de los tweets con ubicación New York son falsos sobre desastres.



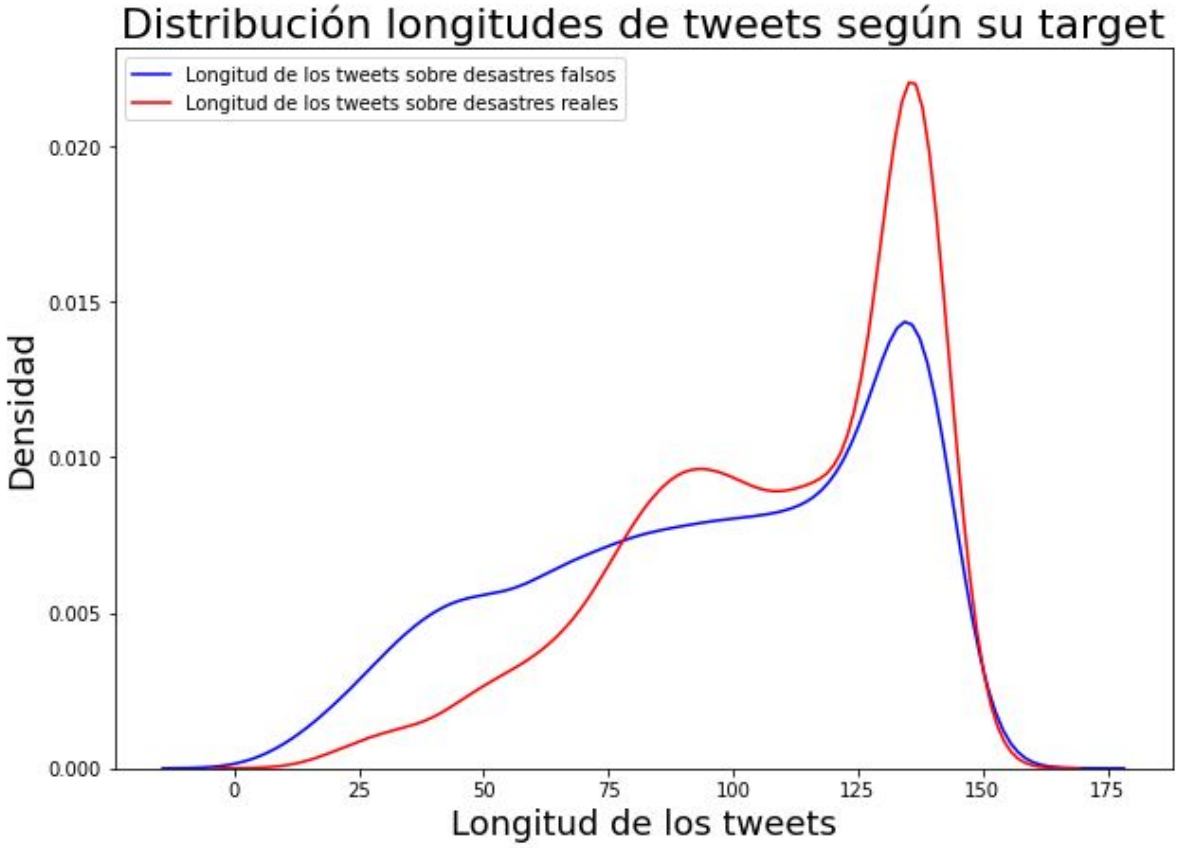
8.¿Cuáles son las longitudes de nuestros Tweets?

La siguiente visualización nos dá información de densidad de los tweets. De esta manera, observamos que hay una gran concentración de tweets que rondan los valores de 125 a 140 caracteres. Además cabe destacar que los tweets con 136 a 140 caracteres, componen el 17% del dataset.



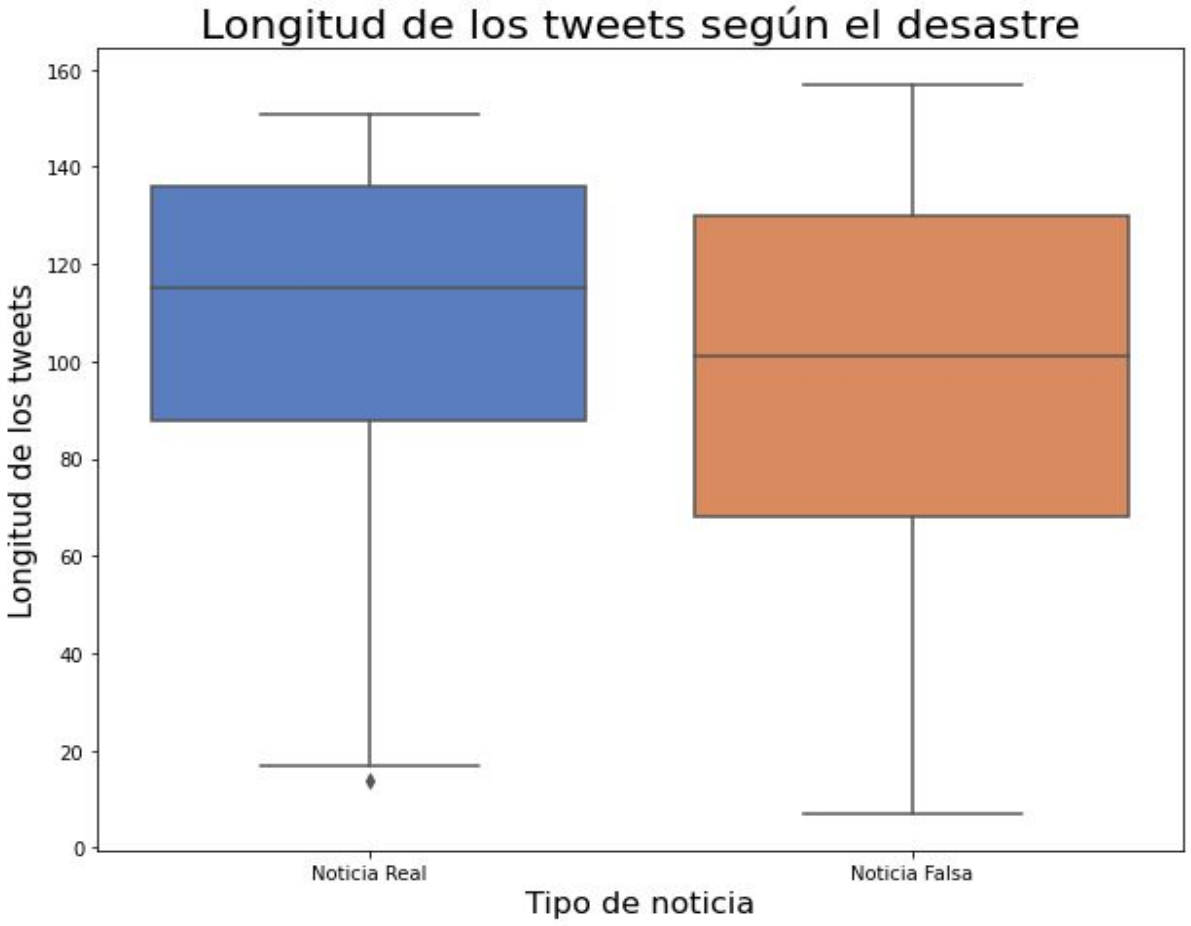
9.¿Cuáles son las longitudes de nuestros Tweets según su veracidad?

En el siguiente gráfico observamos que los tweets verídicos (que tratan sobre desastres reales) tienden a ser más extensos que los NO verídicos. Además podemos observar que siguen compartiendo el mismo patrón que vimos en la visualización de la 6ta pregunta: la mayor densidad de longitud de tweets varía entre los 120 y 140 caracteres.



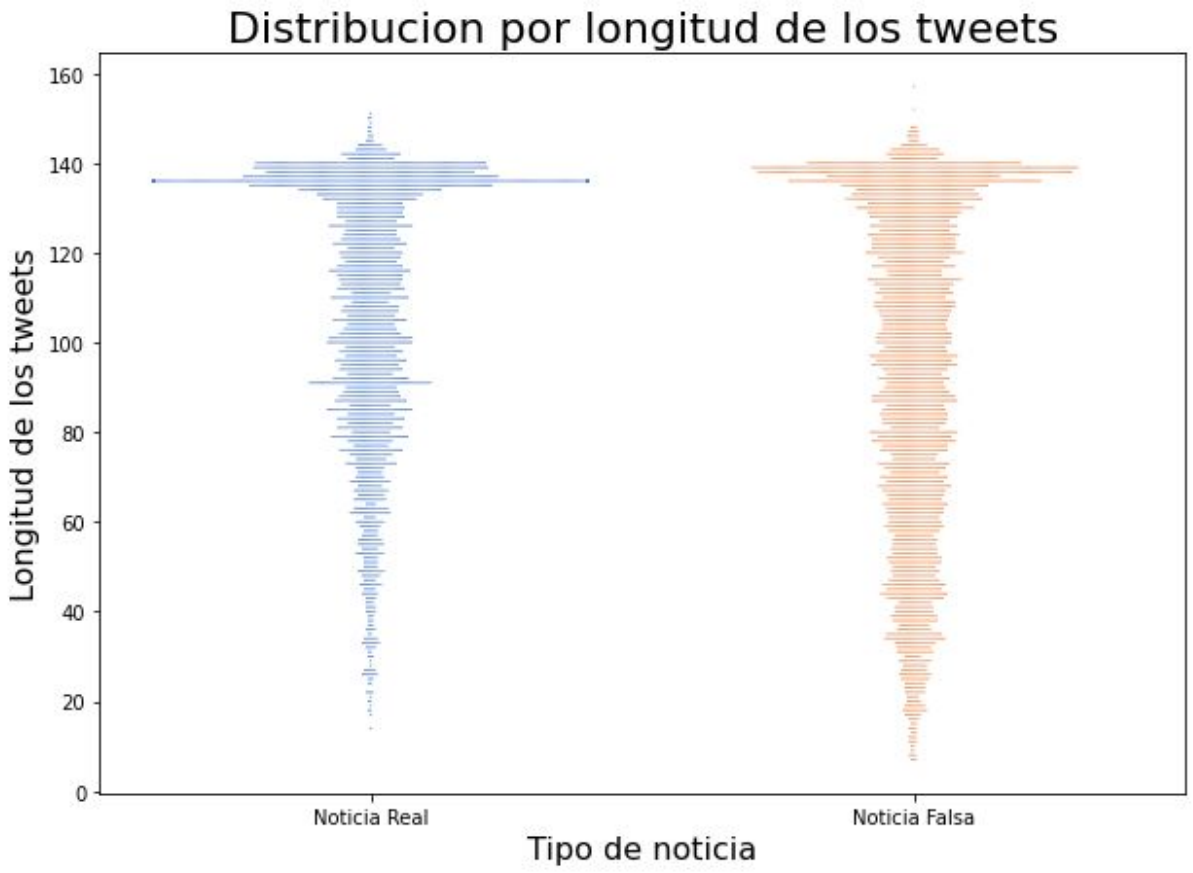
10.¿Cómo está distribuida la longitud de los tweets según su veracidad?

En la siguiente visualización podemos observar que hay una concentración muy grande de tweets verídicos de longitud entre 90 y 135 caracteres aprox.; estos tweets tienen una media de 115 caracteres aprox. En cambio, los tweets NO verídicos tienen una concentración muy grande con longitudes de entre 70 y 125 caracteres aprox.; con una media de 100 caracteres.



11.¿Cómo están distribuidos los tweets según la cantidad de caracteres de cada uno?

Por último, realizamos un gráfico swarmplot para visualizar de manera aproximada la cantidad de tweets que hay de cada tipo de noticia y su longitud. El eje Y representa la cantidad de caracteres que tiene cada tweet, mientras que el eje X representa el tipo de noticia. Podemos notar que las noticias reales tienen tienden a tener una mayor longitud que las falsas, visualizandose un pico mayor llegando al 140, mientras que las falsas se encuentran más distribuidas.



4-Conclusión

Como se mencionó anteriormente, el análisis exploratorio ayuda a entender los datos, buscar características e irregularidades. Esto se logró realizando preguntas interesantes y resolviendolas mediante análisis estadísticos y creando visualizaciones. La finalidad de este primer trabajo fue únicamente realizar el análisis exploratorio del dataset para poder continuar con la segunda parte, que básicamente es crear un algoritmo de NLP para analizar cada tweet. Esto significa que tuvimos que enfocarnos en analizar (sin realizar ningún procedimiento de machine learning) la relación entre el keyword y el texto del tweet. Además analizamos la longitud de los tweets como para saber de antemano con qué longitudes vamos a estar trabajando en la segunda etapa del trabajo, ya que no es lo mismo analizar un texto de 300 caracteres, que otro de 140.

Gracias a todas estas visualizaciones tenemos un contexto muy amplio del dataset con el que vamos a trabajar para la segunda etapa.

En el dataset nos encontramos con un error al tratar de realizar un plot, hay una celda en la columna ‘location’ que contiene M!$$!$$!PP!, por estar escrito con '$' el plot de matplotlib no funciona.

5-Bibliografía

- **Apunte de la materia:**

  https://piazza.com/class_profile/get_resource/k7s41iiajq271y/k7s577n4h1g14s
- **https://numpy.org/**
- **https://matplotlib.org/**
- **https://pandas.pydata.org/**
- **https://seaborn.pydata.org/**