UNIVERSIDAD DE BUENOS AIRES
FACULTAD DE INGENIERÍA

75.06 Organización de Datos

Trabajo Práctico 1
Primer Cuatrimestre de 2020

Grupo oyentes: GMC

Gastón Nuñez
Martín Ríos
Carla Squillace

Link de GitHub: https://github.com/carlachka/7605_TP1

Introducción

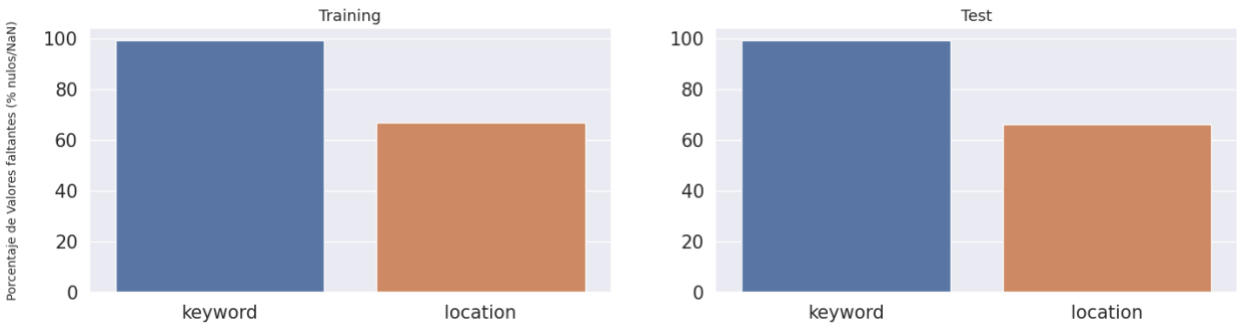
Este informe se encarga de analizar los tweets del set de datos de la competencia: https://www.kaggle.com/c/nlp-getting-started. Nos centramos mayoritariamente en el archivo train.csv, y lo relacionamos con test.csv. En ambos archivos se encuentran las siguientes variables:

- id - identificador único 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)

Procedemos a explicar los análisis de datos más interesantes, relacionado o no, a predecir si un cierto tweet es real o no.

Valores faltantes

Figura 1. Porcentaje de valores no faltantes o no nulos (%)



id	100.00
keyword	99.20
location	66.73
text	100.00
target	100.00

id	100.00
keyword	99.20
location	66.14
text	100.00

Tanto el data frame train como test tienen únicamente valores faltantes para las variables keyword (0,8%) y location (33,27% y 33,86%). Las variables id, text y target no tienen valores faltantes.

Tanto el data frame train como test tienen la misma proporción de missing values para las variables keyword y location.

En principio, a nivel general no estaríamos haciendo de tratamiento de valores faltantes. Sí indagaremos al respecto más adelante cuando analizamos cada variable por separado.

Duplicados

Al momento de analizar duplicados en la base de datos, optamos por estudiar dos metodologías diferentes.

En principio, comparamos cuantos registros se perdían considerando las variables keyword, text y location. En este caso se estarían eliminando 61 registros. También analizamos cuantos registros se perderían si consideramos la variable target.

Al incluir la variable target, vemos que eliminaríamos 52 registros duplicados. Esto quiere decir que hay tweets que son completamente iguales - no solo incluyendo su texto, también location y su keyword -, pero que hay veces en las que están marcados como desastre (target = 1) y otras veces en las que están marcados como no desastre (target = 0).
Consideramos que esta es una inconsistencia en la base de datos, que afectaría el objetivo de predecir si un cierto tweet es real o no en el futuro. Consecuentemente decidimos quedarnos con los registros únicos considerando las variables text keyword y location.
De esta forma, cuando vuelve a aparecer un registro que tiene el mismo text keyword y location, se eliminaría. De acuerdo con esta metodología, eliminaríamos 52 registros repetidos.

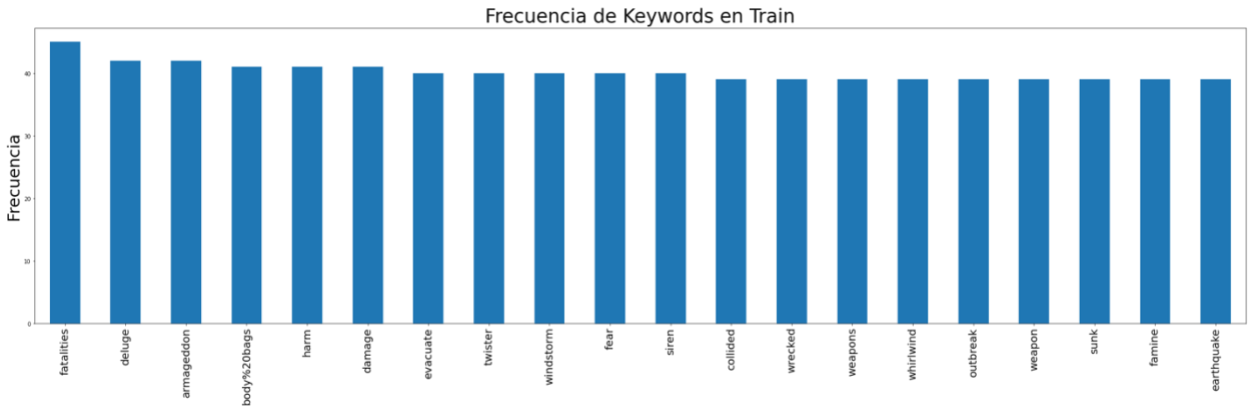
En segunda instancia, propusimos una nueva metodología no automática para la eliminación de repetidos que pone particular énfasis en la variable target. Teniendo en cuenta primeramente el texto de cada tweet repetido, los agrupamos y calculamos la media de target y cantidad de coincidencias por grupo. Si en el grupo hubiera discrepancia de target, decidimos quedarnos con un solo tweet por grupo cuyo target corresponda con la mayoría de target del grupo. De acuerdo a esta metodología se eliminarían 110 registros.

Decidimos seleccionar la segunda metodología, y eliminar 110 registros.

Palabra clave o “keyword”

Figura 2. Palabras claves con mayor ocurrencia

Análisis Exploratorio de Datos (EDA)

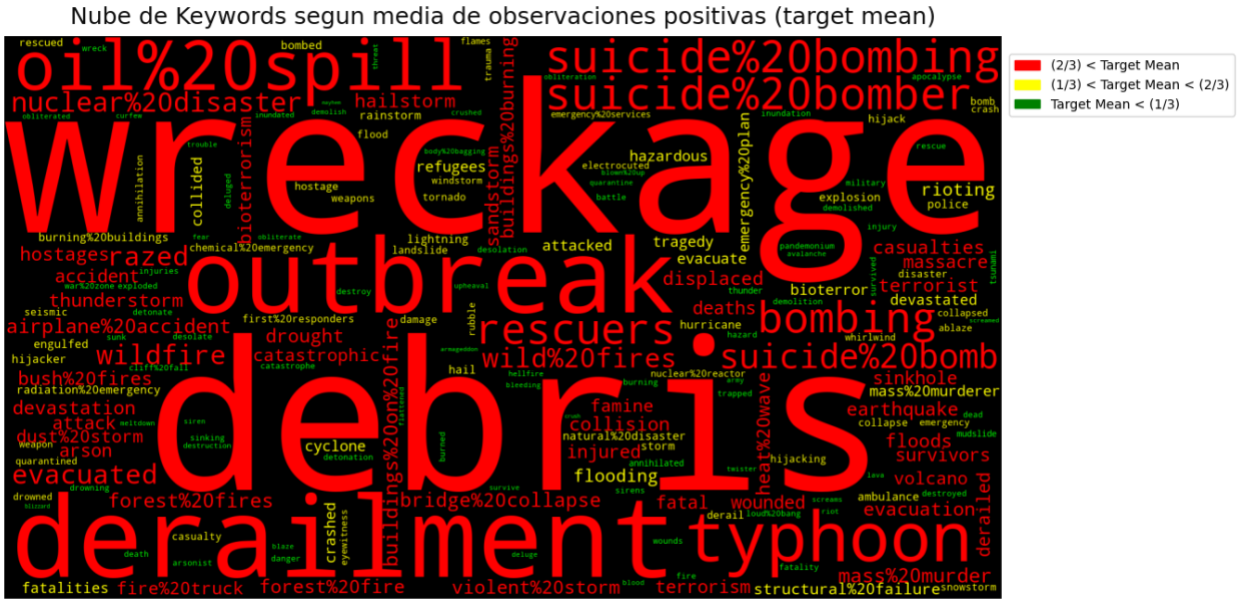


En el grafico anterior se pueden observar las 20 palabras claves más repetidas en el data frame train. Son un total de 221 palabras claves únicas.

Al analizar el dafa frame test, vemos que también posee 221 palabras claves únicas. Esto puede ser útil al momento de predecir tweets reales.

Otra forma de visualizar las palabras claves es la utilización de una nube de palabras.

Figura 3. Nube de palabras claves

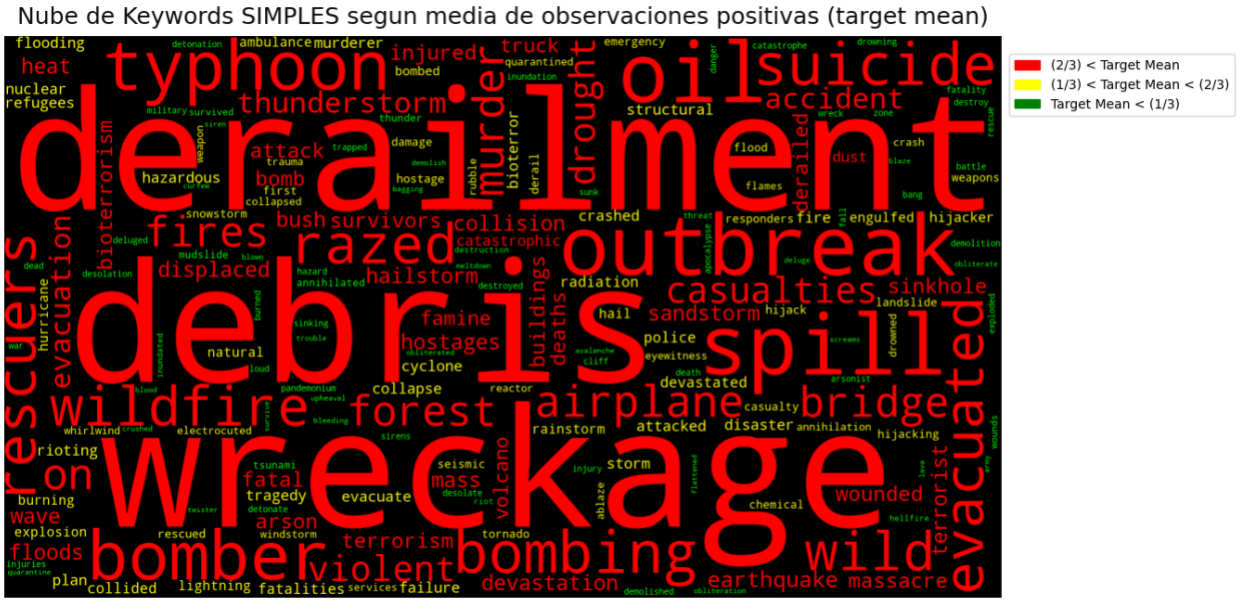


En la nube anterior se puede observar que las palabras claves con mayor probabilidad de ser tweet reales (target=1) de desastre no solo aparecen con mayor tamaño sino que también aparecen en color rojo. Esto se debe a que desarrollamos un sistema de semáforo.
Las palabras clave cuya probabilidad de ser tweets reales de desastres es baja (menor a 33%) aparecen en verde. Las palabras clave cuya probabilidad de ser tweets reales de desastres es media (probabilidad entre 33% y 66%) aparecen en amarillo. Finalmente, Las palabras clave cuya probabilidad de ser tweets reales de desastres es alta (mayor a 66%) aparecen en rojo.

Observamos que las palabras clave con mayor probabilidad de ser tweets reales de desastre son derailment, debris, wrekage, outbreak y los compuestos “oil%20spill” y “suicide%20bombing”.

Con el objeto de tener mayor claridad sobre las palabras claves compuestas, procedimos a separar estos compuestos en palabras claves individuales. Obtuvimos una nube de palabras clave simplificada.

Figura 4. Nube de palabras claves simplificadas

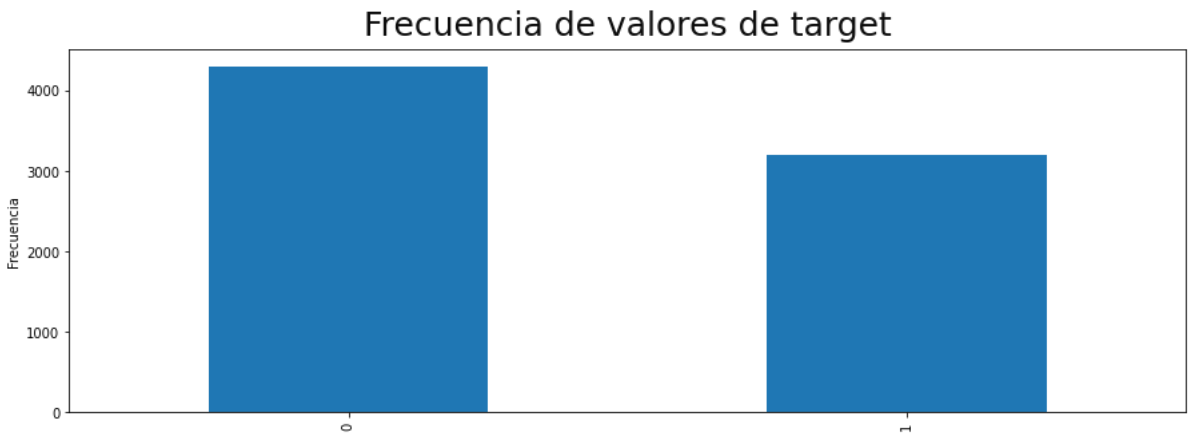


Observamos que a diferencia de la nube anterior, aparecen nuevas palabras separadas como spill, bomber, bombing.

Variable objetivo o “target”

La variable objetivo nos permite identificar si un tweet real de desastre (target = 1) o falso (target=0).

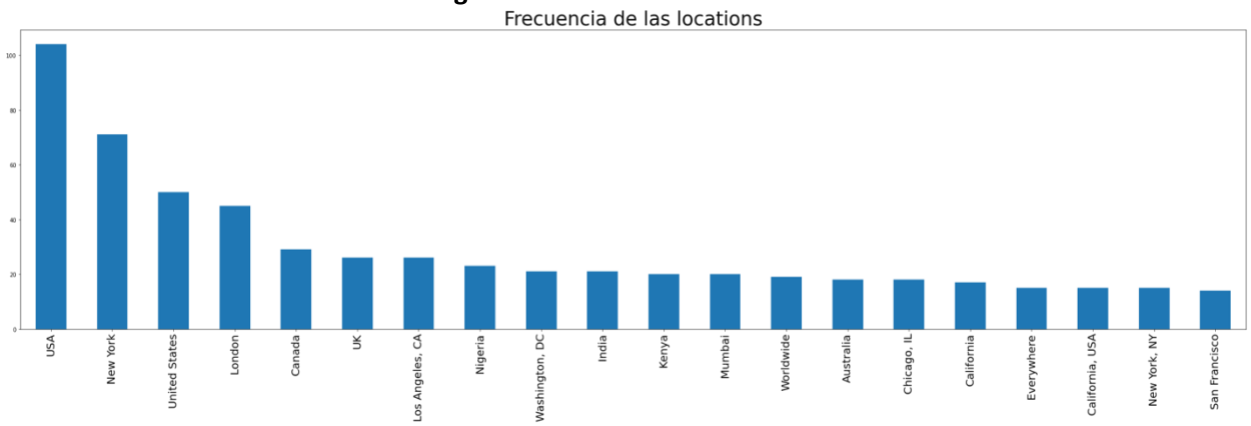
Figura 5. Frecuencia de target reales y falsos



Hay un mayor porcentaje de tweets falsos (57,4%) que reales (42,6%). Esperamos que la cantidad total de observaciones sea suficiente para poder hacer una futura predicción.

Ubicación o “location”

Figura 6. Frecuencia localidades



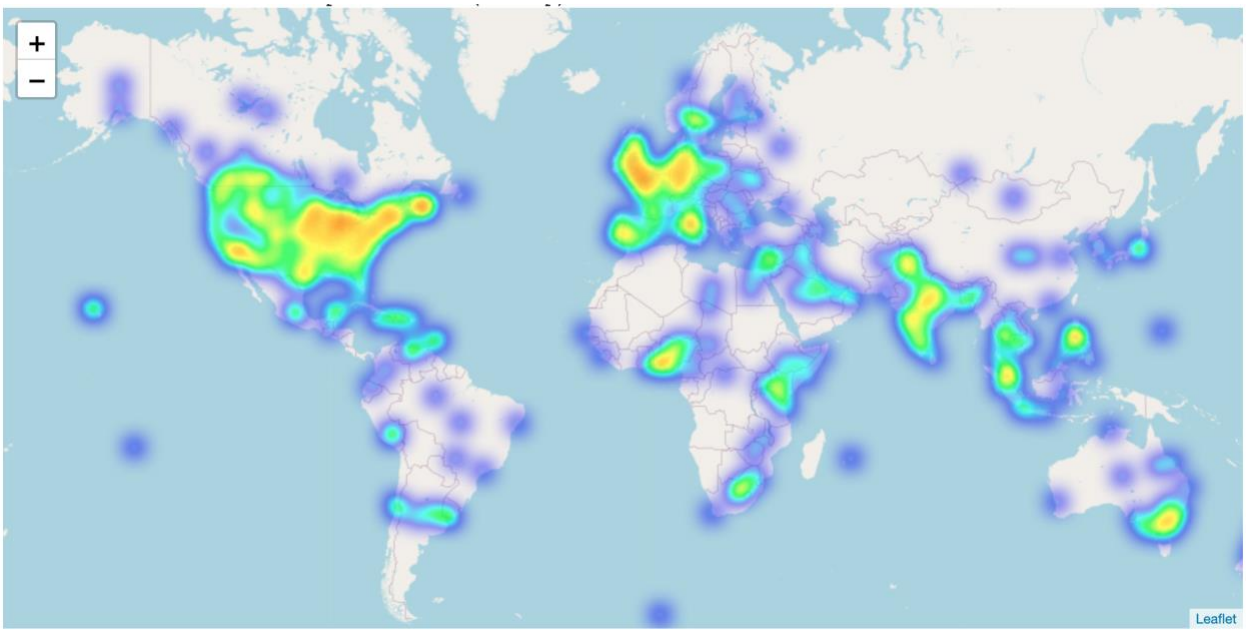
En un primer análisis se registraron tweets desde 3341 lugares. No es correcto hablar de 3341 lugares distintos ya que dentro del listado de locaciones hay paises y estados repetidos, y mezclas de ambos. Por ejemplo, podemos ver a Estados Unidos escrito de dos maneras distintas y considerado como dos lugares distintos. Asimismo, podemos ver que la variable locations no esta automaticamente generada, sino que es un input que ingresa el usuario. Por este motivo, es que hay tantos "lugares distintos".

Hicimos un tratamiento de las localidades para poder realizar “heatmaps”.Para intentar visualizar sobre un mapa la densidad de ubicaciones de los tweets, primero fue requerido convertir esos datos en una notación de coordenadas (latitud y longitud)
Esto fue resuelto usando un geocoder (convertidor de direcciones en coordenadas), separando coordenadas para Desastres y No Desastres. Luego Visualizamos en un mapa interactivo la densidad de la distribución geográfica de los tweets para ambos casos.

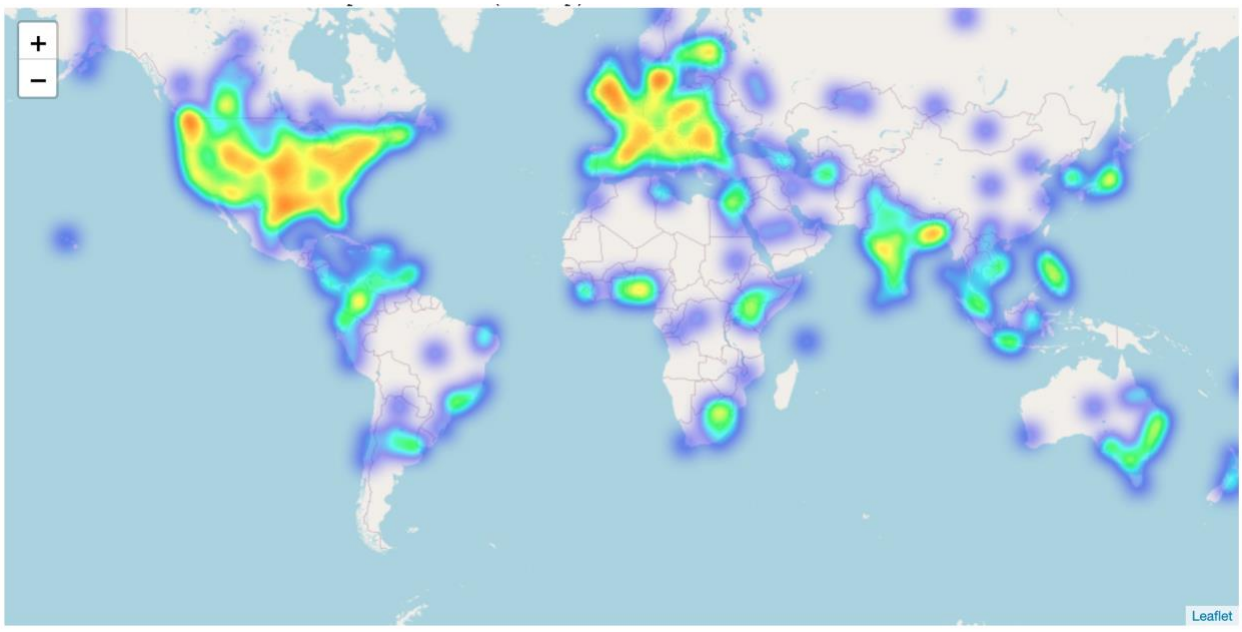
Análisis Exploratorio de Datos (EDA)

Figura 7. Heatmaps por ubicación

Densidad de tweets "Desastres" por ubicación (heatmap)



Densidad de tweets NO desastres por ubicación (heatmap)



Observamos que las diferencias son menores, concentrándose en ambos casos la mayor densidad de tweets en Estados Unidos y Europa.

Texto del tweet

Procederemos con un análisis más detallado del texto de los tweets. Analizamos las frecuencias de las palabras en en general, y como se distribuyen en los dos tipos de tweets.
Como primer paso eliminamos aquellas palabras que se repiten mucho porque se utilizan cómo conectores, se las denomina STOPWORDS. Tambíen eliminamos los símbolos y algunas construcciones que no aportan nigún significado y que son frecuentes en la escritura informal.
En el siguiente listado se pueden observar las palabras mas repetidas con su respectiva cantidad de operaciones.

```
Las palabras más comunes son:
[('via', 211),
 ('get', 184),
 ('fire', 162),
 ('people', 162),
 ('one', 153),
 ('would', 128),
 ("'re", 126),
 ('got', 111),
 ('new', 103),
 ('know', 103),
 ('California', 102),
 ('video', 96),
 ('back', 94),
 ('buildings', 94),
 ('disaster', 92),
 ('going', 92),
 ('News', 92),
 ('burning', 91),
 ('killed', 90),
 ('still', 89)]
```

El análisis de frecuencias, además de saciar la curiosidad, es útil para intentar tener una mirada rápida de lo que trata el total del texto. En este caso, al ser tweets, las temáticas son muy variadas y es difícil encontrarlas y distinguirlas. Consecuentemente para tener mas claridad, la clasificación entre un tweet de un evento trágico o no resulta un buen punto de partida.
En este contexto repetiremos el procedimiento anterior, esta vez diferenciando las que más se repiten en cada uno de los dos grupos. El objetivo es ver si las palabras más frecuentes nos aportan alguna información, alguna estructura o algún patrón propio del grupo al que pertenecen.

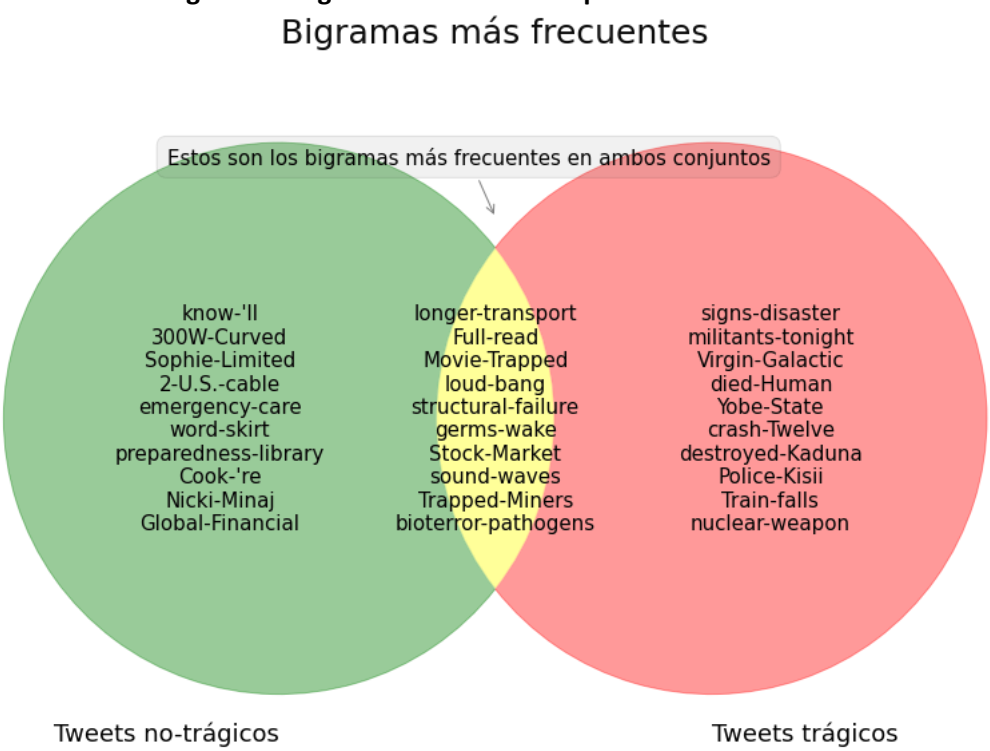
```
Las más comunes para los tweets trágicos:
[('fire', 120),
 ('via', 118),
 ('California', 95),
 ('people', 90),
 ('killed', 87),
 ('News', 74),
 ('Hiroshima', 74),
 ('suicide', 72),
 ('disaster', 68),
 ('fires', 67),
 ('crash', 66),
 ('buildings', 65),
 ('MH370', 64),
 ('Northern', 60),
 ('bomb', 59),
 ('police', 58),
 ('train', 58),
 ('bombing', 58),
 ('Legionnaires', 58),
 ('one', 55)]
```

```
Las más comunes para los tweets no-trágicos:
[('get', 143),
 ('one', 98),
 ('would', 93),
 ('via', 93),
 ("'re", 86),
 ('new', 84),
 ('got', 82),
 ('know', 76),
 ('people', 73),
 ('Full', 69),
 ('see', 68),
 ('going', 68),
 ('back', 68),
 ('gt', 67),
 ('body', 67),
 ('video', 64),
 ('time', 63),
 ('still', 62),
 ('YouTube', 62),
 ('think', 61)]
```

No se llega a identificar una temática en particular, pero si se observa como los tweets durante un evento trágico son, de una forma, más precisos. Por más precisos entendemos a que encontramos nombres propios, sustantivos y verbos específicos. Por el otro lado, los tweets en momentos nos trágicos, parecieran contener palabras más coloquiales, propias de la actividad de “tweetear” contenidos diversos.

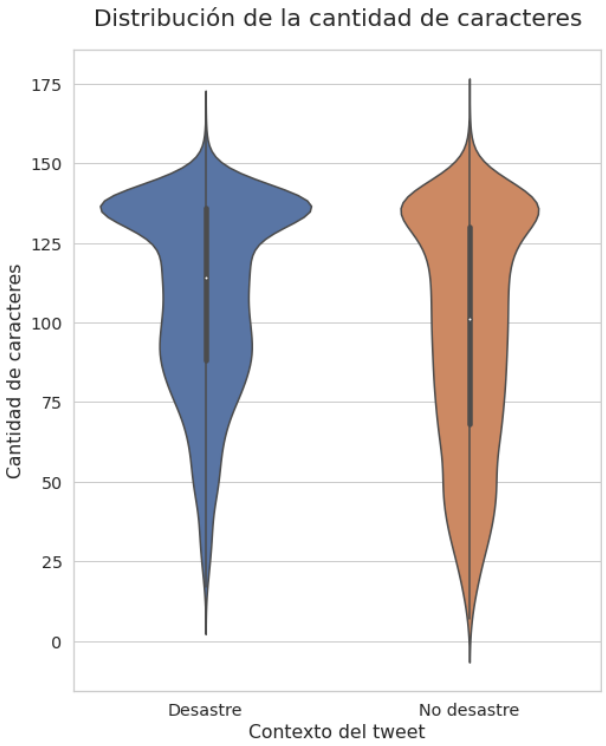
Dado que las palabras pueden tener varios significados según el contexto en el que se utilizan, una práctica casi obligatoria es la de analizar la frecuencia de los n-gramas, es decir, los pares de n palabras. De esta forma, se puede llegar a encontrar alguna noción de idea o concepto y no una simple palabra. Aquí se poder ver bi-gramas (dos palabras) únicos para cada grupo de tweets de desastre reales y no reales. Recordemos que se han quitado las stopwords, por lo que las parejas pueden resultar inconexas.

Figura 8. Diagrama de Venn sobre palabras frecuentes



Podemos concluir que en el conjunto de “tweets trágicos” se observan una mayor frecuencia de palabras con connotación negativa. Esto se acentúa en el análisis de los bi-gramas, donde se les da “contexto”.

Figura 8. Distribución de la cantidad de caracteres por tweet



Los tweets de desastre real son más largos que los falsos o de no desastre.

Conclusiones

Pudimos obtener la siguientes conclusiones:

- Los análisis anteriores confirman que hay diferencias entre los dos conjuntos de tweets de desastre reales y los falsos en base a longitud o cantidad de caracteres por tweet y frecuencia de palabras, bi-gramas y su contexto.
- En cuanto a ubicación, predominan los datos de Estados Unidos y Europa.