Nombre del Grupo : *Equipo Rocket*

integrantes del grupo: **Email :**

-Maria Laura Sousa *marialaurasousa23@gmail.com*
-Guillermo Silva *victorguillesilva07@gmail.com*
-Griselda Maria Veron Bordon
-Matin Coronel

75.06/95.58 Organización de Datos - 1er cuatrimestre 2020

TP 1: análisis exploratorio de datos

Introduccion :

Este analisis se basa en los datos provistos por Twitter en el archivo train.csv, los cuales se encontraban en la pagina https://www.kaggle.com/c/nlp-getting-started bajo el nombre de Real or Not? NLP with Disaster Tweets.
Informacion que nos provee el dataframe

target: nos informa sobre el tweet, si este se trata de un desastre falso (0) o un desastre real (1).
id: es un identificador numerico para cada tweet.
text: nos proporciona el texto de cada tweet.
keyword: es un tipo de clasificador para agrupar los temas de los tweets.
location: nos informa sobre la ubicacion de donde fueron enviados los tweet, esta podria estar o no.

objetivo de este analisis:

El Objetivo de este analisis es obtener una vision general sobre los datos recientemente descriptos

Caracteristicas generales del Dataframe train.csv y preguntas que se realizaron para su analisis:

1_¿Que dimenciones tiene este Dataframe?

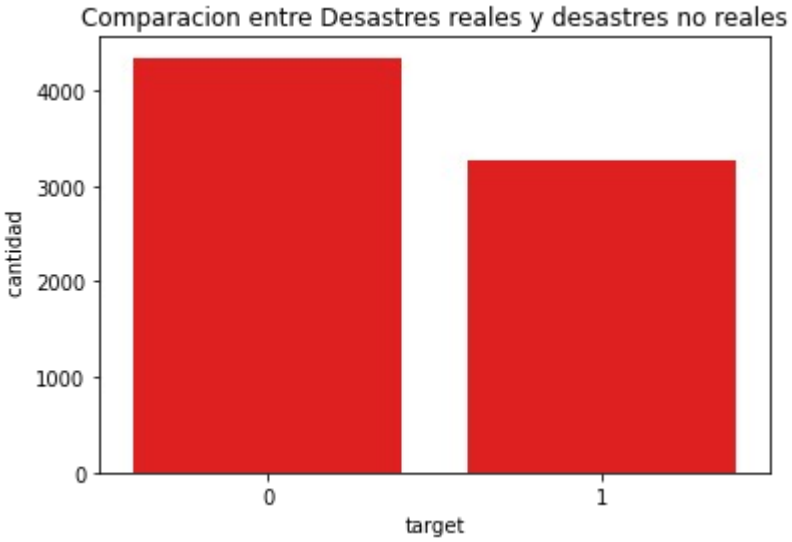
```
Filas 7613, columnas 5
```

```
2_¿hay datos nulos en el dataframe?

id             0
keyword       61
location    2533
text           0
target         0
```

```
 Aqui podemos observar que hay 61 keywords faltantes por lo que habran 61 tweets
que no tendran clasificacion, tambien hay 2533 localizaciones faltantes,
informacion importante para rastrear  de donde fue enviado un tweet
```

```
3_¿Que tipos de datos contienen cada columna?

id           int64
keyword     object
location    object
text        object
target       int64
```

4_¿Que cantidad de tweets anunciando desastres falsos y tweets anunciando desastres reales hay?

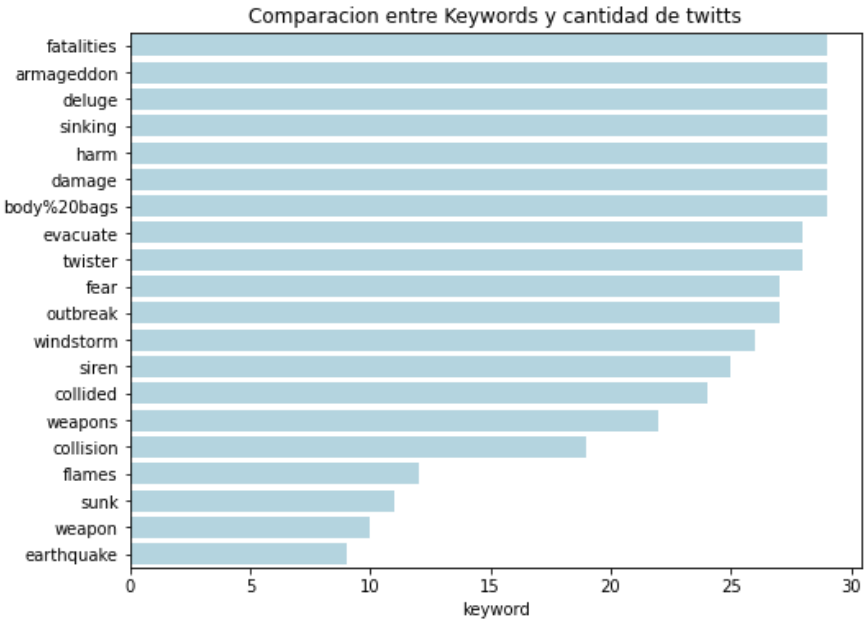


Desastres falsos (0) 4342

desastres reales (1) 3271

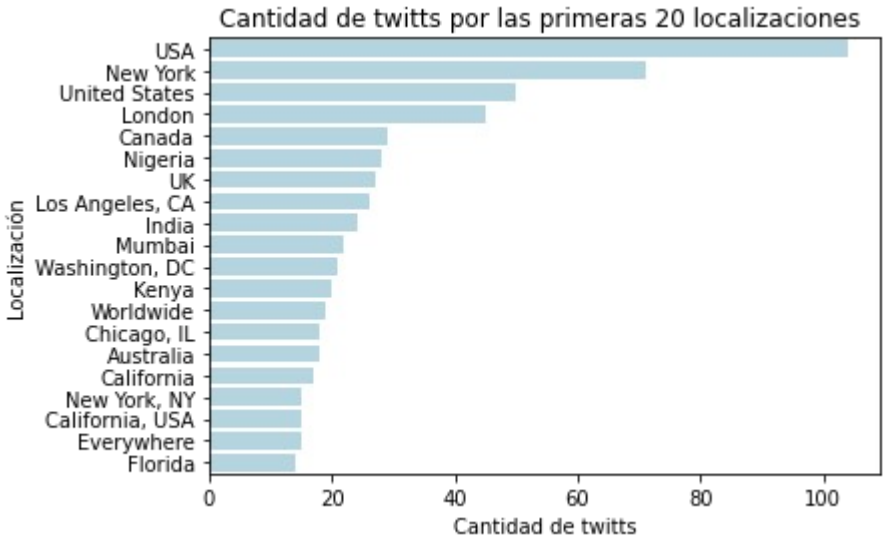
observacion: apesar de que la cantidad de tweets anunciando desastres reales es menor que la de tweets de desastres falsos, es alarmante su numero.

5_¿Cuales son las primeras 20 keywords en la clasificacion de twitts?



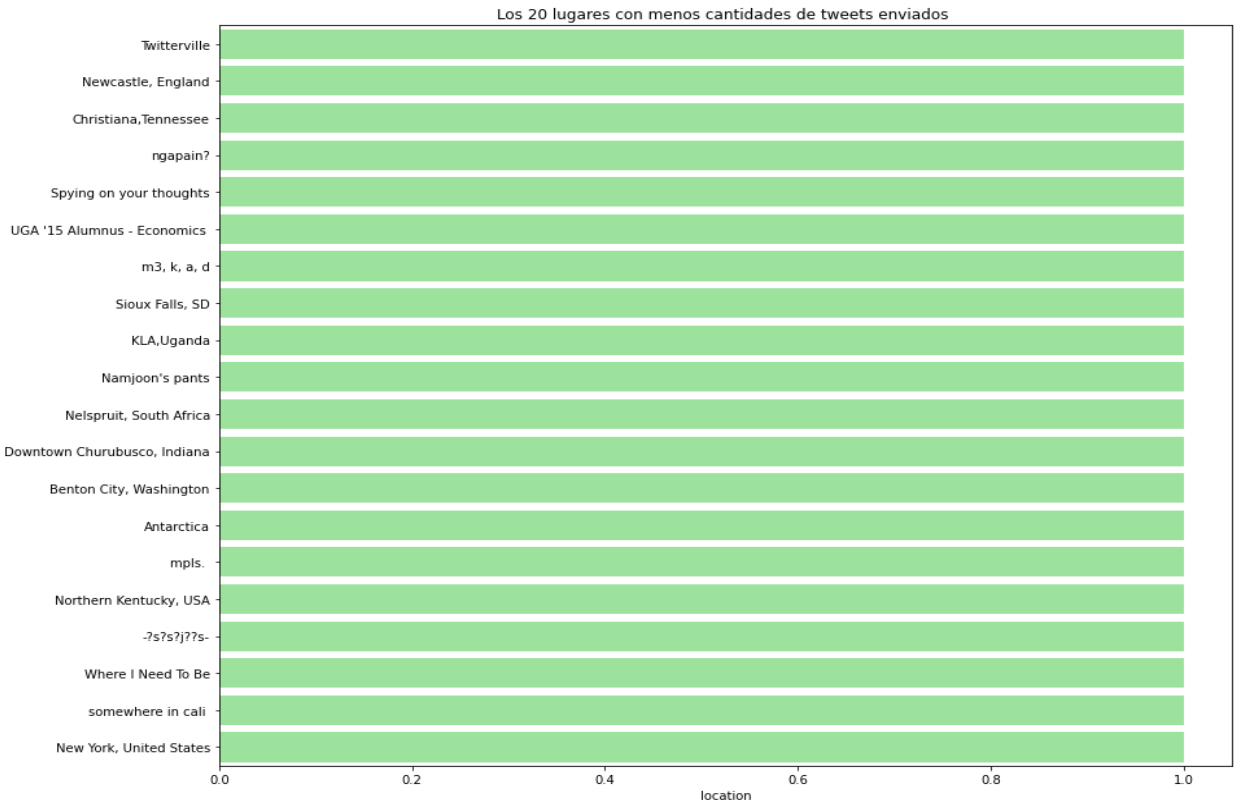
Observacion: podemos ver que la 20 keywords mas utilizadas para clasificar twitts estan relacionadas con palabras que tienen relacion con el daño, como : fatalities, harm, evacuate,colided,weapons, collision.

6_¿Cuales son las primeras 20 localizaciones que envian mayores cantidad de twitts?



Observacion: la informacion provista en esta seccion muestra los nombres de lugares reales, por lo que esta informacion es muy importante si se desea rastrear un twitt anunciando un ataque.

7_¿Cuales son las localisacion con menos cantidades de tweets enviados?



Observacion: En esta seccion observacion que toda la informacion en ‘location’ se refiere a lugares reales, sino a lugares ficticios como : Where I need to be. Por lo que si se desea utilizar la informacion ingresada en ‘location’ para rastrear de donde fue enviado un twitt, no toda la informacion podra ser de utilidad. Al ser lugares ficctios a esto se debe que sea tan poca la cantidad de lugares de donde se envian estos twitts

Conclucion de Analisis:

Esta vision general del Dataframe nos muestra que la red social Twitter es un sitio muy usado por personas para anunciar desastres reales, estos podrian ser anunciados por un espectador o por alguien que va a efectuar un ataque. Tambien podemos ver que las palabras relacionadas a desastres son de uso muy comun en los twitts que se analizaron en esta ocacion.

La informacion provista en ‘Location’ es de mucha utilidad ya que muestra desde donde fue enviado el twitt,pero tambien parte de esta iformacion habla de lugares ficticios. Pero en los casos en los que se pueda determinar si un twitt habla de un desastre real y la ubicacion sea real, podria ser de guia para salvar muchas vidas.