Universidad de Buenos Aires
Facultad de Ingeniería

75.06 Organización de datos

Profesor Titular Argerich, Luis

Trabajo Práctico N^o 1

Alumnos	Padrón
GOMEZ, Marcelo Alejandro	
RODRÍGUEZ MUÑIZ, Maite	94326

1er. Cuatrimestre 2020

22 de mayo de 2020

Índice

1. Objetivos

El objetivo es este trabajo practico es realizar un análisis exploratorio del set de datos. De esta manera nos disponemos a conocer que datos tenemos y cuales no, si alguna de su información necesita ser limpiada, que obtenemos después de eso. También encontrar como se relacionan entre sí los datos.

2. Datos

Vamos a analizar el set de daton "train.cvs"que es el que cuenta con la información de los target. El mismo contiene:

- id - identificador unico para cada tweet
- text - el texto del tweet
- location - ubicación desde donde fue enviado
- keyword - un keyword para el tweet
- target - indica si se trata de un desastre real (1) o no (0)

Vemos que los valores no tienen mucho filtro.

3. Limpieza del data set

Asumiendo que los valores de target están bien puestos y que los valores de textos de los tweets no los queremos modificar, limpiaremos las locaciones y las palabras claves.

3.1. Locaciones

Se encuentran algunas falsas, podemos ver que están escritas a mano sin restricción de caracteres o cantidad de palabras. No son ni de países, ni de provincias ni de ciudades en particular, cada uno ha ido poniendo lo que se le ocurrió como locación, por lo que tenemos un poco de todo y varias que explícitamente son falsas.

Las locaciones que tienen "[#!$ %#/()?*+]"suelen no ser reales, de nuevo hay muy pocas que podrían ser reales pero vamos a decidir considerarlas como falsas.

Algo más que podemos ver son los números, siempre que aparecen son falsas, o tendríamos que rastrear si son coordenadas, pero no es lo más habitual en este set por lo que vamos a descartarlos.

Si bien hay algunas que pueden contener una locación real entre las que tienen la palabra "where"la gran mayoría no aportan información, por lo que podríamos descartarlas de una Locación real.

Como vemos que hay muchas locaciones irreales, donde se escriben frases vamos a filtrar palabras posibles que puedan estar incluidas en frases. Por ejemplo pronombres personales, la palabra World, o live, moon, hell, entre otras.

Para todos los dichos anteriormente creamos una nueva columna "$location - target$" $= False$. Denotando así que los tomamos como falsos.

Por ultimo la idea es dividir entre las comas, ya que es común especificar çiudad, provincia", o "provincia, país". Vemos que entre estos el dato que mas nos va a interesar es el último, ya que es el que tiene el lugar mas conocido, o común para el resto del mundo. Por ende vamos a guardar en una nueva columna "Location2.^{es}tos juntos con los que no tienen coma, ya que tal vez sólo pusieron el país. Dentro de esta columna están sólo los que no registramos como locaciones falsas anteriormente.

Como seguían apareciendo frases, y no podemos filtrar una por una palabras posibles, lo que hicimos fue poner como falsos los que tengan mas de tres palabras.

3.2. Palabras claves

Podemos ver que las palabras claves tienen %20 cuando son dos palabras, elegimos sacarlo y dejar que las dos palabras se junten. Salvo las que tienen de segunda palabra "disaster", ahí elegimos sacar disaster ya que es el tópico en común del que sabemos que estamos hablando.

Por otro lado hay muchas que son las mismas palabras pero conjugadas o en plural. Por lo tanto deberían ser parte de una misma palabra clave. Vemos que estas palabras derivan de la la misma por lo cual empiezan igual y tienen diferente terminación. Vamos unificar eligiendo las palabras claves mas chicas de esas familias , definiendo como familias las que tienen las 5 primeras letras iguales.

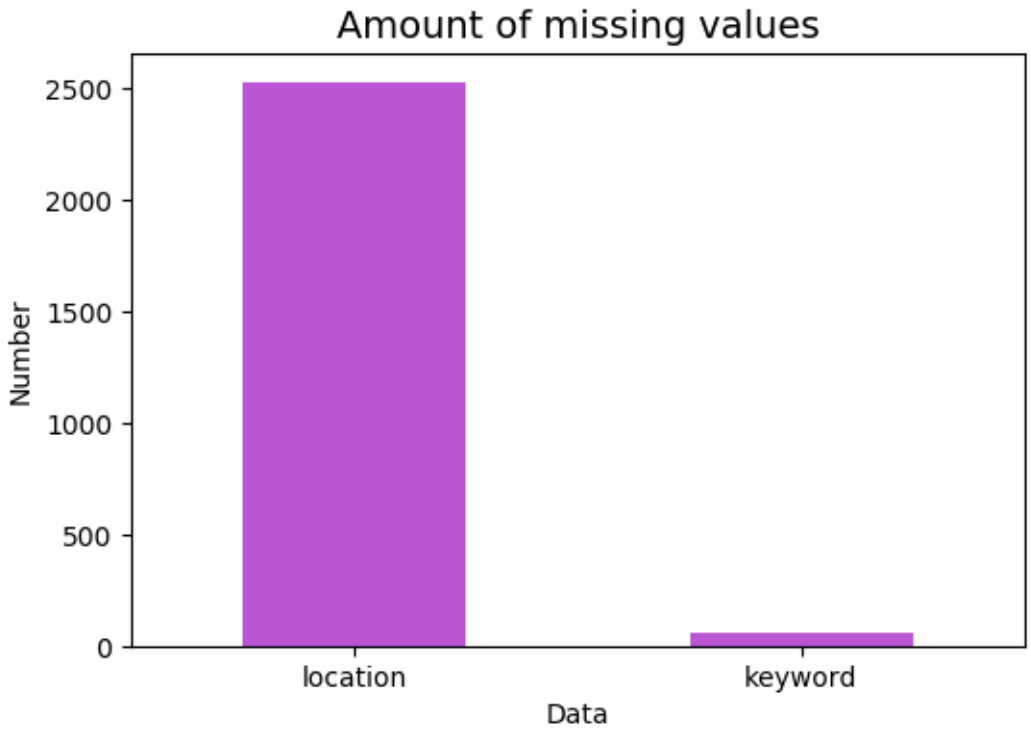
De esta manera achicamos la cantidad de palabras claves únicas de 220 a 150.

4. Preguntas

A partir de acá vamos a ver qué preguntas le podemos hacer al set de datos para conocerlo mejor.

4.1. Que datos nos faltan

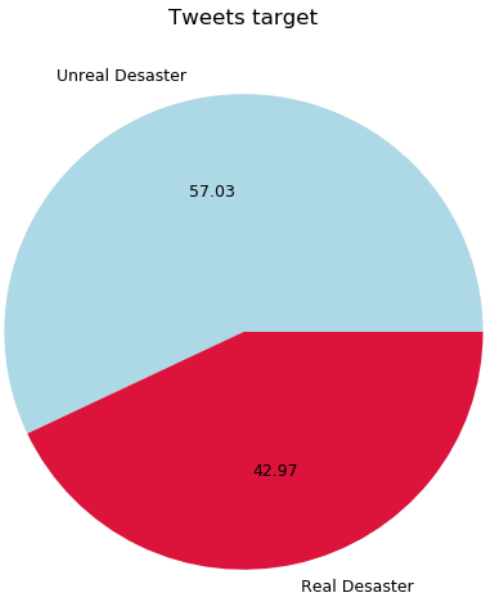
Tenemos datos en todas las filas (7613) de target y de texto, pero no en las demás, por lo que vamos a ver que cantidad de datos nos faltan.



Podemos ver que nos faltan muchísimos datos de locaciones, casi 2500, un 30 % de los datos y muchísimos menos de palabras claves, 61 valores, un 0.8 % de los datos.

4.2. Target

Podemos ver cuantos tweets tenemos como desastres reales y cuales que no hablan sobre desastres reales.



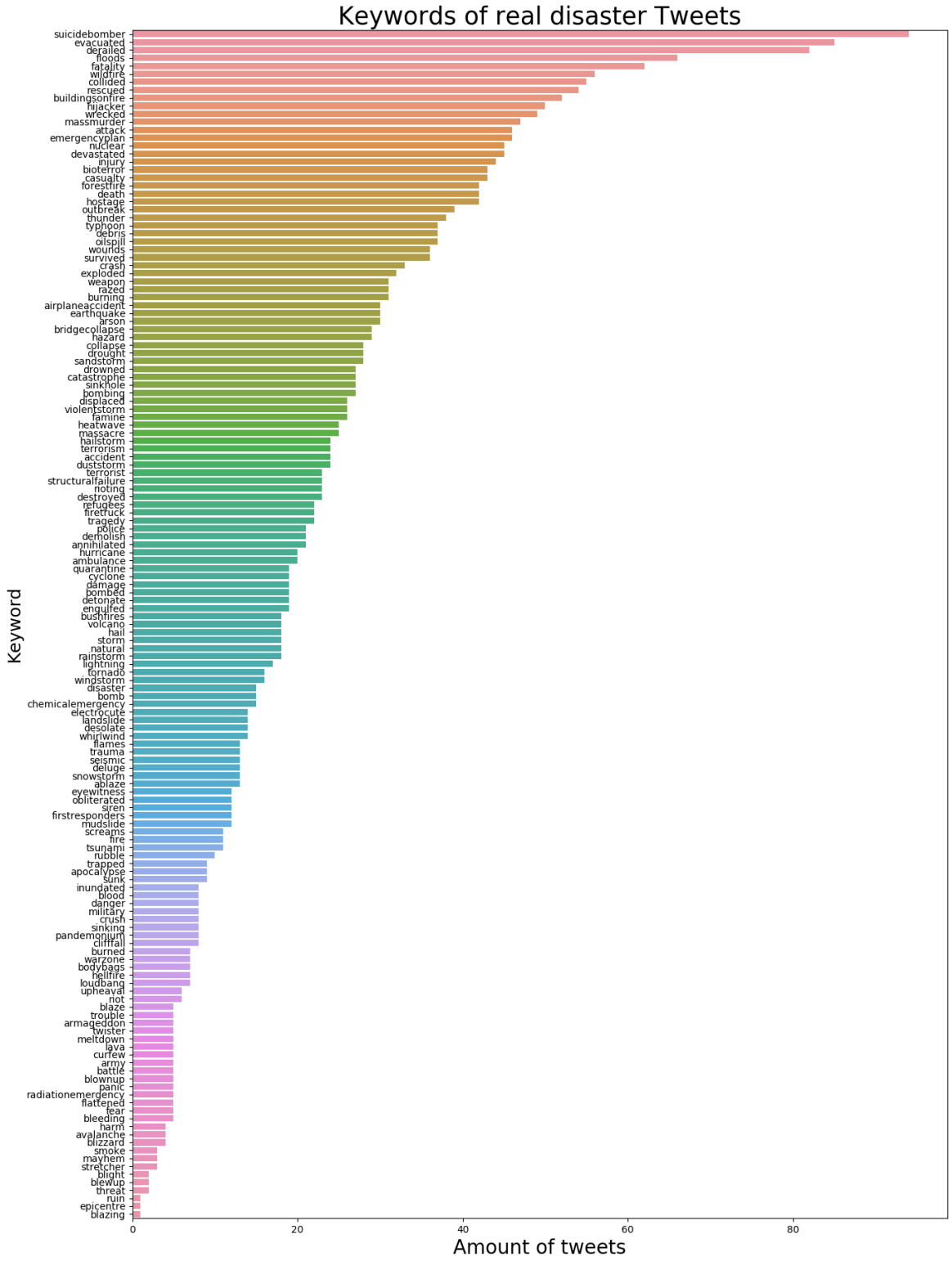
Podemos ver que mas de la mitad de los tweets no habla sobre desastres reales.

4.3. Palabras claves

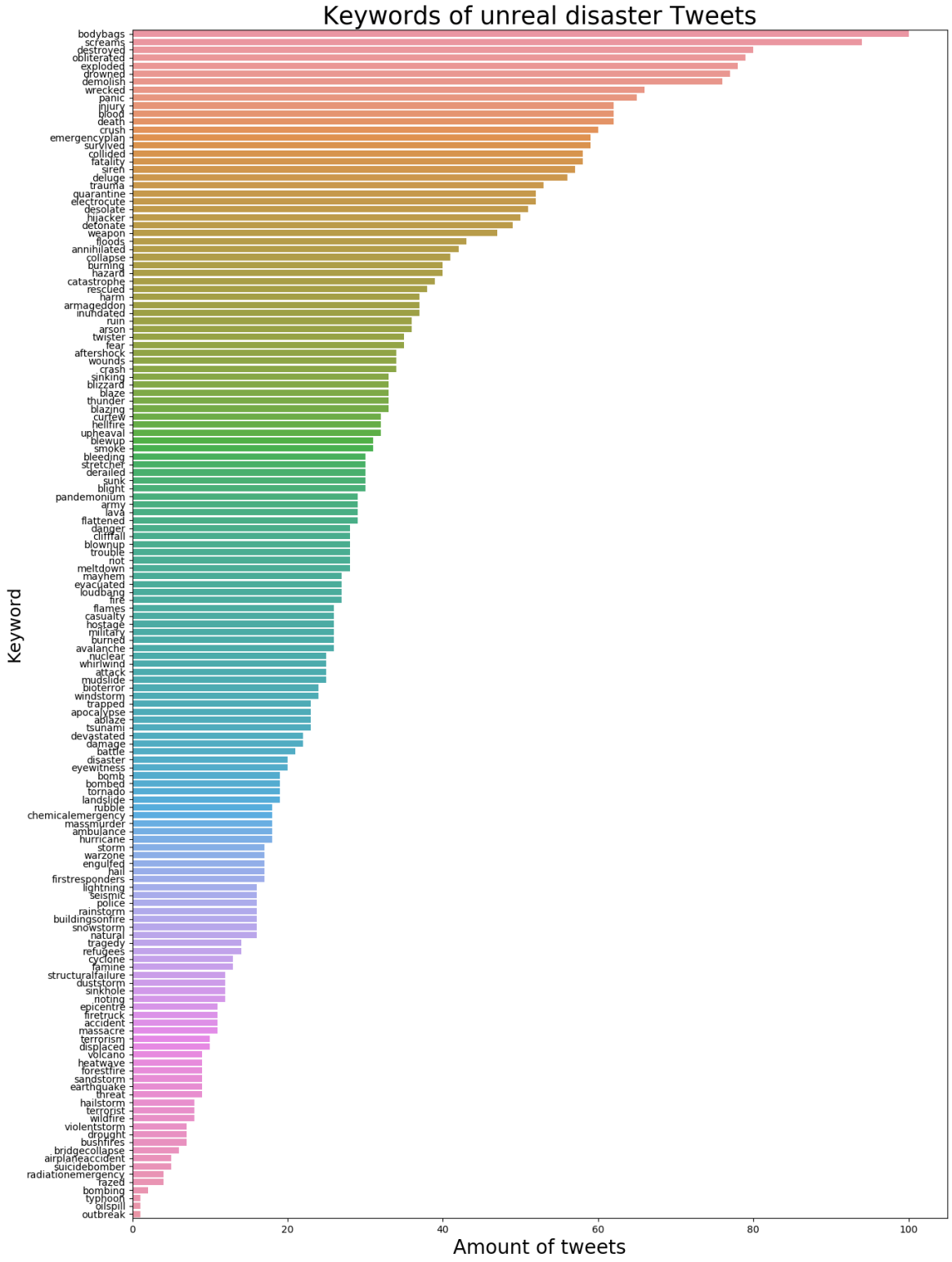
Vamos a analizar ahora las palabras claves y como varian si hablan de desastres reales o no.

Podemos ver cuanto varian, palabras claves que son comunes en desastres reales o no lo son en desastres que no son reales aunque tal vez hablan de temas similares.

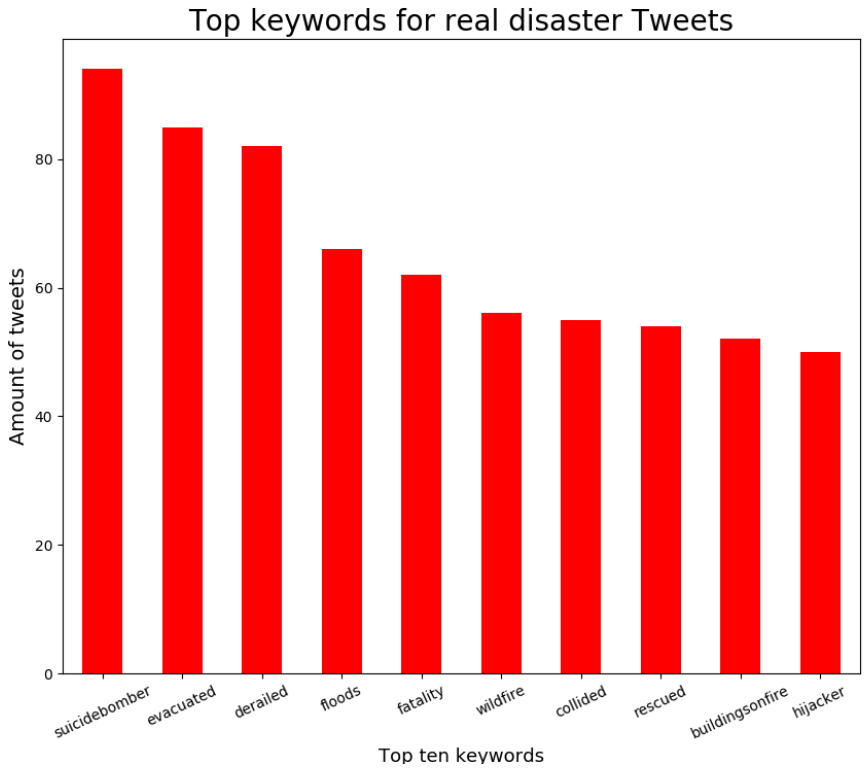
Podemos ver cuanto aparece cada palabra clave en tweets que son sobre desastres reales:

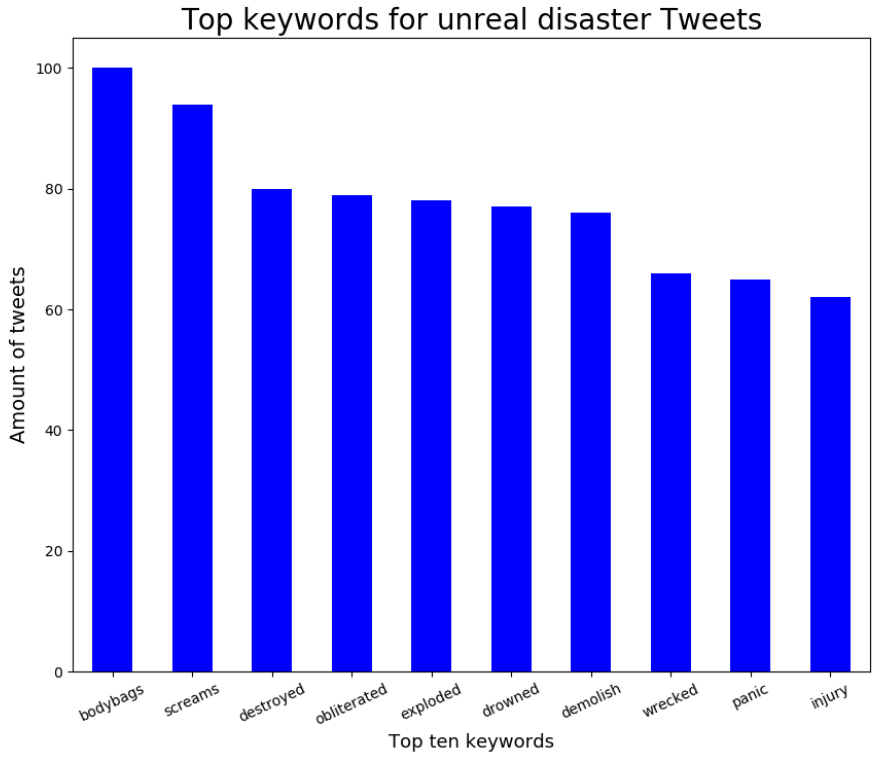


Podemos ver cuanto aparece cada palabra clave en tweets que no son sobre desastres reales:



Veamos ahora las diez palabras claves más utilizadas para tweets de desastres reales y los que no lo son.





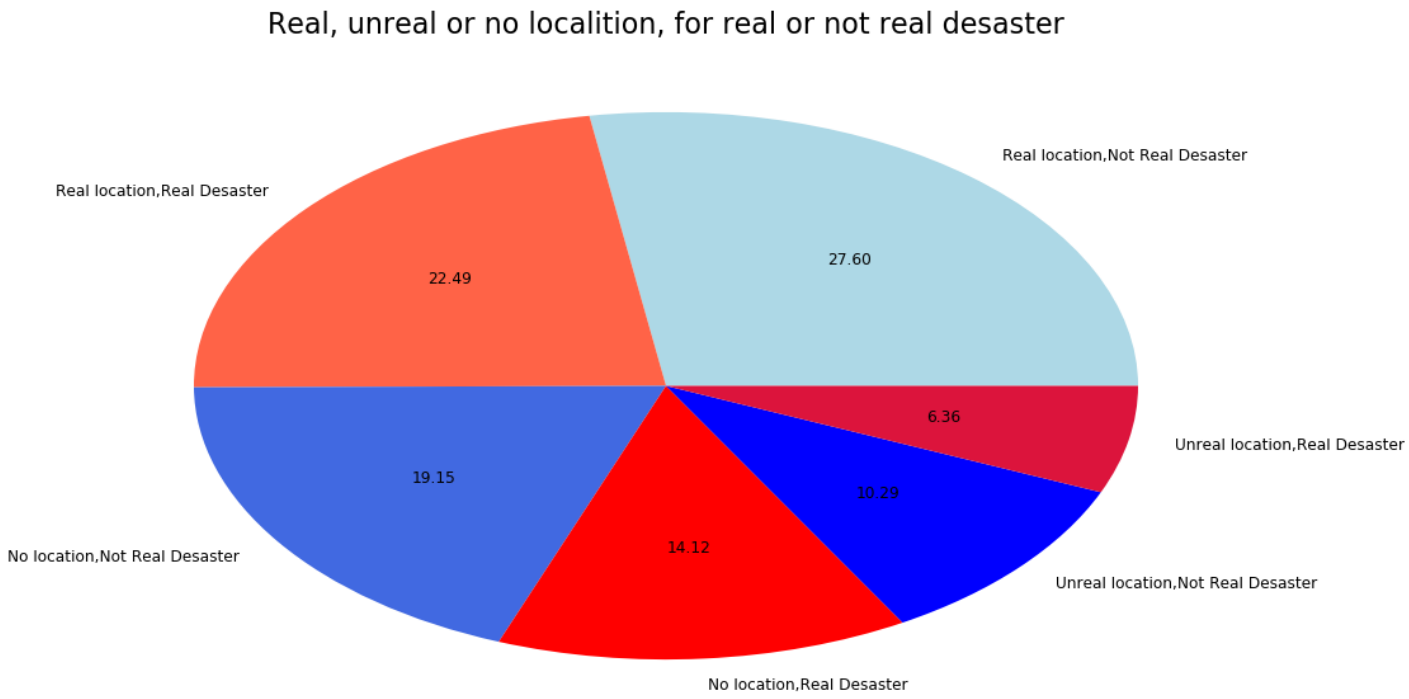
Podemos ver que las palabras claves usadas cambian muchísimo. Por ejemplo la primera palabra clave de los que no son desastres reales es "bodybags"que aparece 7 veces contra 100 en tweets sobre desastres reales y no Lo mismo sucede con screams que figura mas de 90 veces mientras que en

desastres reales no figura mas de 15 veces. Viendo los casos contrarios, los primeros de los casos reales, "suicidebomber.ªparece 94 veces contra 5 en los tweets no reales.

Como se puede ver en estos gráficos y veremos en los siguientes hemos elegido la gama de los rojos para hablar de desastres reales ya que es el color de la alerta, y los colores azulinos los dejamos para los que no son reales, esta misma gama se verá en los siguientes gráficos.

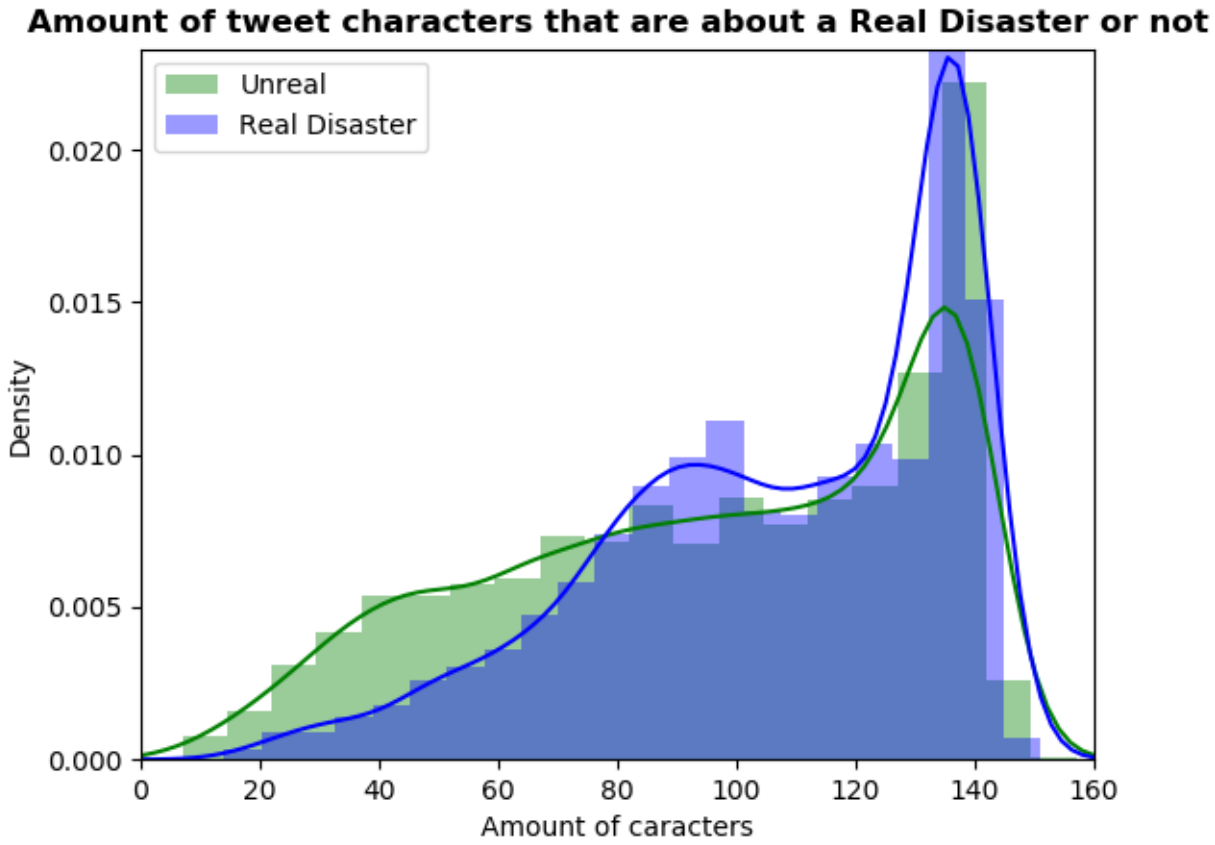
4.4. Locación

Dentro de los datos de locaciones vemos que hay varios que no tienen valor, otros que los filtramos como valores irreales y valores reales, o que no se se han filtrado como irreales. Para todas esas opciones vamos a ver cuales que cantidad son de desastres reales y cuales no.

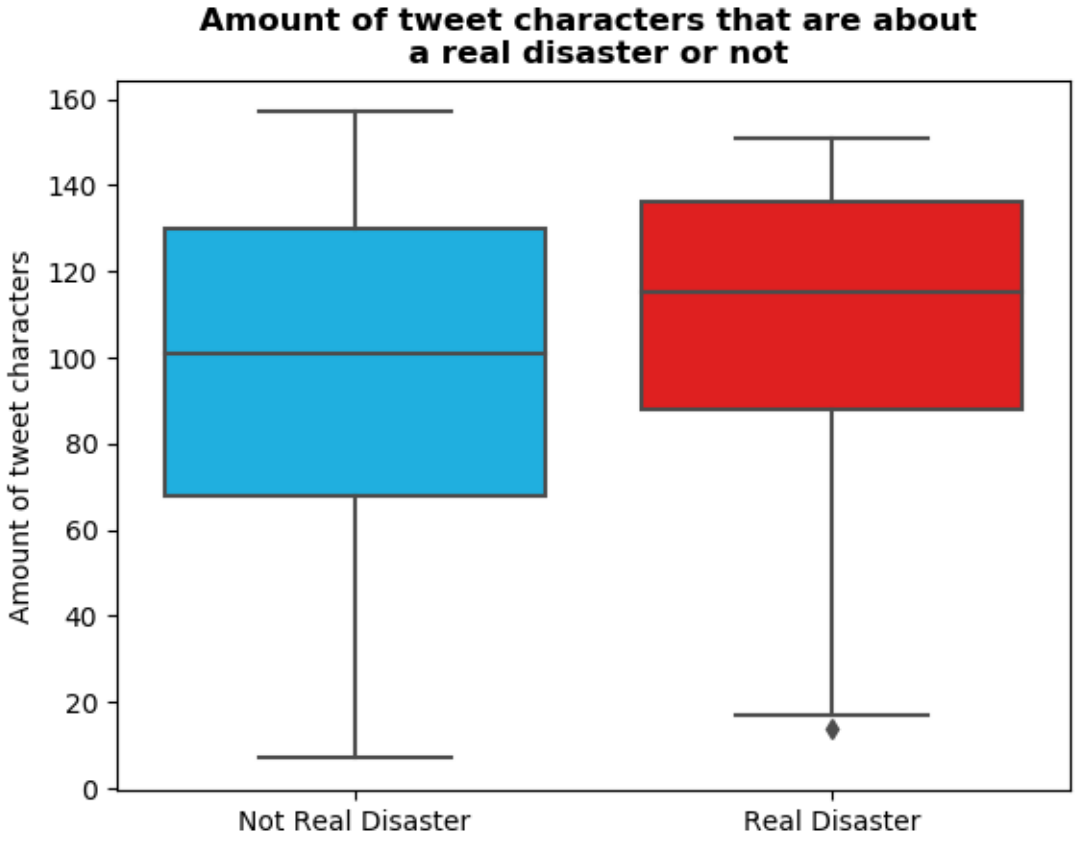


Tenemos que tener en cuenta que no se ha hecho todavía un filtrado con datos de países y estados de Estados Unidos para definir mejor cuales son locaciones falsas.

4.5. Largo del tweet vs si habla sobre un desastre real o no

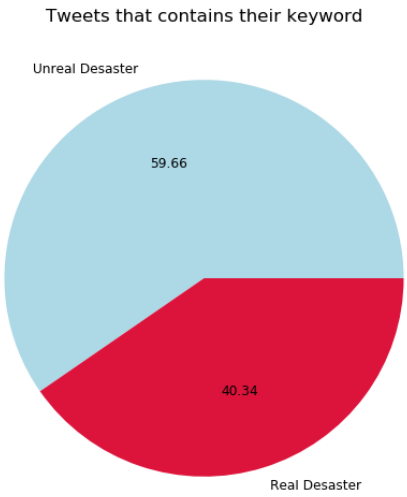


Podemos ver que la longitud promedio de los tweets que no son sobre desastres reales es mas baja (100) que los que si lo son (115). También que los primeros tienen longitud mínima menor, de 7, mientras que los segundos de 20 caracteres. La longitud máxima es mayor en los que no hablan sobre desastres reales, y la densidad esta mas concentrada alrededor de la media en los que sí hablan sobre desastres reales.



4.6. Tweets que tengan su palabra clave

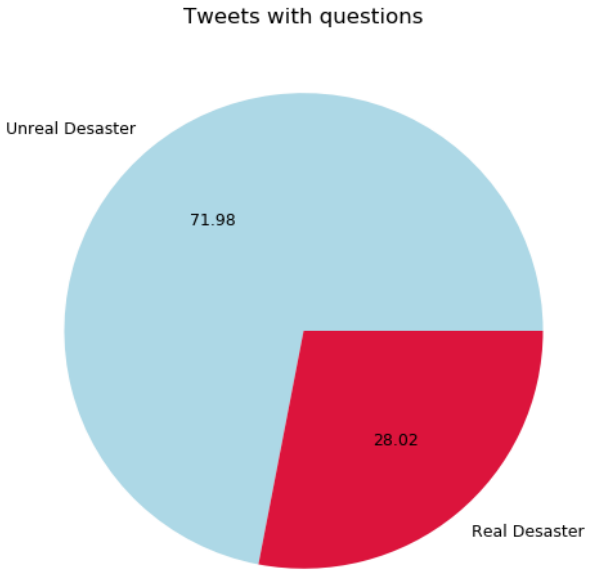
Vamos a ver ahora si los tweets que tienen su palabra clave dentro del texto son sobre desastres reales o no.



Podemos ver que tiene aproximadamente el mismo porcentaje que presentan todos los tweets en general, aproximadamente 60 % no son sobre desastres reales.

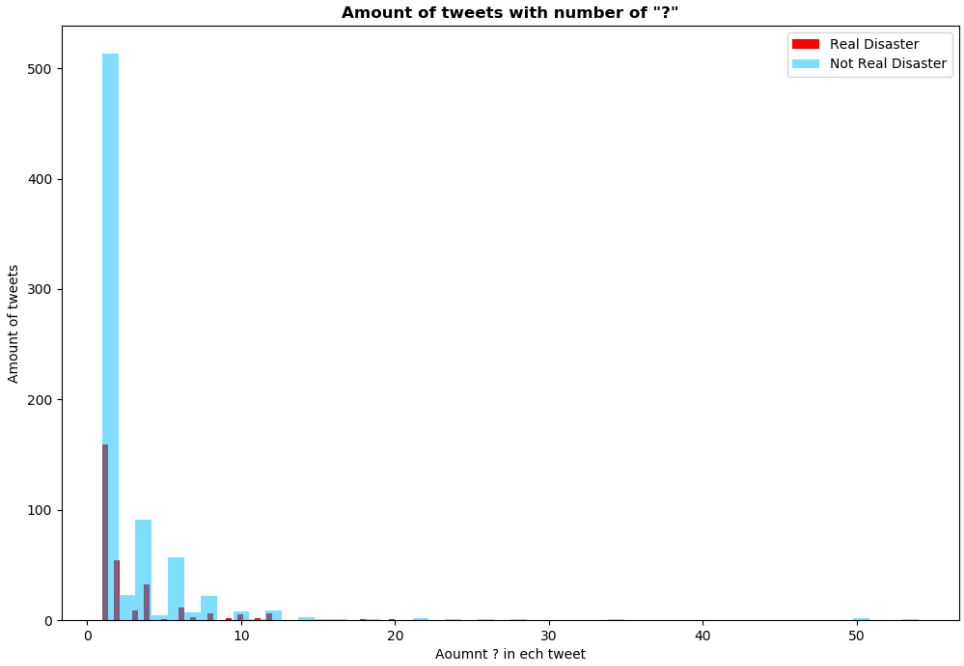
4.7. Tweets con preguntas

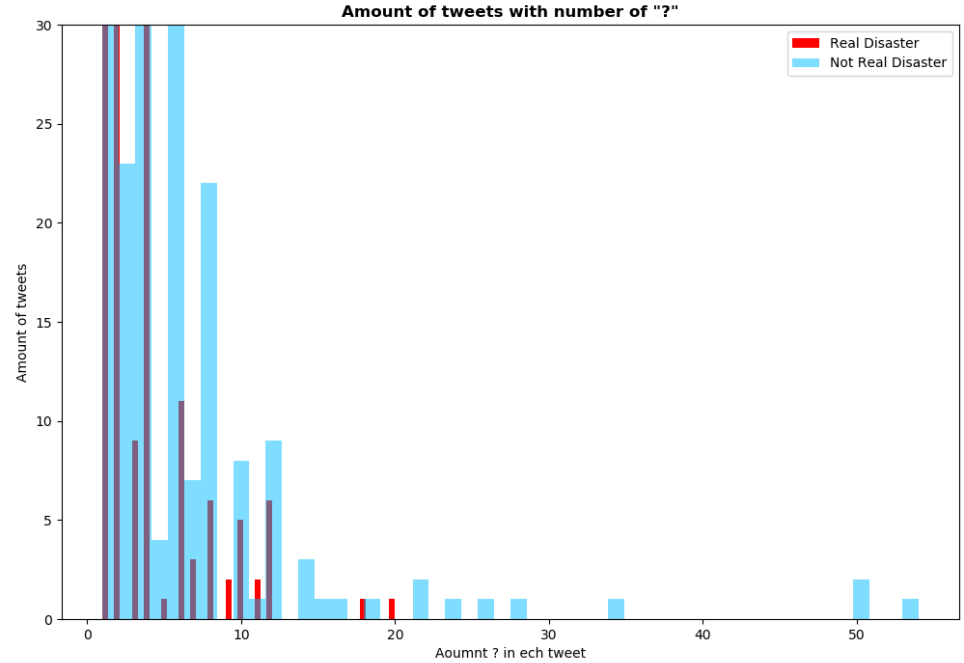
De los tweets en los que hay signos de preguntas, veremos cuantos son sobre desastres reales o no.



4.8. Cantidad de preguntas de Tweets

Veremos ahora la cantidad de signos de pregunta ” ? ”que aparecen en cada tweets.

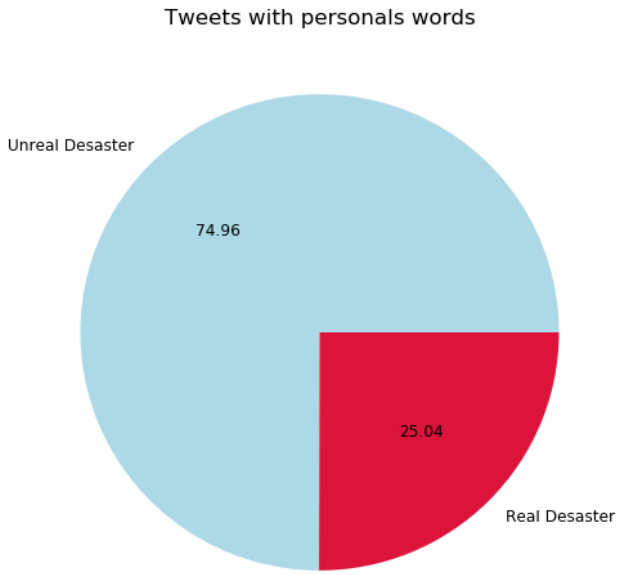




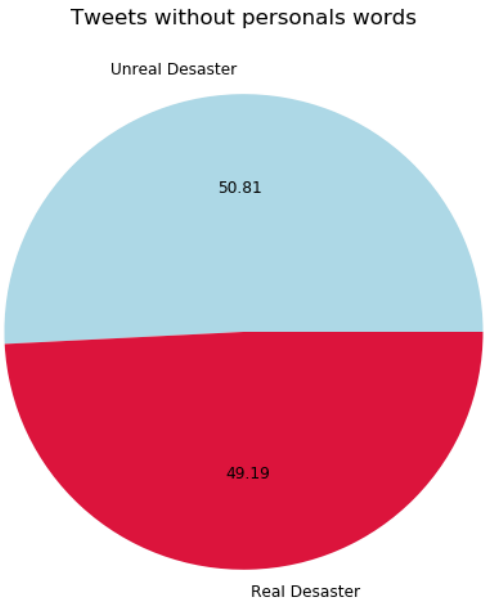
Podemos ver que la mayoría de tweets con preguntas son los que no son sobre desastres reales y los que más preguntas tienen también.

4.9. Tweets con palabras personales

Nos pareció interesante preguntarnos cuando hablan desde la primera persona, desde sus vivencias u opiniones. Por lo que vamos a preguntarnos que tweets tienen las palbras (I , My, Our, me, we, mine) y la veracidad de los desastres de los cuales hablan.

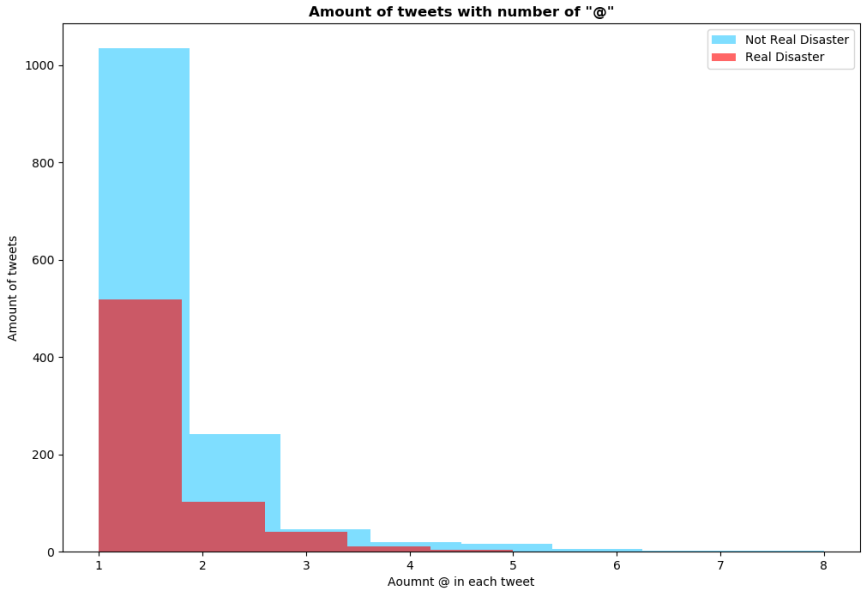


Vemos que la mayoría no hablan sobre desastres reales.
Por el contrario, lo que no tienen estas palabras estan divididos.



4.10. Tweets con citas

Vemos que para citar a otra persona en esta red se utiliza "@"por lo que vamos a ver que tanto se cita por tweet y como varia eso con la veracidad el desastre del que habla.



Podemos ver que la mayor cantidad de citas se hacen tweeteando sobre desastres no reales.

5. Conclusiones

En primer lugar, tal como anticipamos al comienzo, nuestro análisis estuvo centrado en las columnas locations, text y keywords, y sus relaciones con target, asumiendo que el carácter real o no de los textos viene ya dado por el valor de esta última columna.

En el proceso de limpieza del dataset, advertimos una significativa falta de locaciones, ello nos llevó a utilizar este elemento como relevante al momento del análisis.

La primera conclusión es la significativa disparidad entre las palabras claves más relevantes entre los desastres reales y los no reales, conforme los dos primeros cuadros del punto 3.3.

Dicho extremo se puede ver con mayor claridad a partir del desagregado del top ten de los mismos en los dos cuadros siguientes, en los que puede advertir que no existe ninguna coincidencia entre ambos. Esto da una idea de las palabras que son más utilizadas en los casos de tweets que muestran desastre no reales; palabras como badybags, screams, destroyed, obliterated, exploded, drowned, demolish, wrecked, panic y injury estarían, entonces, relacionadas en general con este tipo de tweets.

En el caso de las locaciones discriminamos entre locaciones reales, no reales y donde no tiene, buscando las relaciones que podría haber entre ellas y el tipo de desastre, concluyendo que hay una mayor cantidad de tweets en las locaciones reales (alrededor del 50 %) mientras que las restantes se desagregan entre el restante 50 % con predominancia de los casos de faltas de location, pero ello está relacionado con la cantidad de casos con esta columna vacía que mencionáramos más arriba y a la falta de una mayor limpieza, contando con mas datos.

Con relación a la cantidad de caracteres por cada tweets, según punto 3.5, podemos concluir que en el caso de los tweets que muestran desastres no reales hay una distribución más uniforme en la longitud de los mensajes, a contrario de los que aluden a desastres verdaderos, donde hay una predominancia de los que tienen una mayor extensión.

También nos preguntamos si habría alguna relación entre las palabras clave y si se trataban de desastres reales. Así logramos advertir cierta tendencia en el casos de tweets que aluden a desastres que no son reales. No obstante ello, se trata de una diferencia de solo un 10 % sobre la media (59.66 % vs. 40.34 %), con lo cual concluimos que no se trata de una variable por si sola significativa y es aproximadamente la misma diferencia que hay en el set de datos en general.

Para finalizar, comparamos la cantidad de preguntas, a partir de la cantidad de signos de interrogación en el texto de los tweets, para inquirir sobre su relación para con los desastres reales o no, y logramos concluir el uso abusivo que se hace del signo '?' en los casos de desastres no reales, pudiendo inferir que podría tratase de una de sus características definitorias junto con la longitud y locations no reales. También vimos que mas tweets tienen pocas preguntas puando no son sobre desastres reales.

Las palabras que aluden a opiniones o vivencias personales se encuentran mucho mas en tweets que no hablan de desastres reales. Y lo mismo sucede con las citas que son mucho mas frecuentes en estos.

6. Código

https://github.com/maiterm/OrgaDeDatosOyentesEnCuarentena