75.06/95.58 Organización de Datos - 1C 2020

Trabajo Práctico 1

Análisis Exploratorio

Grupo 33: "DataTravellers"

Integrantes:

- Andrés Pablo Silvestri: 85881 (silvestri.andres@gmail.com)
- Juan Manuel González: 79979 (juanmg0511@gmail.com)
- Patricio Pizzini: 97524 (pizzinipatricio@yahoo.com.ar)

Link a repositorio de GitHub:
https://github.com/patopizzini/Organizacion_Datos_1C2020/tree/master/TP1

Fecha de entrega: 21/05/2020

Contenido

1 - Introducción

Este trabajo está enfocado en hacer un primer análisis de los datos ofrecidos por Figure-Eight, de manera que encontremos particularidades que nos permitan resolver el problema planteado para la competencia de machine learning de Kaggle “Real or Not: NLP with Disaster Tweets[1]”. La misma consiste en determinar si un tweet está relacionado a un desastre o no.

Siendo este el caso, lo primero que vamos a hacer es interiorizarnos de uno de los pilares que tiene la ciencia de datos, el negocio, para luego realizar un análisis sobre los datos, a fin de contestar algunas preguntas que podrían resultar curiosas o de utilidad para nuestro objetivo.

1.1 - ¿Qué es Twitter[2]?

Twitter es un servicio de microblogging, con sede en San Francisco, California. Desde que Jack Dorsey lo creó en marzo de 2006, la red ha ganado popularidad mundial y se estima que tiene más de 300 millones de usuarios, generando 65 millones de tuits al día y maneja más de 800.000 peticiones de búsqueda diarias. Ha sido denominado como el "SMS de Internet".

La red permite enviar mensajes de texto plano de corta longitud, con un máximo de 280 caracteres, llamados tuits o tweets, que se muestran en la página principal del usuario. Los usuarios pueden suscribirse a los tweets de otros usuarios – a esto se le llama "seguir" y a los usuarios abonados se les llama seguidores. Por defecto, los mensajes son públicos, pudiendo difundirse privadamente y mostrándose únicamente a unos seguidores determinados. Los usuarios pueden twitear desde la web del servicio, con aplicaciones oficiales externas (como para teléfonos inteligentes), o mediante el Servicio de mensajes cortos (SMS) disponible en ciertos países.

En la actualidad Twitter factura más de 2.500 millones de dólares anuales y tiene un valor en bolsa superior a los 10.000 millones de dólares.

[1] Real or Not? NLP with Disaster Tweets: https://www.kaggle.com/c/nlp-getting-started
[2] Texto extraído Wikipedia: https://es.wikipedia.org/wiki/Twitter

2 - Estructura y manejo de los datos

A fin de mejorar la performance de ejecución de las operaciones realizadas, y por otro lado facilitar las tareas de análisis, se ha hecho un tratamiento previo sobre el set de datos proporcionado por la cátedra.

Se puede consultar el tratamiento completo, así como también el código utilizado para generar el análisis y los gráficos, en la carpeta “Entrega” del repositorio de GIT presentado en la carátula de este informe.

Como primer medida, se consideró tipar las columnas de los archivos en forma manual, a fin de mejorar el tiempo de carga y de uso de memoria. Esta estrategia fue recomendada en clase para datasets grandes, pero luego de cargar y analizar las características básicas del set del trabajo práctico se decidió que no era necesario ya que el tamaño era fácilmente manejable por nuestros equipos.

Contamos con un solo archivo para analizar en esta entrega, de aproximadamente 1MB de tamaño y 7.613 registros:

- **TRAIN.CSV** - tweets seleccionados de la plataforma twitter, con información de si pertenecen o no a un desastre (columna “target”).

Un tema que surgió durante el análisis fue el tratamiento de los registros con valores NaN en las columnas siendo investigadas. En estos casos fue necesario adoptar un criterio para solucionar esto: por lo general descartamos estos registros para la consulta que estábamos realizando, pero por ejemplo para el caso de location fueron completados con el dato “Unknown”. Para keywords, al ser un porcentaje muy bajo de NaN decidimos por no considerarlos para ciertos análisis, es decir que fueron dropeados al trabajar con dicha columna.

Otro aspecto importante del manejo de los datos se dió con el análisis realizado sobre las palabras de los tweets, más precisamente la columna “text”. Para ello utilizamos una biblioteca de procesamiento de lenguaje natural, NLTK[3]. Esto nos permitió separar las palabras (tokenizar), filtrar palabras comunes y otros tratamientos propios de este tipo de datos con relativa sencillez.

Por último, vale agregar que en algunos casos nos hemos tomado la libertad de considerar algunos valores como despreciables en cuanto a relación de cantidad sobre el total, como así también hacer un análisis sobre los casos que más cantidad de registros nos brindan, lo que nos permite entender la situación con un volumen de datos mayor lo cual nos acerca un poco más a la realidad.

[3] Natural Language ToolKit: https://www.nltk.org

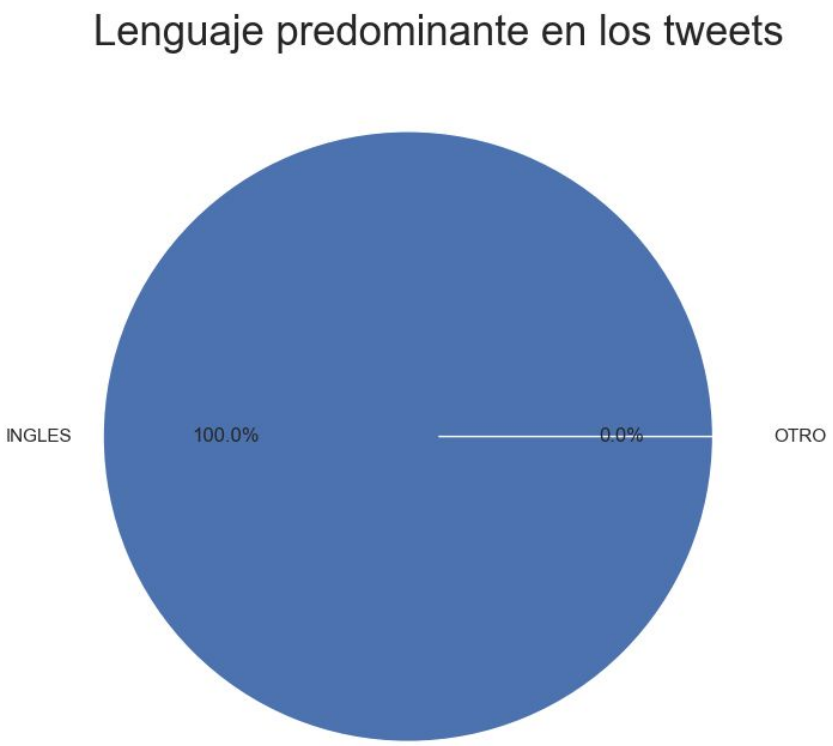
3 - Análisis general sobre los tweets y su tamaño

En esta sección analizaremos aspectos generales sobre los tweets que forman parte del dataset, como por ejemplo su distribución según el valor del target, su longitud o el contenido de la columna *location*.

3.1 - ¿Cuál es la distribución de los tweets respecto a su target?

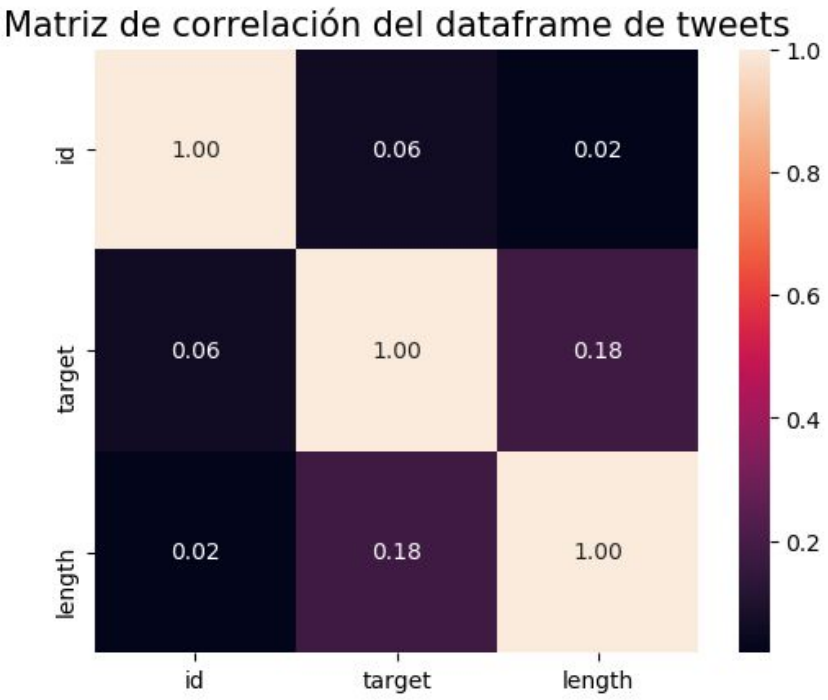
Un primer aspecto que nos propusimos analizar, es ver cómo están repartidos los tweets respecto a su valor de target. Como se puede apreciar, esta distribución es bastante pareja, con una ventaja de tweets con target 0, es decir que no están relacionados a desastres.

3.2 - ¿Cuáles son los idiomas predominantes en los tweets?



Podemos ver, o más bien confirmar, que el idioma en los tweets es siempre el inglés, al menos en este muestreo que manejamos, con lo cual todo trabajo sobre el contenido de los tweets se puede y debe hacer pensando en el idioma inglés.

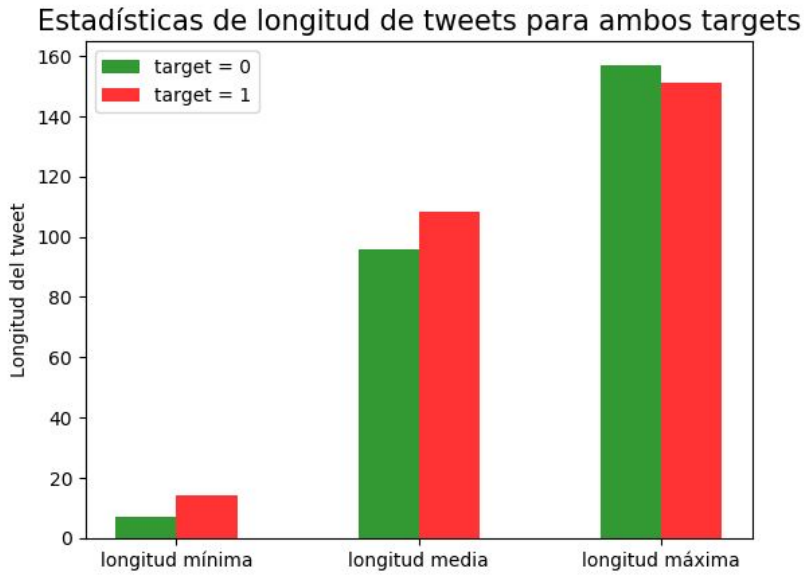
3.3 - ¿Existe correlación entre la longitud de los tweets y su target?



Para comenzar a responder esta pregunta graficamos la matriz de correlación entre las variables del dataframe de tweets, adicionando la longitud de los mismos, tomada como el *length* de la columna text.

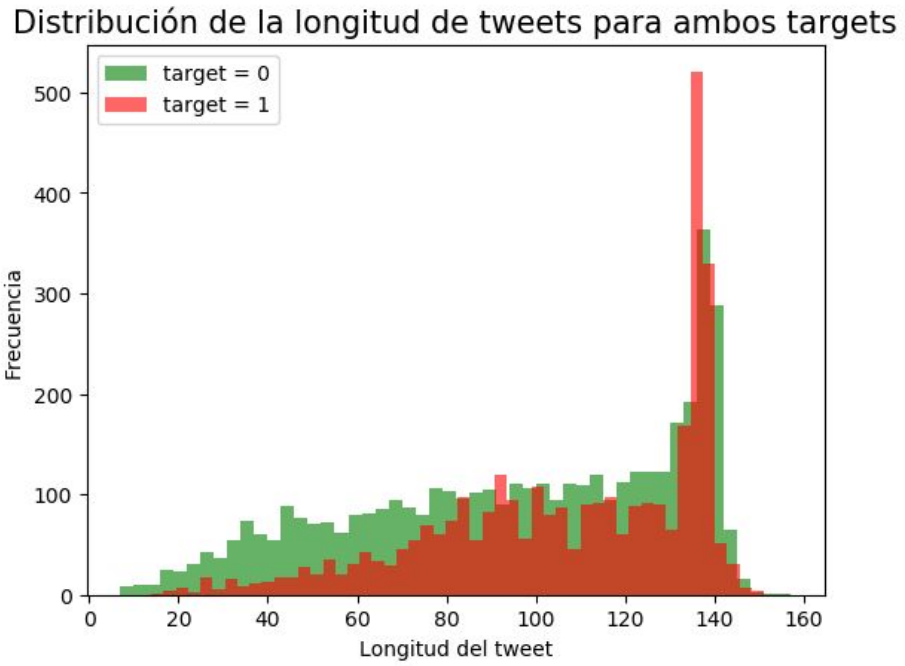
Podemos ver que existe una correlación débil entre ambas variables, entraremos más en detalle en las preguntas subsiguientes.

3.4 - ¿Cuáles son las longitudes mínimas, medias y máximas, para cada target?



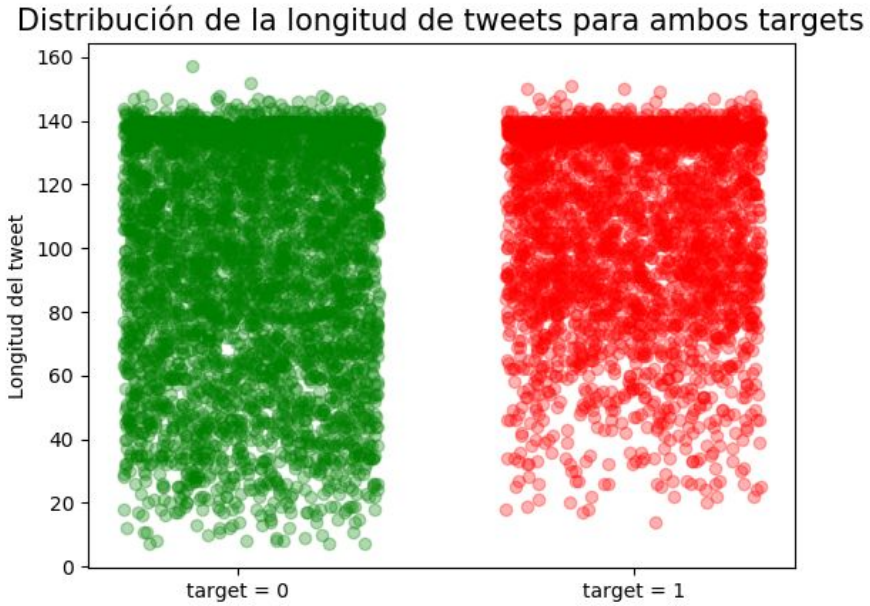
En esta visualización podemos observar las diferencias en los valores mínimos, máximos y medios de longitud par los tweets del dataset, separadas por target. No se evidencia una diferencia que nos permita sacar conclusiones acerca de la relación entre longitud y target, pero nos da una idea de la dimensión de los textos.

3.5 - ¿Cómo se distribuye la longitud de los tweets según su target?



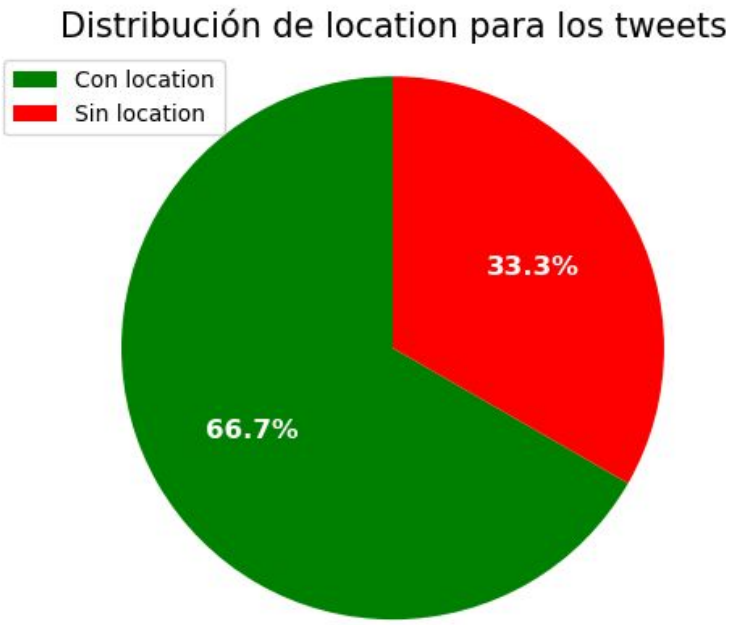
El gráfico nos muestra cómo para target igual a 1, los tweets tienden a ser ligeramente más largos, mientras que para target igual a 0 se encuentran más uniformemente distribuidos.

Sin embargo, vemos que el pico se da en el mismo intervalo, con una mayor cantidad para el caso positivo.



El plot mostrado arriba confirma lo evidenciado en el histograma: puede verse cómo la densidad para casos negativos es más uniforme, mientras que para el caso en el que se confirma un desastre las longitudes de los tweets se encuentran más amontonadas en la parte superior, indicando que hay más registros con longitudes mayores.

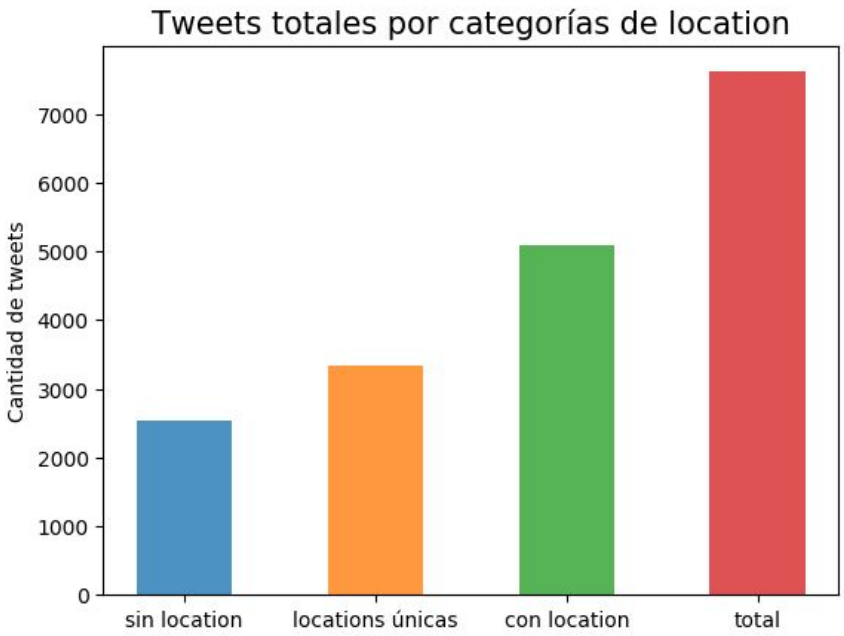
3.6 - ¿Qué cantidad de tweets incluye información sobre location?



Otro aspecto interesante del dataset que investigamos, fue el contenido de la columna location. Para comenzar, al ver en un análisis preliminar que la columna contaba con campos nulos, decidimos mostrar qué porcentaje de tweets tiene location nula, y encontramos que aproximadamente ⅓ de los registros del dataset no incluye esta información.

Como ya vimos, estos valores fueron rellenados con el valor “Unknown”.

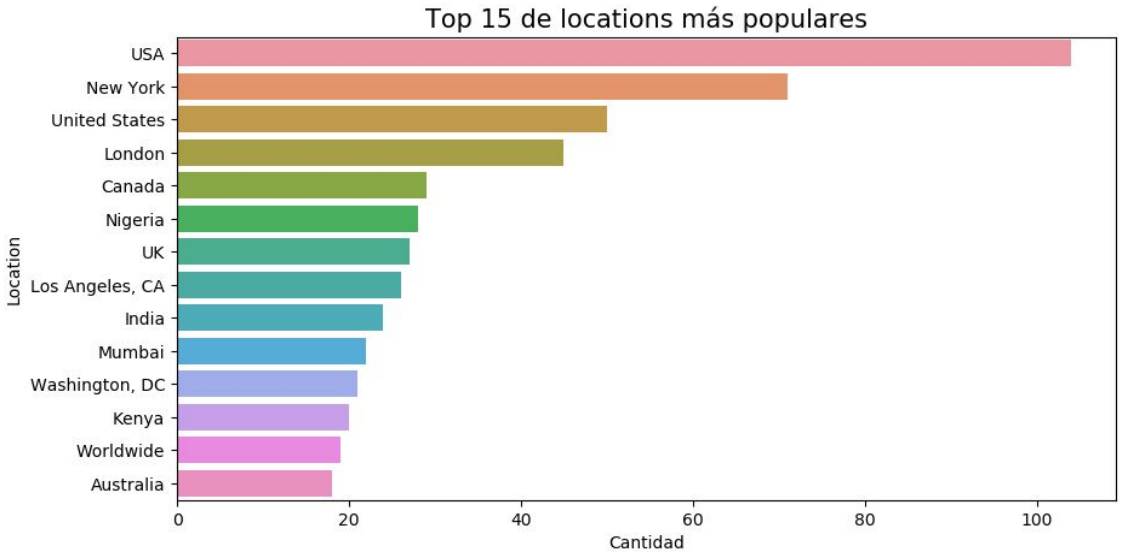
3.7 - ¿Cuántas location únicas existen?



Un siguiente análisis fue ver la calidad de esta información. Nuestra idea era tratar la columna como categoría para poder extraer la mayor información posible, pero como vemos en el gráfico, de un total aproximado de 5000 tweets con location, unas 3500 locations son únicas.

Esto nos da la pauta que los datos probablemente no están generados a partir de valores comunes, ni normalizados.

3.8 - ¿Cuáles son las 15 locations más populares?

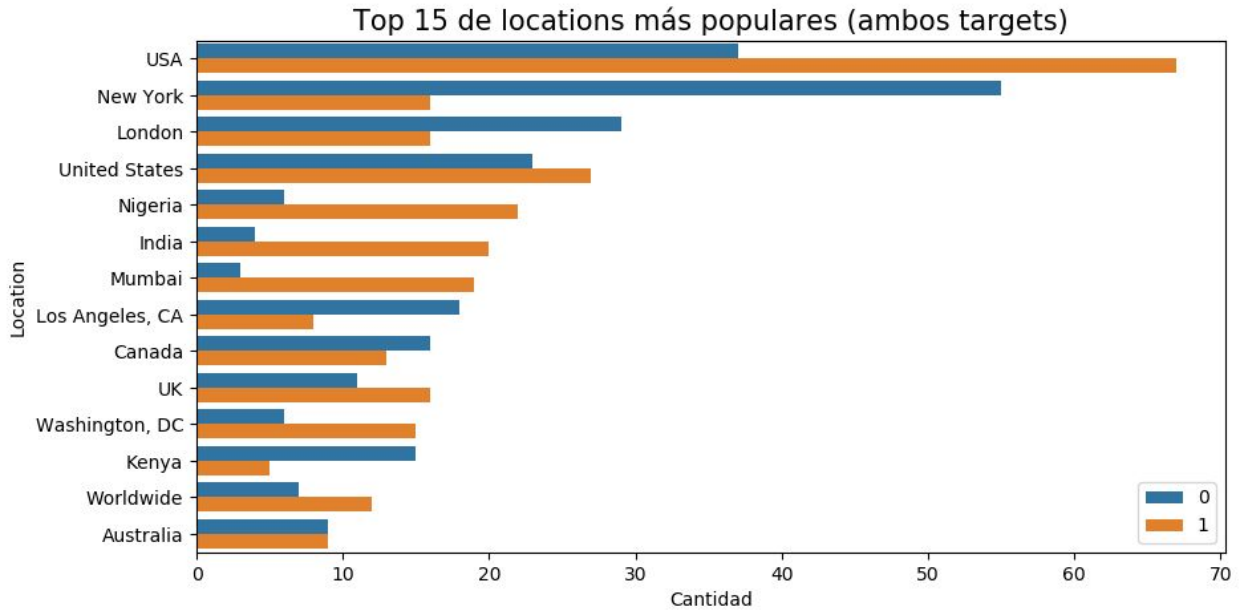


Al mostrar los valores para las 15 locations con mayor cantidad de registros (eliminando el "Unknown" rellenado por nosotros), confirmamos la sospecha del punto anterior, ya que tenemos países, provincias/estados y ciudades mezclados entre los datos, así como también el valor "Worldwide".

De estos 15, hay 5 que pertenecen a Estados Unidos, y en particular el primero y el tercero son equivalentes. Que el primer valor apenas supere los 100 registros nos permite confirmar, además, la gran heterogeneidad presente en los datos.

Entendemos que podría hacerse un trabajo de saneamiento y unificación, lo que nos permitiría llegar a la distribución real (por lo menos por países): este sería un trabajo principalmente manual, dada la cantidad de registros se necesitaría una motivación que justifique tal esfuerzo, hecha esta investigación preliminar.

3.9 - ¿Cómo se distribuyen las 15 locations más populares según el target?



Finalmente, repetimos el plot anterior para las mismas 15 locations, pero contando las ocurrencias por valor de target.

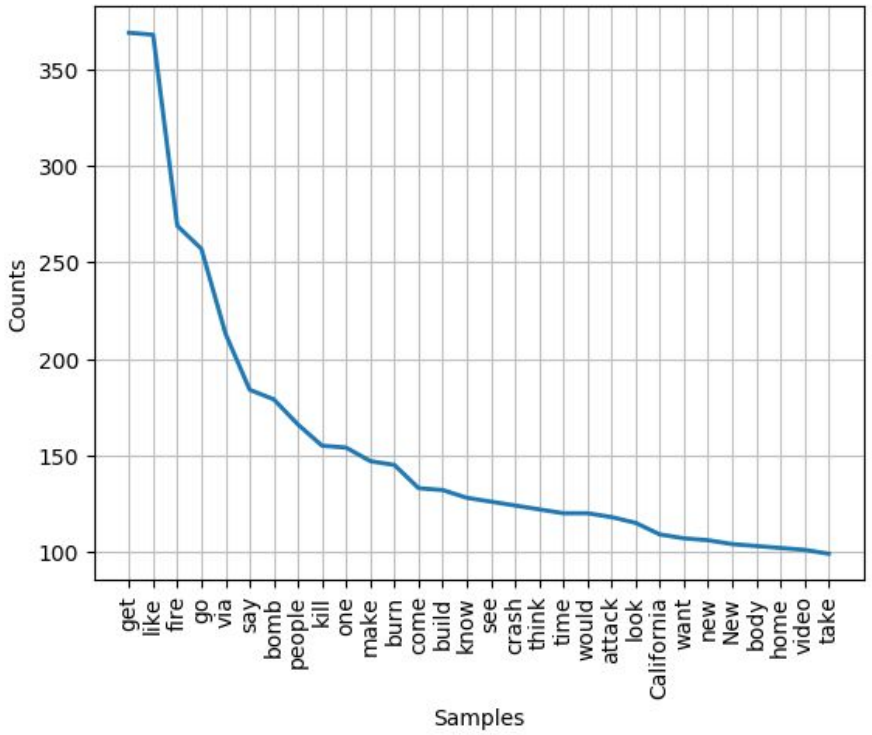
Se observan algunos casos interesantes, como la India o Nigeria, donde prevalece una mayor cantidad de registros con target positivo.

4 - Análisis general sobre las palabras que forman los tweets

En esta sección se analizarán las palabras que forman parte de los tweets, excluyendo los *hashtags* y *keywords*, que serán analizados en otras secciones, y las referencias a otros usuarios de la plataforma.

Para efectuar esta parte del trabajo utilizamos una biblioteca de procesamiento de lenguaje natural, como ya mencionamos se trata de NLTK.

4.1 - ¿Cuáles son las palabras más populares en el set de datos?



Como primera aproximación, decidimos investigar la frecuencia de aparición de las palabras en el set de datos.

Esto necesitó de un procesamiento previo, que fue en gran parte facilitado por el uso de la librería de procesamiento de lenguaje, para evitar resultados obvios o de poco valor.

Los pasos realizados fueron:

1. Separación de los tweets en sus palabras individuales mediante el *tokenizer* especial para twitter de NLTK.
2. Eliminación de *hashtags* y referencias a usuarios.
3. Limpieza de caracteres especiales o no imprimibles.
4. Eliminación de signos de puntuación.
5. Remoción de stopwords, para idioma inglés.
6. Lematización de las palabras.

Los últimos dos pasos, en particular, resultan cruciales para filtrar y unificar las palabras presentes en el set resultante, a fin de obtener un resultado donde las palabras tengan valor real para el análisis a realizar.

El trabajo realizado también se puede visualizar en forma de nube de palabras, como se puede apreciar a continuación.

Palabras más populares

Observando los resultados, vemos que existen varias palabras indicativas de desastres naturales, como “fire”, “flood”, “spill” y “storm”. Por otro lado, tenemos algunas otras que también indicarían una situación de desastre, pero por situaciones generadas por el hombre, como “attack”, “bomb” y “kill”.

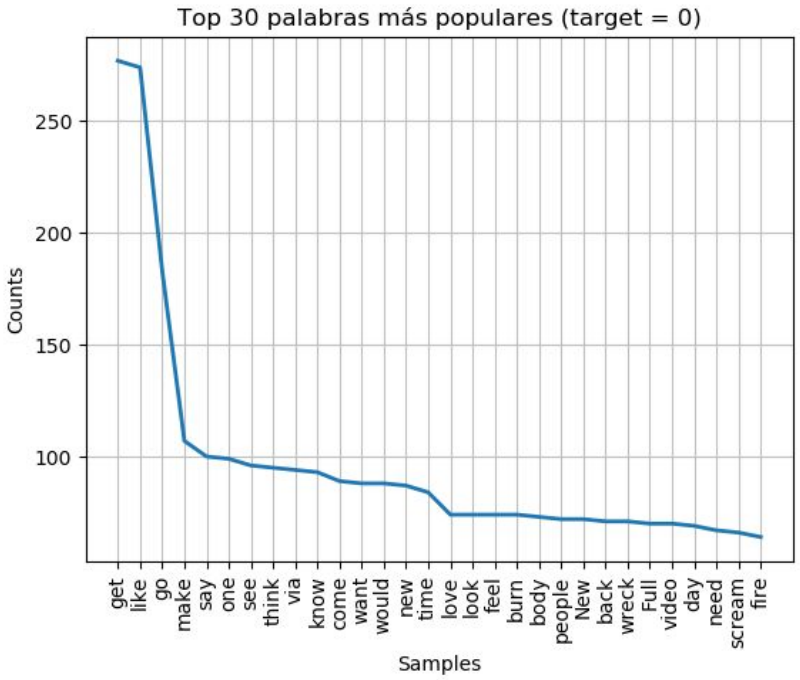
Esta clasificación no puede tomarse como definitiva, recordemos que tenemos las palabras lematizadas, y por ejemplo “kill” puede referirse tanto a muertes por un desastre natural como por una acción humana. Asimismo, palabras como “fire”, o “flood” pueden utilizarse para referirse a otras cosas o como parte de una frase, por lo que sin analizar el contexto no permiten determinar si son referidas a un desastre o no.

Finalmente, hay palabras relativamente comunes que podrían ser eliminadas para mejorar aún más el resultado. Si bien se eliminaron las denominadas stop words, durante la etapa de preprocesamiento, podemos nombrar como ejemplos a “go”, ”like” y “say” entre las que quedaron y podrían filtrarse.

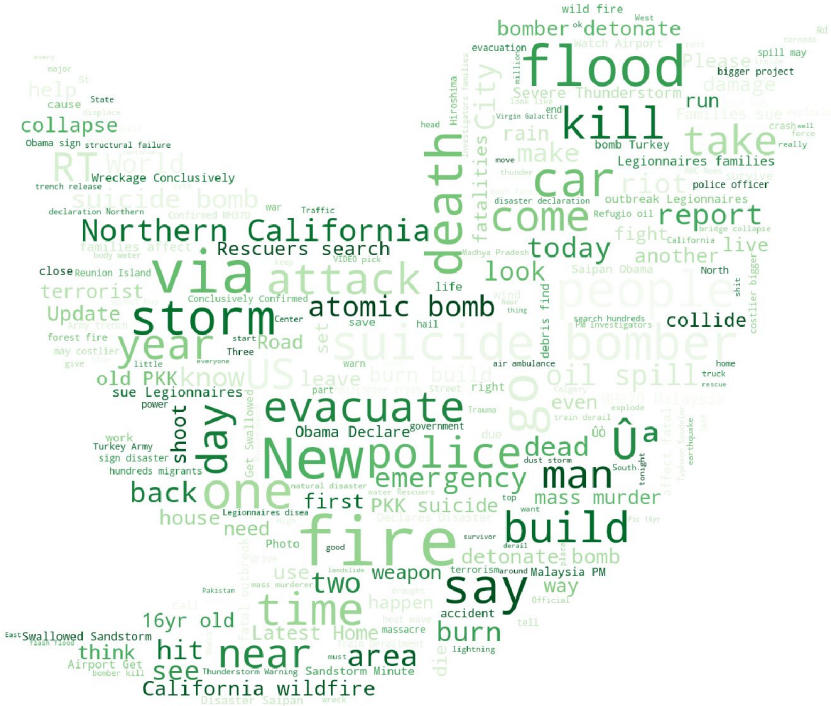
Sin embargo, nos pareció interesante observar globalmente qué palabras tenemos en el dataset objeto de estudio. Interesante es también la aparición de “California” entre los resultados, uno de los estados de Estados Unidos con un problema recurrente de incendios forestales, o “MH370[4]” un vuelo desaparecido en Marzo del año 2014 de la compañía Malaysia Airlines.

[4] Vuelo 370 de Malaysia Airlines: https://es.wikipedia.org/wiki/Vuelo_370_de_Malaysia_Airlines

4.2 - ¿Cuáles son las palabras más populares en el set de datos, para target = 0?

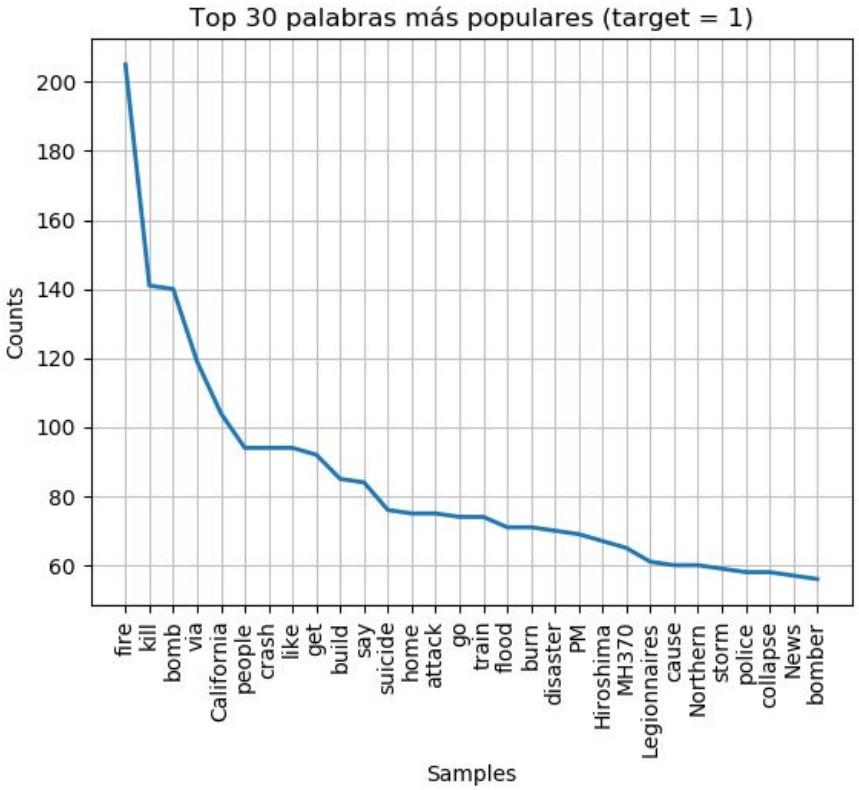


Palabras más populares (target = 0)



Repitiendo el análisis, pero filtrado para tweets que no representan desastres, vemos que aparecen igualmente palabras asociadas a desastres, en consonancia con lo explicado en la pregunta anterior.

4.3 - ¿Cuáles son las palabras más populares en el set de datos, para target = 1?



Palabras más populares (target = 1)

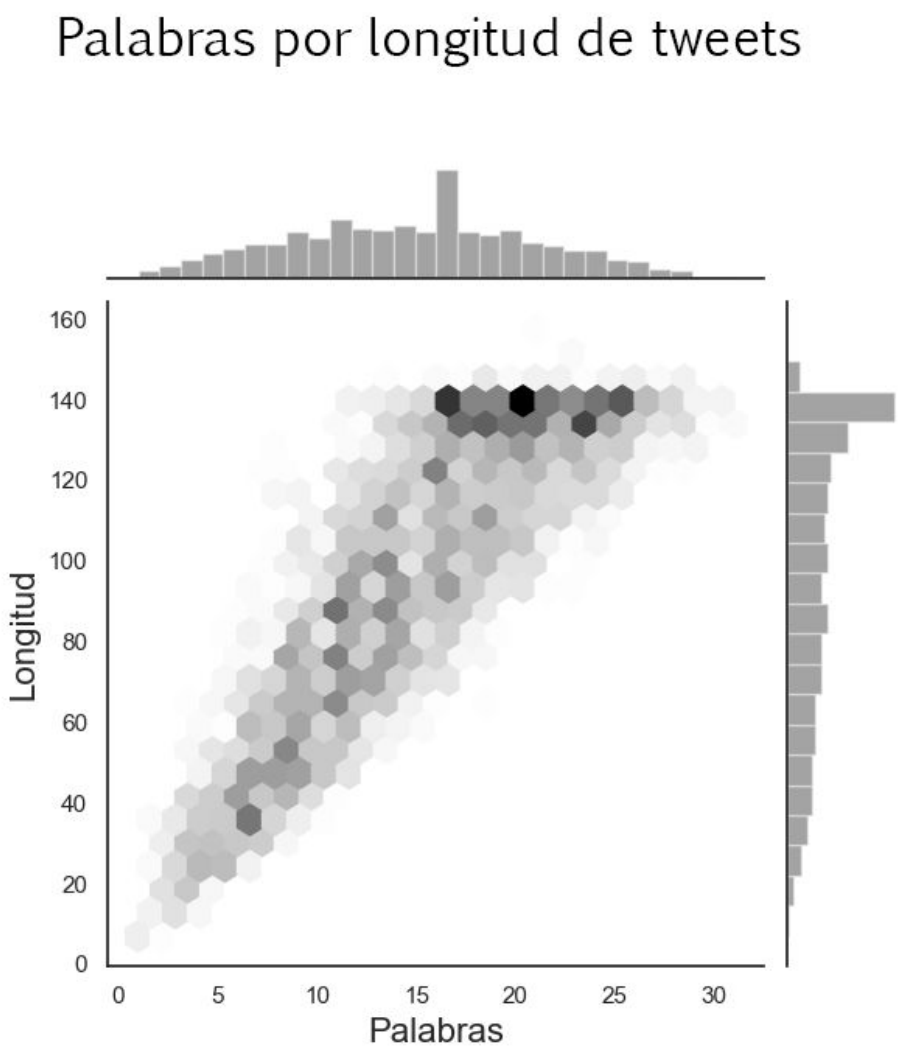
Finalmente, repetimos el análisis, pero quedándonos con los tweets que sí representan un desastre.

Vemos que también están presentes palabras como “kill”, “flood” o “fire”, que como explicamos pueden depender del contexto, pero como valor adicional en este caso tenemos la certeza de que sí están referidas a un desastre natural.

La visualización es sin embargo mucho más valiosa que las anteriores, dado que prácticamente todas las palabras que vemos destacadas hacen referencia a desastre.

A modo de conclusión, podemos decir que no existe una sola palabra que pueda determinar fehacientemente si estamos ante la presencia de un tweet que hace referencia a un desastre, pero si existe una lista de palabras que nos puede hacer llamar la atención para analizar el registro con técnicas más avanzadas, como un análisis contextual, de *mood*, o bien los valores de los datos presentes en el resto de las columnas del dataframe, como por ejemplo los *hashtags* o *keywords*.

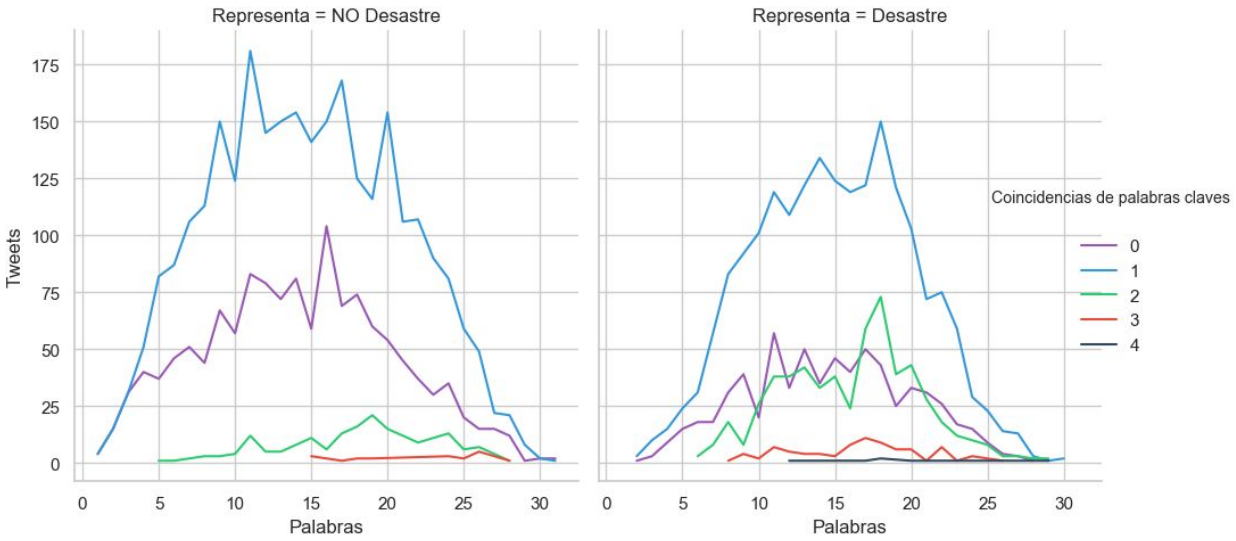
4.4 - ¿Cómo se relaciona la longitud de los tweets con la cantidad de palabras?



Como podemos observar tenemos, una gran concentración de tweets que tienen una longitud de entre 120 y 140 caracteres, como así también la gran mayoría tiene un rango de entre 15 y 25 palabras, lo que nos permite observar que esto es lo más normal a la hora de escribir tweets.

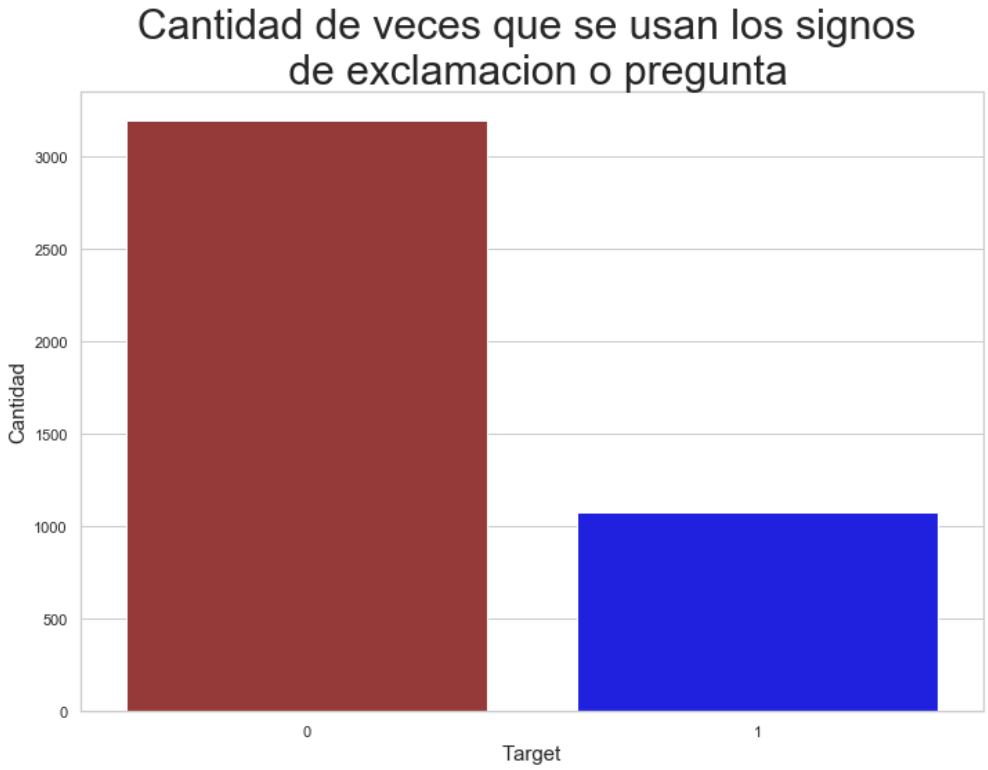
4.5 - ¿Cuál es la interpretación entre tamaño, cantidad de palabras y las palabras claves?

Vale aclarar que para este gráfico la idea es tener una aproximación de qué tan relevantes son algunas palabras claves y sus apariciones, a fin de distinguir un tweet como desastre o no desastre.



Se puede ver que cuando es un desastre, las apariciones de palabras claves en los tweets es mucho más frecuente, y que cuando no es un desastre la curva que representa 0 coincidencias es mucho más amplia. Esto seguro será mucho más relevante cuantas más palabras claves agreguemos. Para ver todo el listado de palabras usado, por favor dirigirse al repositorio de github.

4.6 ¿Hay alguna relación entre los signos de exclamación (!)/pregunta (?) y el target?



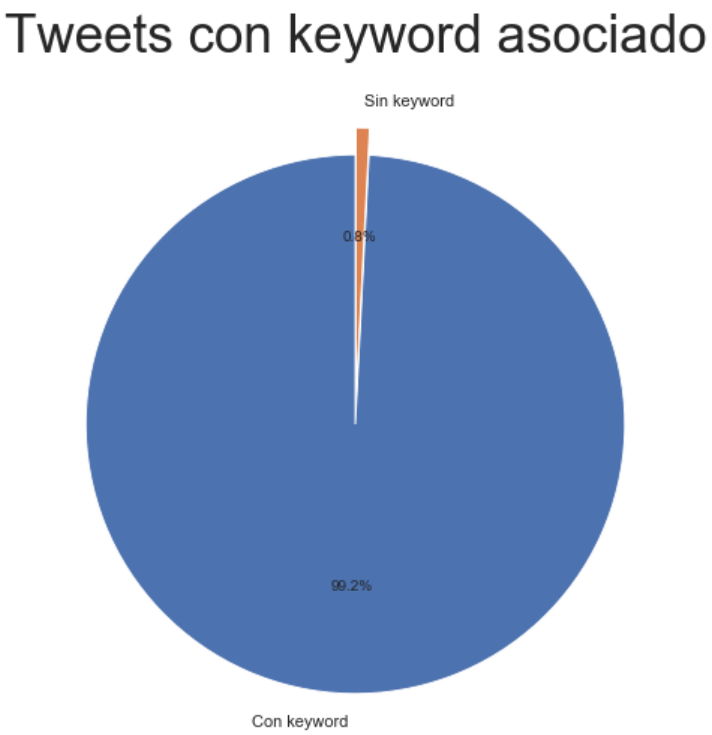
Como podemos ver, en los tweets que no hablan de desastres se usan casi 3 veces más los signos de pregunta o exclamación.

Esto puede deberse a que la gente escribe con más seriedad y veracidad, es decir que se hacen menos preguntas, en los tweets que si hablan de desastres.

Contrario a lo que uno pudiera pensar, que una persona usaría los signos de exclamación para alertar sobre algo desastroso que está sucediendo, en este caso vemos que no es tan así.

5 - Análisis general sobre las keywords asignadas a los tweets

5.1 - ¿Cuántos tweet están asociados a una keyword?

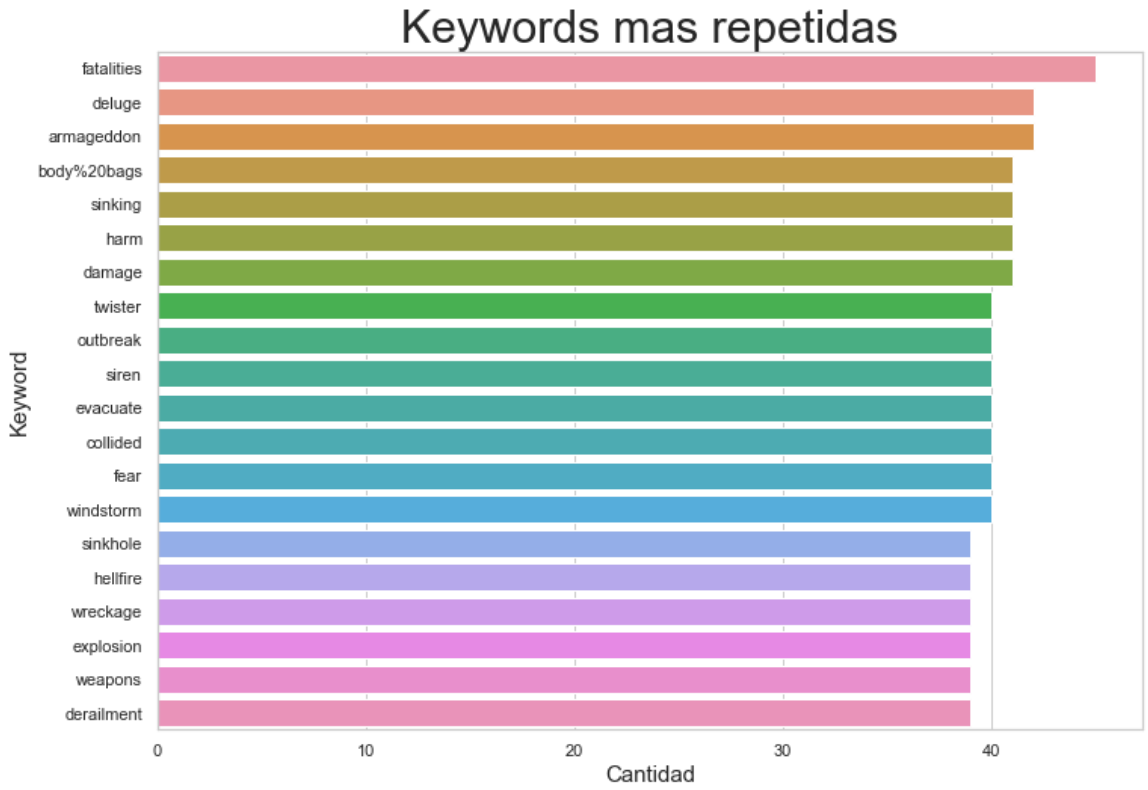


Como se puede ver en el gráfico, podríamos decir que casi todos los tweets están asociados a una keyword, es demasiado pequeño el porcentaje que no está asociado a una palabra clave como para tenerlo en cuenta.

En este análisis no hacemos diferencia si dicho tweet está referido a una catástrofe o no.

5.2 - ¿Cuáles son las keywords más utilizadas?

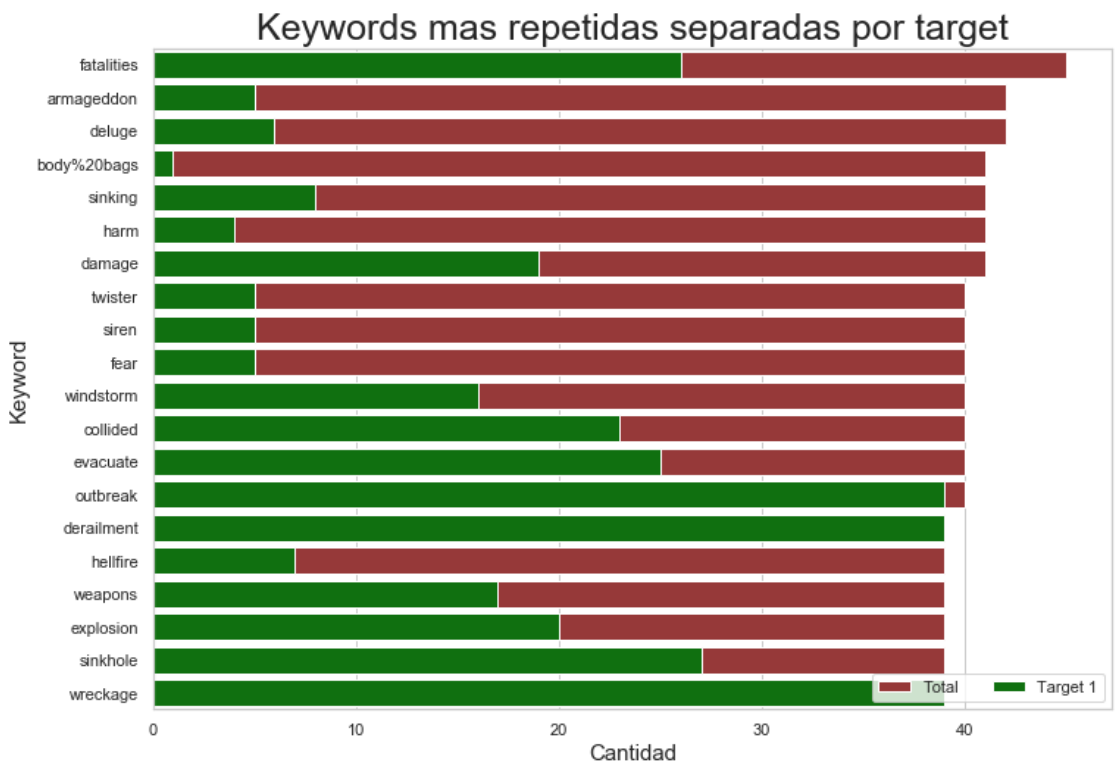
Aclaramos primeramente que se tomó una muestra de las 20 keywords más usadas, ya que si usaban muchas más el gráfico empezaba a perder visibilidad.



Analizando el gráfico, podemos decir que muchas de las keywords hacen referencias a palabras más generales, como es el caso de por ejemplo “fatalities” (fatalidades) o “harm” (daño) por nombrar algunas.

Es decir, muchas de ellas son términos que uno podría utilizar para varios desastres, también se podrían usar para tweets no relacionados con los mismos. Es decir, que un tweet hable de “fatalities” puede estar refiriéndose a un hecho doméstico o cotidiano y no propio de un desastre.

5.3 - ¿Cómo se distribuyen por target dichos keywords?

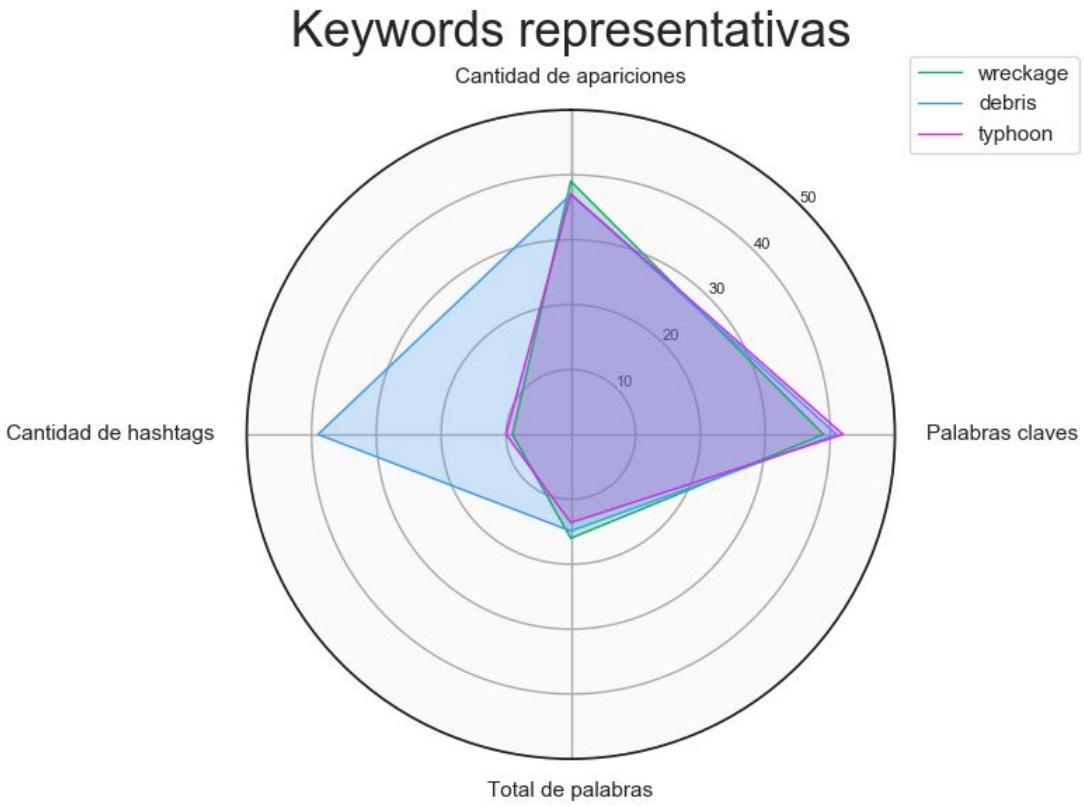


Como podemos ver, muchos de los keywords anteriores están referidos a hechos que no representan desastres, sin embargo hay algunos muy marcados donde notamos que si se trata de desastres, por ejemplo “evacuate”, “derailment” o “wreckage”, donde casi el 100% de los mencionados hace referencia a un desastre que sí ocurrió.

En otros aspectos, tenemos palabras en donde muy pocas ocurrencias son referidas a desastres y que uno capaz imagina que sí podrían haber sido, como “siren” o “fear.

5.4 - ¿Cómo trabajan las keywords más representativas?

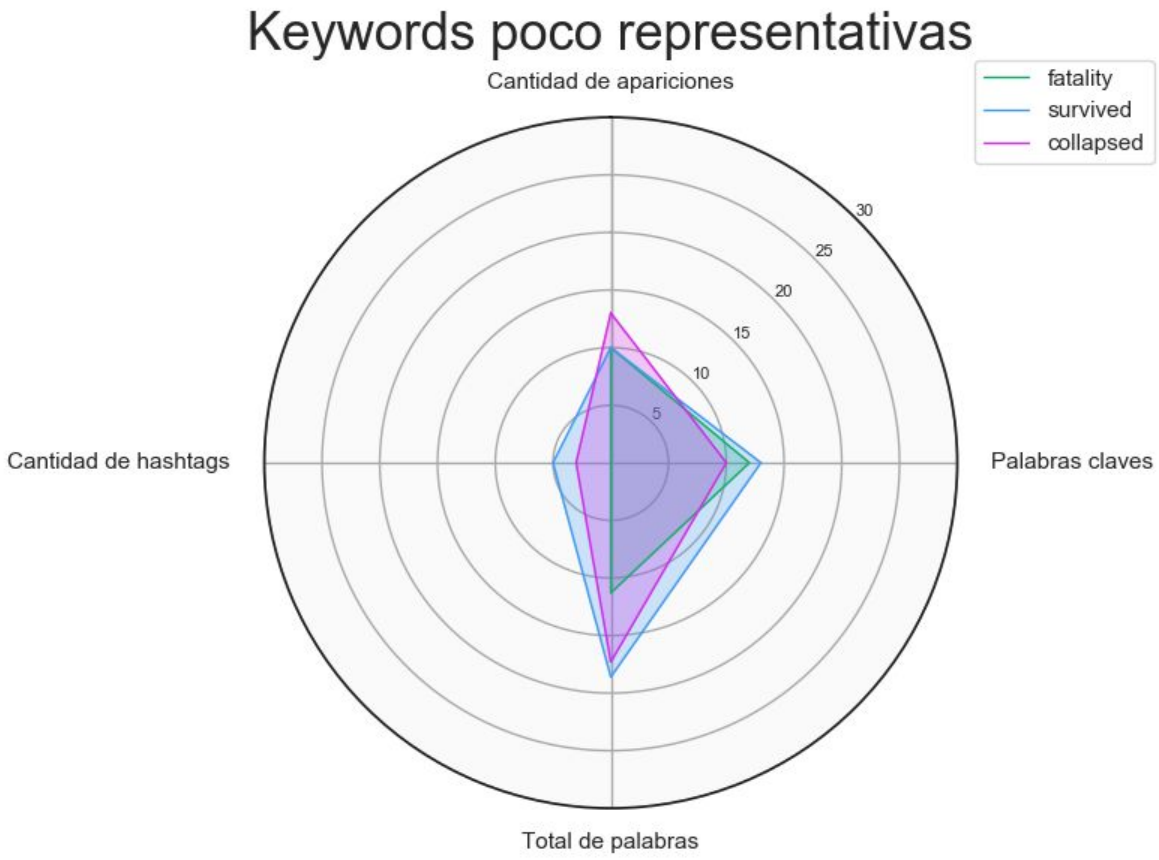
Se han tomado 3 de las keywords más representativas en los tweets que representan un desastre, y lo que buscamos es ver como varían diferentes valores de manera que nos permita obtener algún patrón o diferencia sustancial.



Podemos ver que en la mayoría de los casos el comportamiento es similar, salvo por ejemplo para debris (escombros), donde hay un interesante incremento en el uso de hashtags pero que no necesariamente puede estar vinculado a un desastre, sino que también puede haber casos de demolición, construcción o similares.

5.5 - ¿Cómo trabajan las keywords menos representativas?

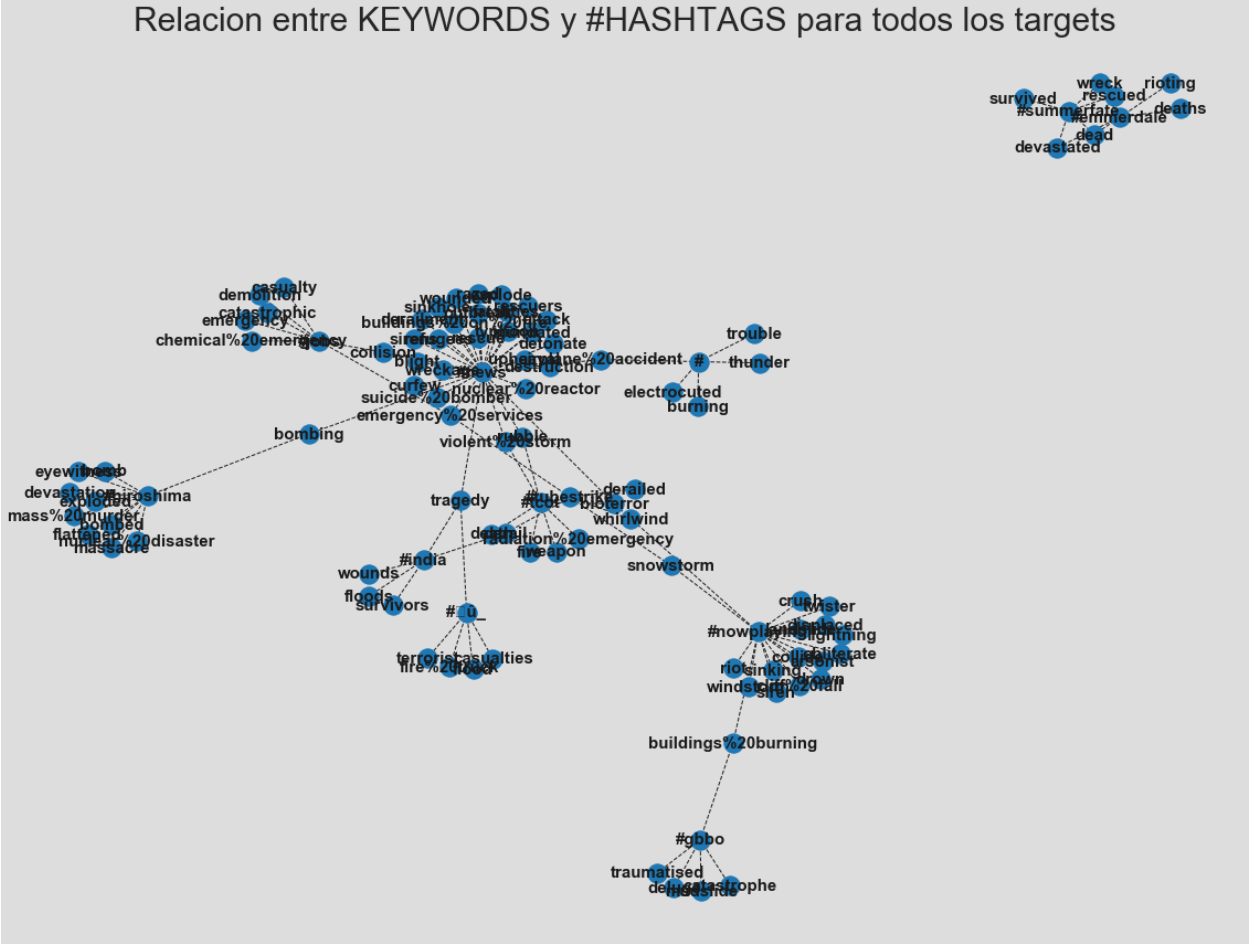
Se han tomado 3 de las keywords menos representativas en los tweets que representan un desastre, y lo que buscamos es ver como varían diferentes valores de manera que nos permita obtener algún patrón o diferencia sustancial.



Como en el caso anterior, vemos que el comportamiento es muy similar entre un caso y otro, para este ejemplo podemos ver que fatality (fatalidades) suele tener menos cantidad de palabras, pero tampoco es algo tan representativo, por lo que podemos sacar que los promedios se mantienen.

5.6 - ¿Qué relación hay entre los keywords y los #hashtags?

Para ver dicha relación, tomamos los hashtags que estaban en 4 o más diferentes keywords, ya que sino nos generaba una dispersión muy grande.

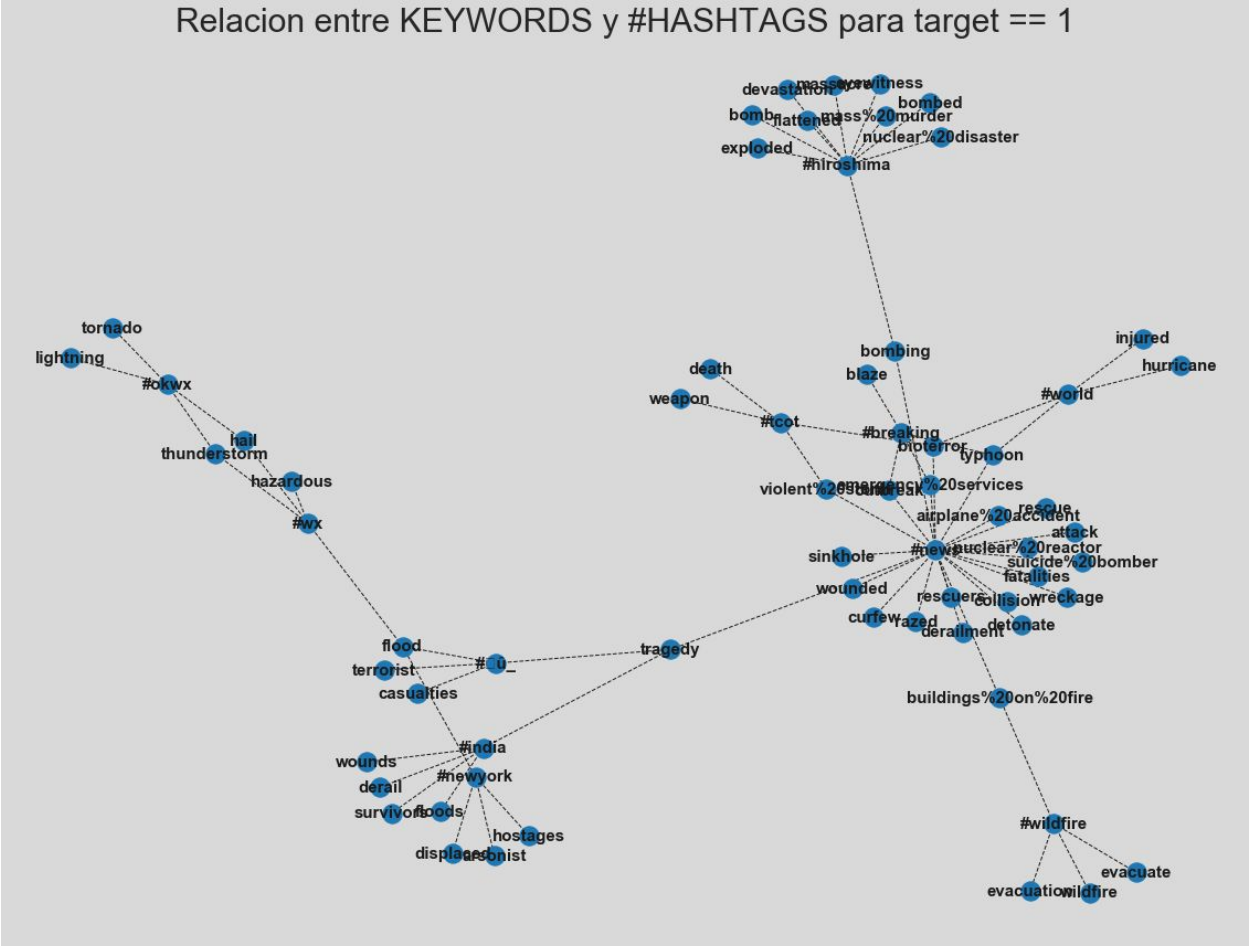


Como se puede ver en el gráfico hay 4 grupos grandes de hashtags que atraen a varios keywords, son los casos de “#News”, “#Hiroshima”, “#NowPlaying” y “#summertime”-“emmerdale”.

A simple vista podríamos asociar algunos, como #Hiroshima a algo que realmente ocurrió y otros como #NowPlaying o #Summertime nos parece que son cosas que no condicen con desastres, pero lo analizaremos en los próximos puntos.

5.7 - ¿Qué relación hay entre los keywords y los hashtags en los tweets que hablan de desastres?

Tomando la idea del anterior, acá también utilizamos solo los hashtags que estaban en más de 3 keywords diferentes.



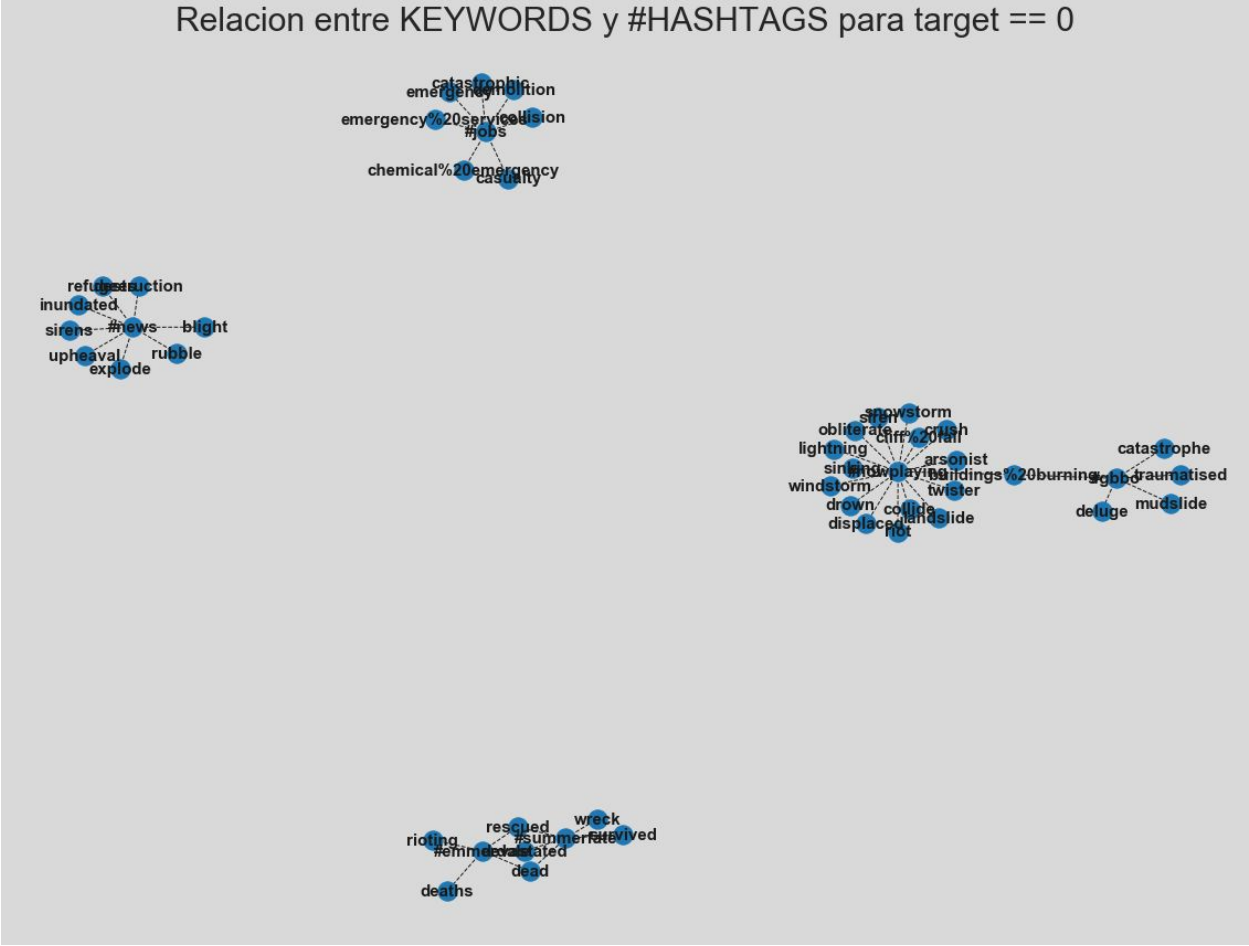
Podemos observar varias uniones, una muy importante es la de “#News”, también estaba en el anterior punto, que acapara la atención de varios keywords, podríamos decir que la gente suele querer informar algo que ocurrió diciendo esa palabra.

También nos dice que “#Hiroshima” es muy mencionada y, si bien el hecho real fue una bomba atómica, se puede mencionar que la gente los relaciona con diferentes keywords pero siempre relacionados a bombas o explosiones.

Entre otros hashtags destacados vemos que hay nombres de ubicaciones (“#New York” e “#India”), podemos inferir que cuando la gente quiere tweetear sobre algún desastre, lo hace refiriéndose a la ciudad donde ocurrió.

5.8 - ¿Qué relación hay entre los keywords y los hashtags en los tweets que NO hablan de desastres?

Aca tambien tomamos los hashtags que estaban relacionados a los keywords, de forma única, en más de 4 veces.



Podemos ver que son cinco grupos muy marcados de un solo hashtag cada uno relacionado con keywords separadas.

Acá se destaca “#NowPlaying”, que como mencionamos anteriormente no se referiría a un desastre. En ese mismo sentido se encuentra la combinación de tweets con los hashtags “#Summertime”-”#emmerdale”, que era fácil suponer que tampoco hacía referencia a dichos eventos.

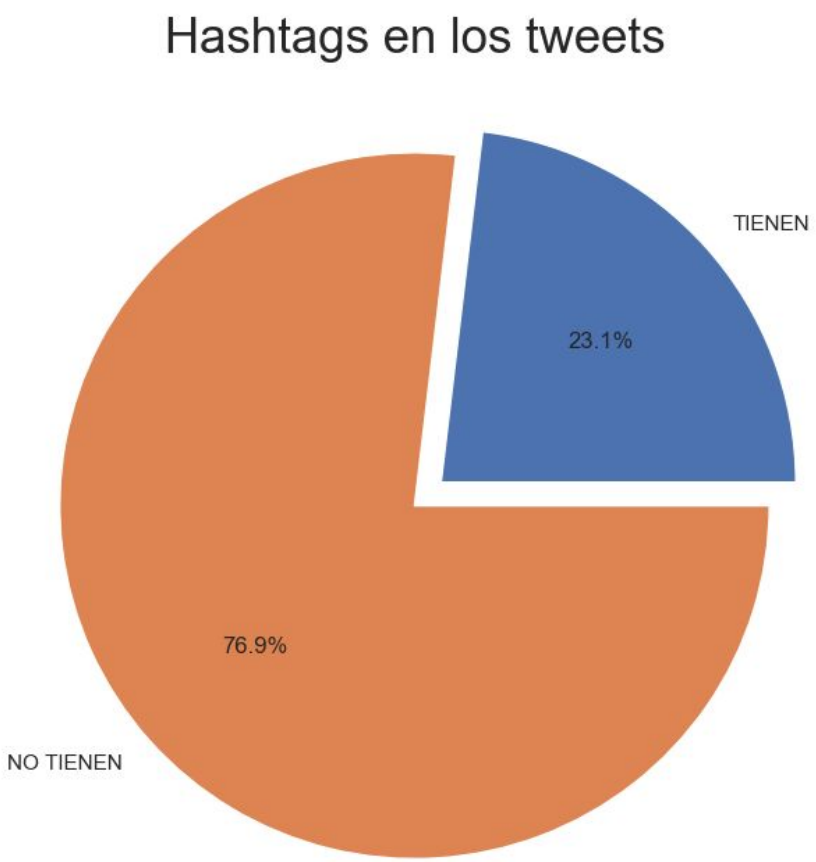
También se encuentran otros como “#News” o “#Jobs” que pueden hacer referencias a cosas cotidianas.

6 - Análisis general sobre los hashtags utilizados en los tweets

Un **hashtag** de **Twitter** es simplemente una frase de palabras clave, escritas sin espacios y precedidas por un numeral (#) o una sola palabra en lugar de varias. Los **hashtags** permiten agrupar los contenidos u opiniones en distintas temáticas.

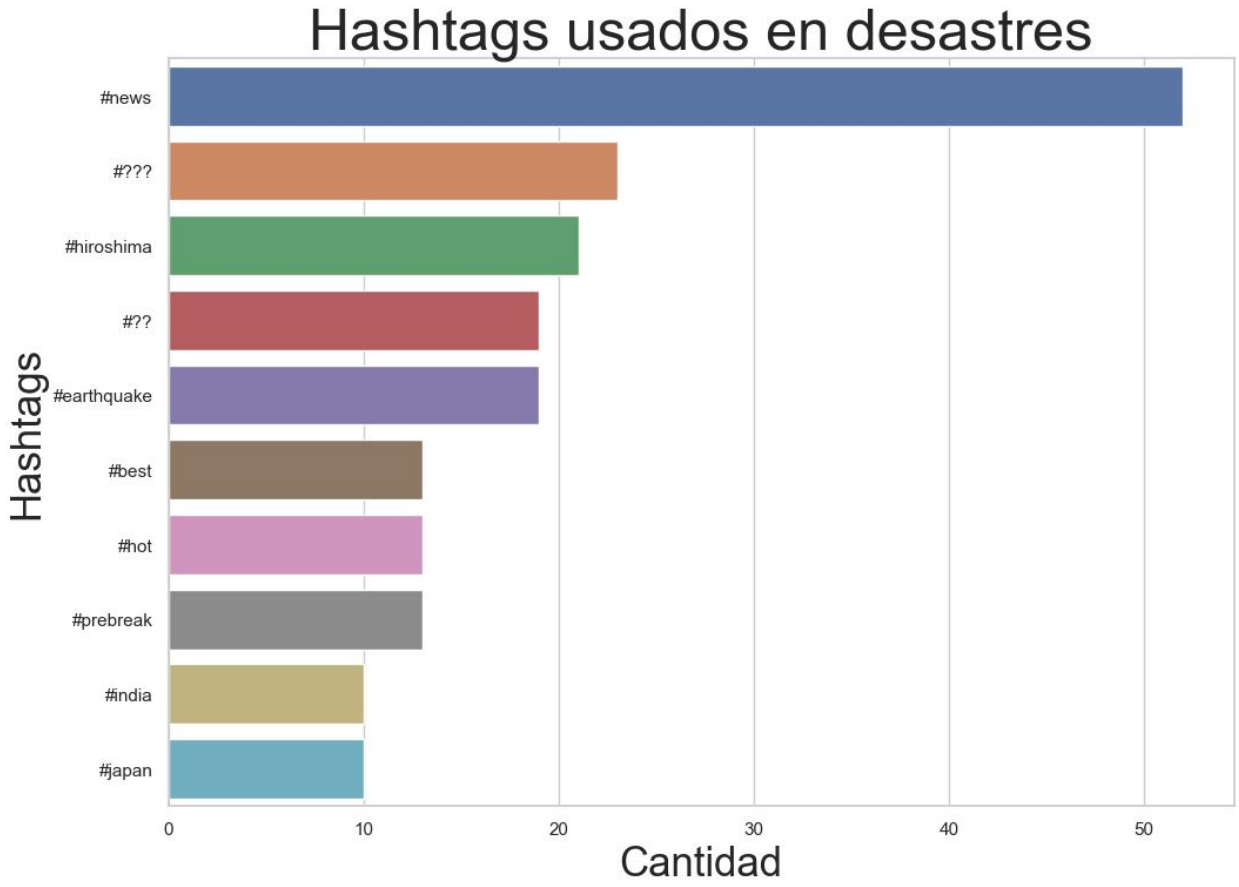
6.1 - ¿Cuál es la proporción de hashtags sobre el total de tweets?

En el siguiente gráfico buscamos obtener como relación que tanto son usados los hashtags en relación al uso general de los tweets.



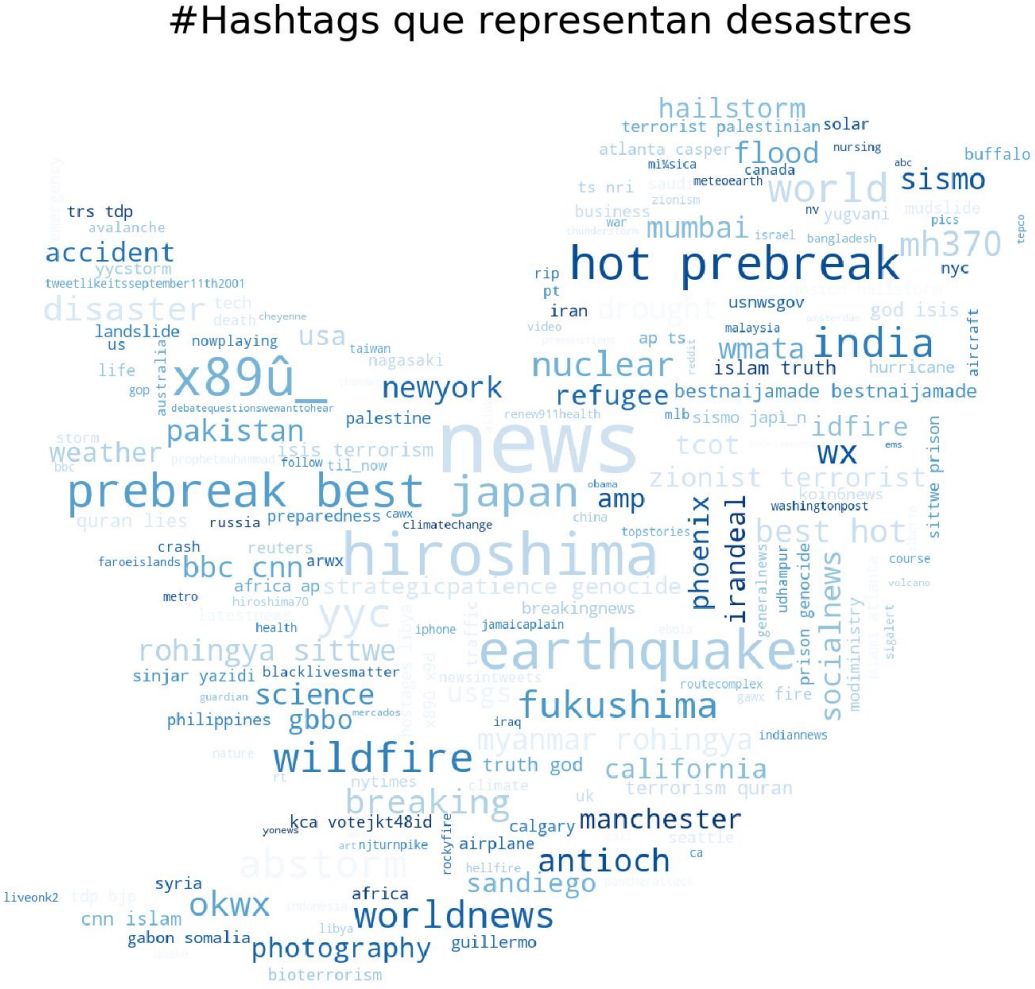
Podemos ver que solo en un cuarto de los casos se usan hashtags, lo que nos da a pensar que no son tan relevantes en el contenido del tweet, de todas maneras se entiende que es una funcionalidad bastante extendida y que a pesar de no ser tan usada puede ser clave para informar algo.

6.2 - ¿Cuáles son los hashtags más usados para tweets que representan desastres?



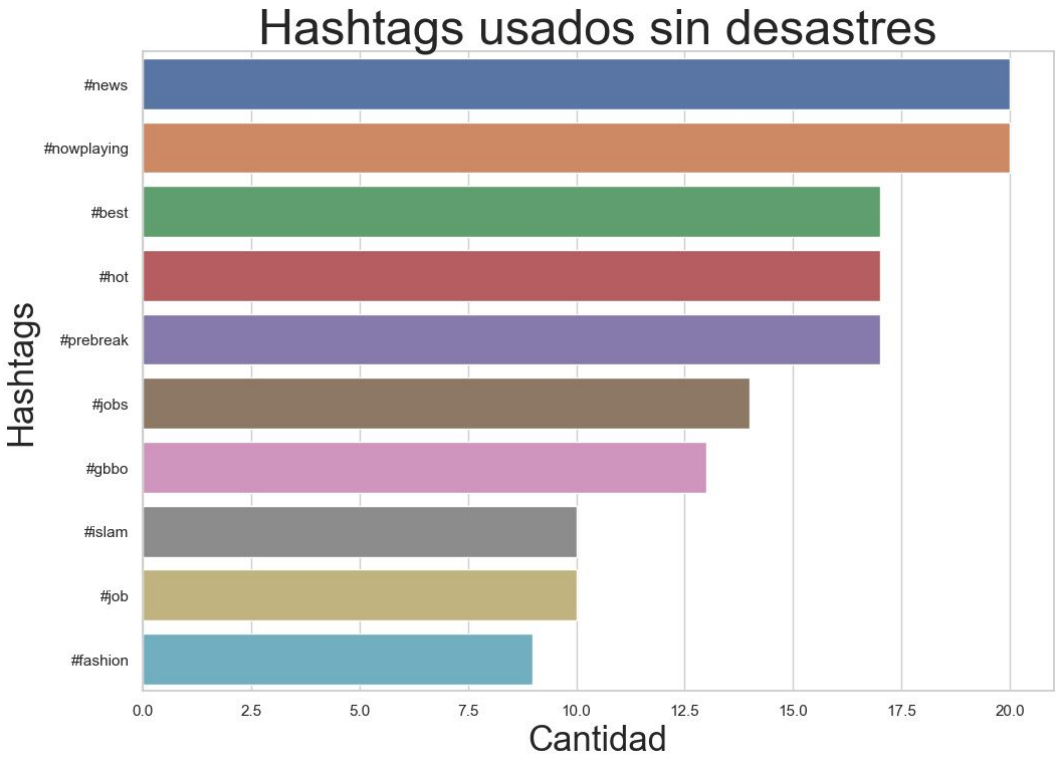
Podemos ver que por lejos el más usado es “news” lo cual hace mención a la intención de informar una noticia, pero además tenemos signos de interrogación que nos hace alusión a que ha pasado algo desconocido, así también vemos algunos nombres de lugares, y de fenómenos lo cual es entendible para aquellos tweets que buscan informar algún tipo de desastre.

6.3 - ¿Cómo se presentan los hashtags que están atribuidos a desastres?



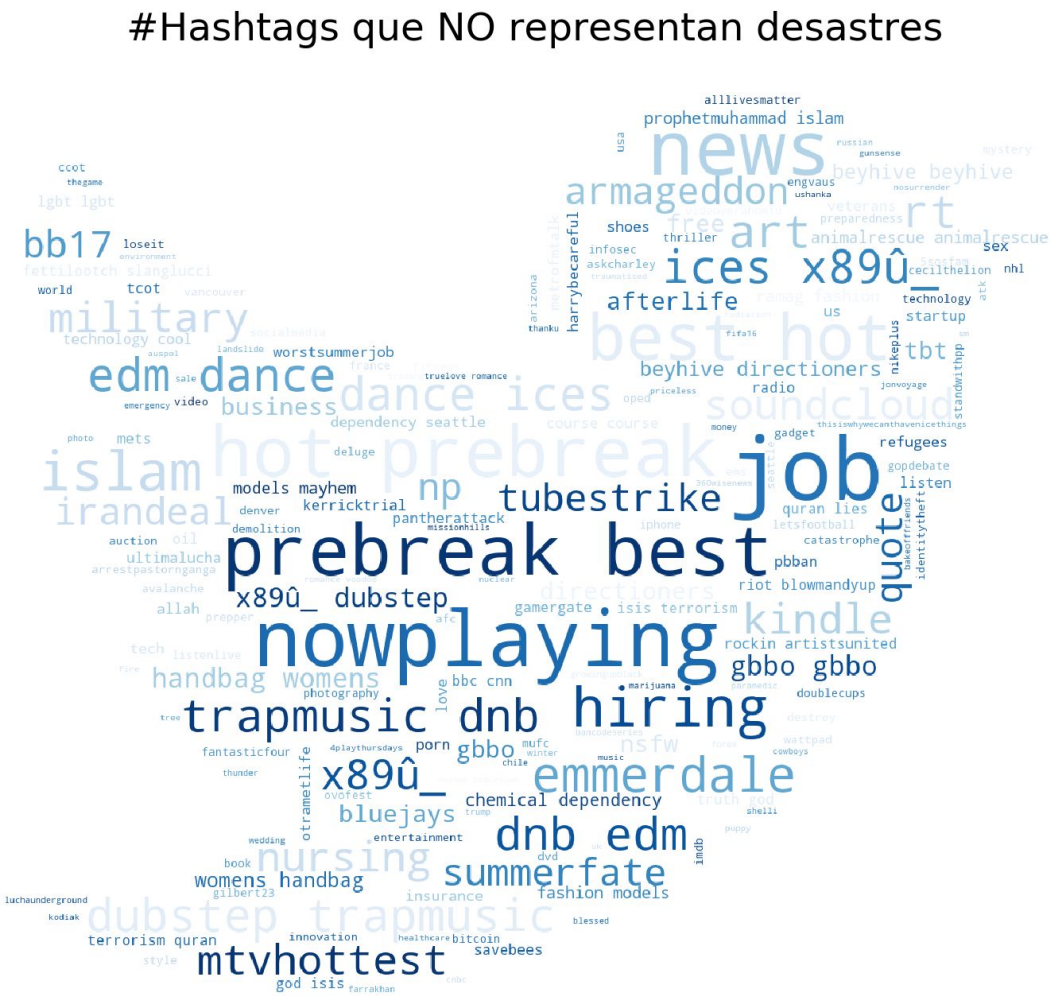
En la nube anterior podemos ver que para los tweets que representan desastres la mayoría de los hashtags que se usan y que tienen un mayor peso son aquellos que indican alguna de las siguientes tres alternativas, una noticia, un desastre o un lugar, aunque también encontramos algunos términos que no tienen un significado tan notable pero son los menos, en su mayoría se aplica lo analizado anteriormente.

6.4 - ¿Cuáles son los hashtags más usados para tweets que NO representan desastres?



Podemos ver que los tweets que no representan desastres tienen hashtags mucho más parejos en su cantidad de apariciones, así también el tipo de hashtags varía bastante desde novedades, búsquedas de trabajo, moda, videojuegos, lo que nos permite entender que los hashtags utilizados suelen estar muy relacionados con lo que se busca expresar en el tweet o lo que se busca informar.

6.5 - ¿Cómo se presentan los hashtags que NO están atribuidos a desastres?



En la nube anterior podemos ver que para los tweets que no representan desastres, los hashtags están mucho más dispersos en su temática y no es fácil distinguir un tema que los englobe. Podemos ir desde videojuegos y novedades a trabajo. También su peso es mucho más parejo en relación a otras palabras, no es que haya una o dos muy predominantes como es el caso anterior.

7 - Conclusiones

A lo largo de este trabajo, hemos estudiado diversos aspectos del set de datos proporcionado por Figure-Eight con la intención de identificar patrones o características que nos den una pista sobre como identificar aquellos tweets que representan un desastre o no. Para esto, en este informe, fuimos pasando por varias secciones analizando por ejemplo el tamaño, el idioma, las palabras claves, los hashtags y así poder obtener una visión más detallada de la situación.

Es bueno entender que los Tweets están repartidos bastante equitativamente entre aquellos que representan un desastre y aquellos que no, lo cual nos permite analizar por separado estas condiciones y así comparar los resultados obtenidos a sabiendas de que las cantidades representadas son similares.

Por otro lado la ubicación de los tweets no representa una información que nos pueda aportar un salto diferencial en el análisis, sino que éste se termina enfocando en el análisis del texto o más bien del contenido del tweet en sí mismo, donde la idea es reconocer en base a ese texto si representa o no un desastre.

Sobre los tweets y su tamaño vemos que la distribución entre aquellos que representan un desastre y los que no es bastante pareja cosa que no pasa con la longitud, donde los que representan un desastre suelen tener mayor cantidad de caracteres mientras que los otros tienen un tamaño mucho más variable. Ahora bien, si hacemos una distinción por palabras, vemos que en la mayoría de los casos van entre quince y veinticinco palabras pero manteniendo la misma lógica, es decir que los tweets que representan desastres suelen tener mayor contenido. Por otro lado y como dato de color, encontramos que el lenguaje de los tweets ha sido enteramente en inglés, cosa que nos facilita el análisis, mientras que para la ubicación tenemos algo totalmente distinto: son datos muy variados que requieren mucho trabajo para poder hacer una unificación, al menos conceptual, y no parece brindar información reveladora.

Respecto a las palabras que forman los tweets, realizamos un estudio utilizando una biblioteca de procesamiento de lenguaje natural. Encontramos que no existe una sola

palabra que pueda determinar fehacientemente si estamos ante la presencia de un tweet que hace referencia a un desastre, pero sí que existe una lista de palabras que nos puede hacer llamar la atención para analizar el registro con técnicas más avanzadas, como un análisis contextual, de *mood*, o bien los valores de los datos presentes en el resto de las columnas del dataframe, como por ejemplo los hashtags o keywords.

En relación a los keywords asignados a los tweets, notamos primeramente que más del 99% de los tweets tienen asociado un keyword, lo que nos permite hacer un buen análisis al respecto. Si nos adentramos en cada uno, notamos que muchos de esos keywords (sin diferenciar si se trata de un desastre real o no), son palabras que no hacen referencia a un tipo de desastre en particular, sino que son términos más generales, lo que nos hace difícil entender a que se refiere; aún así pudimos obtener algunas conclusiones, sobre todo con su combinación con otros parámetros, como por ejemplo los hashtags y targets, que nos indica una clara combinación entre ellos cuando se trata de algo real y una dispersión grande cuando no lo son.

Finalmente, en cuanto a los hashtags utilizados en los tweets, percibimos que solo un cuarto de estos tienen al menos un hashtag dentro de su contenido, lo que nos da a entender que no es tan prioritario su uso en relación a la cantidad de tweets total, pero sí podemos confirmar que en los casos en que se haya decidido emplear un hashtag, este suele marcar parte de lo que se quiere informar. A modo de ejemplo, hemos encontrado que para los desastres se ven hashtags referentes a noticias, lugares o desastres, mientras que para los casos que no representan desastres vemos cosas mucho más variadas como ser la búsqueda de trabajo, videojuegos, música y demás. Queda entonces bastante claro que a pesar de no ser tan usados, tienen mucha relevancia a la hora de guiarnos sobre el contenido de los tweets.

A modo de cierre, podemos decir que el análisis realizado nos ha permitido entender la estructura del set de datos y sus principales características, así como también prepararnos para encarar la participación en la competencia objeto del segundo trabajo práctico.

Asimismo, nos gustaría agregar que el estudio de un set de datos de estas carácteristicas, podría utilizarse, mediante el monitoreo de Twitter y el uso de machine learning, para detectar o enterarnos de noticias vinculadas a desastres prácticamente en tiempo real, con un alto grado de certeza.