FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico 1
Análisis exploratorio de datos

[75.06/95.58] Organización de Datos
Primer cuatrimestre de 2020
Grupo Cloud

Nombre	Padrón	E-mail
Davérède, Agustin	98.540	agusdaverede@yahoo.com.ar
Garófalo, Nicolás	100.952	nico.garofalo98@gmail.com
Chogri, Ramiro	100.499	rchogri@fi.uba.ar
Ortiz, Benjamín	100.585	benjaortiz969@gmail.com

Repositorio: https://github.com/NicoGarofalo/TPDatos

Índice

1. Resumen

El presente informe fue desarrollado con el objetivo de mostrar los resultados obtenidos del análisis exploratorio del set de datos de la competencia de Kaggle, bajo el nombre de Real or Not? NLP with Disaster Tweets". En primer lugar, se realizó el análisis individual, por columna, de los datos con los que se contaba, para luego considerar las interdependencias e interrelaciones que pudieran existir entre los mismos. Se intentó aplicar el ingenio y la creatividad a la hora de analizar los datos, a la vez que se buscó incorporar los criterios y conocimientos cientificos que nos ha aportado hasta ahora la asignatura.

1.1. Utilización de recursos

El análisis de los datos se efectuó haciendo uso del lenguaje *Python* y de las siguiente bibliotecas:

- Pandas
- Matplotlib
- Numpy
- Seaborn
- WordCloud

2. Análisis Introductorio

Se comenzó realizando los imports con los que trabajó: **Pandas**, para realizar el análisis de datos, y **matplotlib** y *seaborn* para realizar el plot de los gráficos. Luego, se efectuó la lectura de train.csv, creando un *DataFrame* de Pandas con todos los datos del set.

Se llevó a cabo una inspección general de los datos, lo que incluye realizar un *.head()* para observar la apariencia del *DataFrame*, y *.shape()* para saber la cantidad de registros totales. De esta primera inspección se obtuvo que el set de datos cuenta con las siguientes columnas o categorías principales:

- **id**: Valor numérico único de identificación del tweet.
- **Keyword**: Palabra clave. Se empleará para categorizar a los tweets.
- **Location**: Ubicación desde la cual se realizó el tweet.
- **Text**: Texto del tweet.
- **Target**: Vale 1 si la emergencia sucedió, 0 en caso contrario.

Se prosiguió definiendo los tipos de cada columna, donde **id** pasará a ser de tipo *int*, **target** será *bool*, y los demás serán *strings*.

Para la limpieza de datos se eliminarán los duplicados, en caso de que existan. Al menos al comienzo, se decidió conservar los registros que no continen información debido a que, principalmente en la columna *location*, hay un gran porcentaje de éstos que contienen *NaN*, por lo que se perderían muchos registros. De hecho, sólo un 66,73 % de los registros poseen una *location*.

2.1. Conteos destacables

Durante el análisis de introducción se observaron los siguientes resultados parciales:

- **Total Registros**: 7613
- **Total Keywords que tienen al menos un tweet falso**: 219
- **Total Keywords que tienen al menos un tweet verdadero**: 221
- **Total unique Keywords**: 222
- **Total unique Location**: 3342
- **Porcentaje de datos no nulos para id, text y target**: 100 %
- **Porcentaje de datos no nulos para keyword**: 99.19 %
- **Porcentaje de datos no nulos para location**: 66.73 %

3. Análisis General

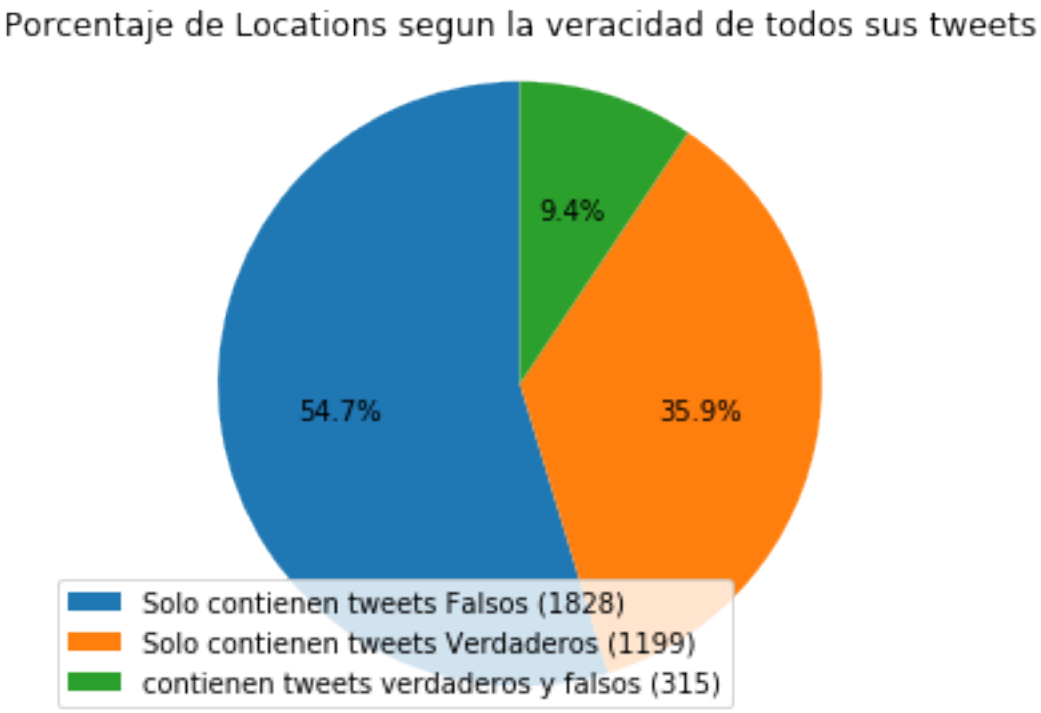
3.1. Análisis en base a las Locations

En un principio se planteó lograr un diccionario de locations en base a países, con el objetivo de luego visualizar, por país, cierta *keyword* destacable en base a sus ocurrencias, o cuál es el país que tweetea más tragedias falsas o verdaderas. Sin embargo, al trabajar con los diferentes valores que tomaban las *locations*, se comprendió que resultaría casi imposible poder incluir algunas de estas locations como parte de otras, como por ejemplo la *location* 'New York' en 'USA'. La dificultad residía en que se necesitaría alguna variable que contenga, por ejemplo, todos los estados o ciudades existentes en USA para utilizarla de guía, lo cual es complejo de lograr. También, se dieron casos donde las *locations* eran universidades o localidades de un país específico, por lo que complejizaría el proceso de armado de un diccionario.

Tras analizar los registros cuyas *location* fuesen distinto de *Null*, se observó que las *keywords* asociadas a los registros correspondientes también resultaron ser distinto de *Null*.

Se buscó realizar el conteo de cuántas *locations* twittea solamente textos verdaderos, o falsos. Para lograrlo, se realizó una agrupación de tweets totales por *location*, para luego contar cuántos de estos eran falsos. Por lo que, si coincide con el total de tweets de la *location*, éste solo tiene tweets falsos. Luego, se realizó lo mismo para los tweets verdaderos, por lo que podemos obtener un Top 5 *locations* con sólo tweets verdaderos, e ídem falsos (El top se hace en base a cantidad de tweets).
Luego de contabilizar cuántas *locations* cumplen con estos requisitos, se obtuvo un *Pie Chart* que expresa de modo porcentual lo analizado.

3.1.1. Porcentaje de *Locations* según la veracidad de todos sus tweets



Pie Chart: Porcentaje de *Locations* según la veracidad de todos sus tweets

3.1.2. ¿Por qué *Pie Chart*?

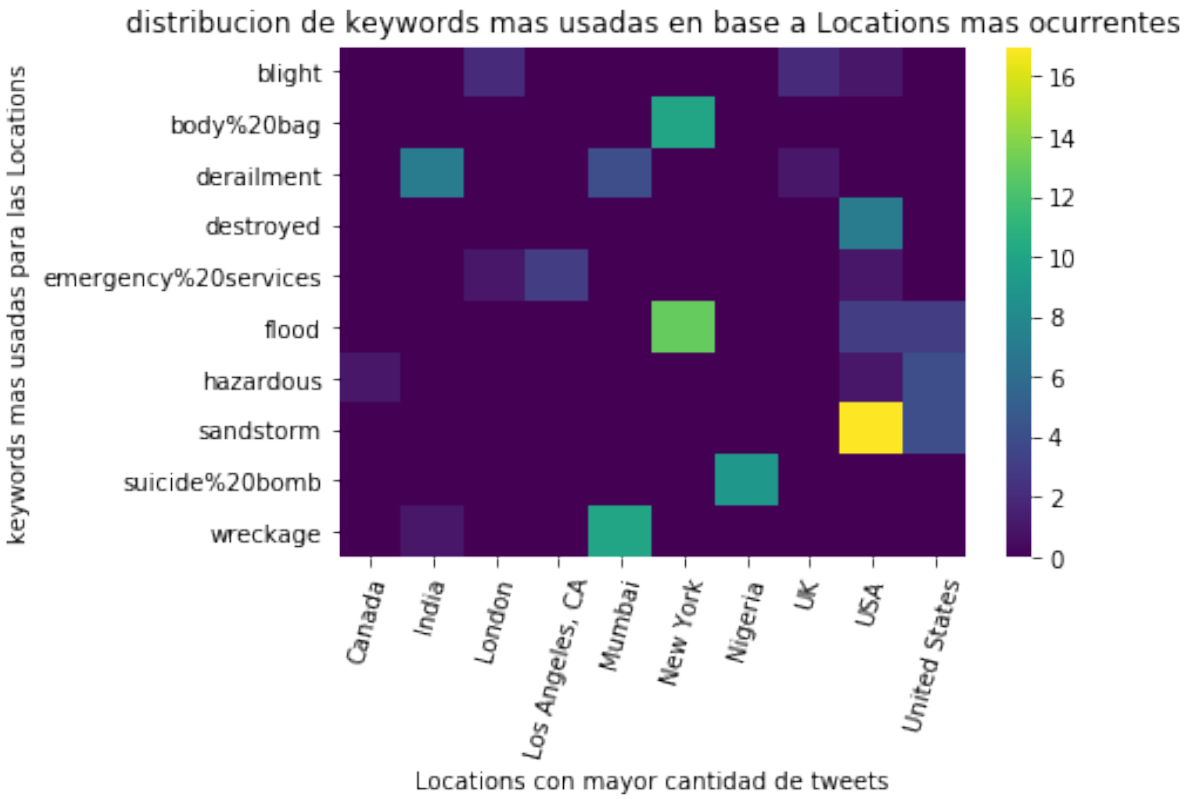
Si bien esta visualización no es de las favoritas a utilizar, se la consideró efectiva para la idea que se buscaba transmitir: la **diferencia de porcentajes** entre las *locations* que tienen sólo tweets falsos, sólo tweets verdaderos, y los que tienen tanto verdaderos como falsos. Además, al no particionar en exceso el *Pie Chart*, resulta simple de interpretar.

3.1.3. Análisis de keywords para las Locations más recurrentes

Al haber observado las *locations* mas recurrentes en el *dataset*, surgió una pregunta: ¿Habrá alguna relación entre las *keywords* correspondientes a las *locations* más recurrentes y estas mismas *locations*?

Para lograr responder dicha pregunta, se filtró el *dataframe* original, eliminando cualquier registro cuya *location* fuese distinta de las 10 más usadas. Una vez obtenidos los registros correspondientes a estas 10 *locations*, se obtuvieron las 10 *keywords* más recurrentes y se filtró nuevamente, dejando sólo los registros cuya *location* y *keywords* estuviesen dentro de dichos 'top 10'. Luego de una re-organizacion de los datos, se obtuvo el *heatmap* que se muestra en la siguiente sección.

3.1.4. *Heatmap*: distribución de keywords más usadas para las *locations* más recurrentes



Heatmap: distribución de las *keywords* mas usadas en base a las *locations* mas recurrentes

3.1.5. ¿Por qué *Heatmap*?

El heatmap parecía ser la mejor opción para visualizar de manera instantánea la posible relación entre dos variables categóricas, tales como lo son las 10 *locations* más recurrentes y las 10 *keywords* más usadas dentro de dichas *locations*.

3.1.6. Conclusión del *Heatmap*

Observando el *heatmap*, se logra concluir que no hay ninguna relación clara entre las *keywords* más usadas para las *locations* más recurrentes. Cabe destacar que ninguna de estas *keywords* está presente en más de 3 de las *locations*. Desde el punto de vista de las *locations* es destacable el hecho de que, a excepción de 'USA', las demás no cuentan con un gran número de *keywords* "populares" sino que tienden a tener 2 *keywords*.

3.2. Análisis en base a la cantidad de caracteres del tweet

Esto surgió del cuestionamiento que se hizo el equipo sobre si existía una posible relación entre el largo de los tweets y las *keywords*. Podría suceder que para *keywords* como *derailment o suicide bomb*, la cantidad de caracteres de los tweets fuera más grande para poder dar mayor detalle de la tragedia sucedida, y para *keywords* como *screaming o fear*, ser más cortos debido a que podría tratarse de un hecho no trágico. De hecho, se han encontrado tweets de tamaño pequeño en los que el usuario simplemente comenta algo de poca importancia.
Para esto, se decidió calcular el promedio del largo de los tweets, el máximo y el mínimo. Establecer 3 cotas diferentes:

- El propio promedio
- El valor medio entre el promedio y el máximo
- El valor medio entre promedio y el mínimo

Luego, se contabilizó cuántos tweets estaban por debajo de cada cota para cada *keyword*. Los resultados obtenidos no fueron los esperados, ya que no se encontró un posible patrón del que se pueda definir un comportamiento específico.

3.2.1. Cantidad de caracteres para tweets sin location ni keyword

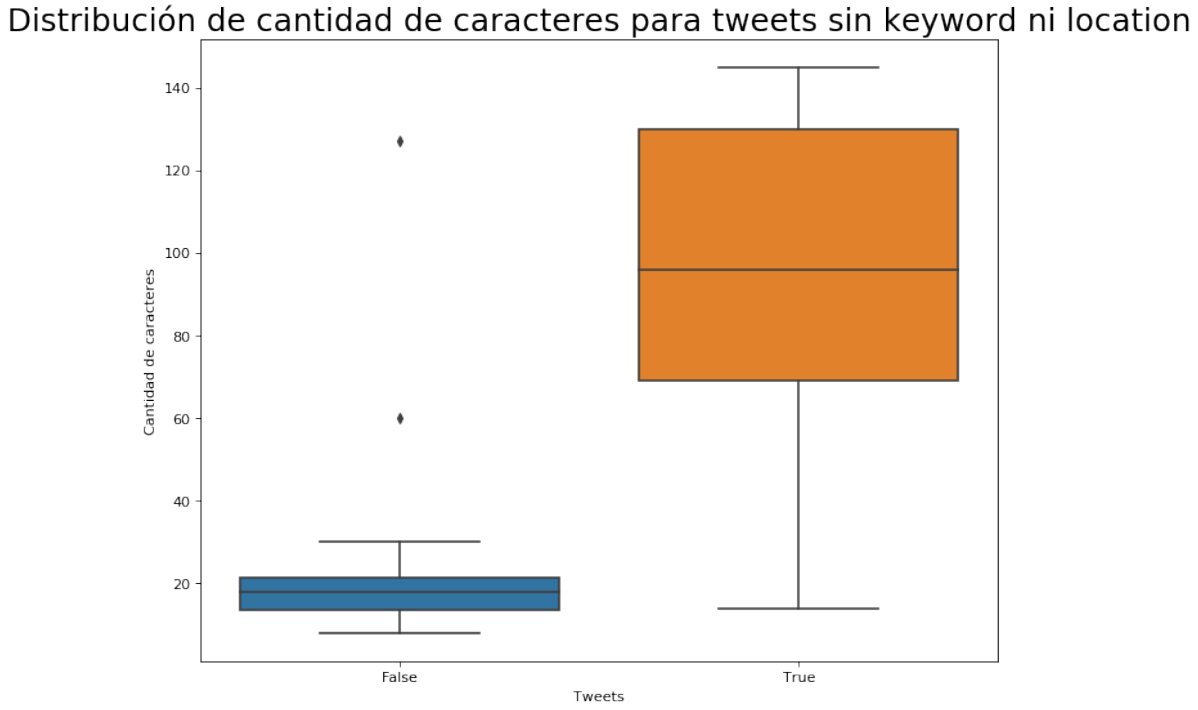
Por otro lado, se calculó también el promedio de los tweets verdaderos que no tienen ni *keyword* ni *location* asignada, lo mismo para los tweets falsos. Los resultados obtenidos son destacables: para los tweeets verdaderos, el promedio es de **96** caracteres, mientras que para los falsos, **25** caracteres. Esto podría suceder debido a lo que se buscaba obtener al momento de hallar una relación entre keyword y el largo del tweet: Los tweets verdaderos utilizan más caracteres para informar acerca de una tragedia, mientras que los tweets falsos sin *keyword* ni *location* suelen ser tweets de usuarios que ni siquiera aportan una tragedia. Para ejemplificar un poco, algunos de los tweets falsos que no aportan una tragedia serían los siguientes: *'My car is so fast'*, *'I love fruits'*, *'What a wonderful day!'*, entre otros.

Se decidió realizar una visualización donde se observó la distribución de la cantidad de caracteres para estos tweets que no cuentan con *keywords* ni *locations*, de manera que se logre clarificar lo mencionado anteriormente. Por esta razón, se decidió realizar un **Boxplot**.

3.2.2. ¿Por qué Boxplot?

Se decidió utilizar un *boxplot* ya que es una de las mejores opciones al momento de poder visualizar la variación de la cantidad de caracteres, logrando también observar el promedio y los casos particulares de ambas.

3.2.3. Boxplot: Distribución de caracteres para tweets sin keyword ni location



Boxplot: Distribución de caracteres para tweets sin *keyword* ni *location*

3.3. Análisis en base a las keywords

3.3.1. Keywords más utilizadas

Al observar las distintas keywords, una de las primeras interrogantes que se plantearon fue '¿Cuáles son las keywords más utilizadas u ocurrentes?' '¿Qué pasa si filtramos por tweets verdaderos y falsos?'. Es natural pensar en que hay que realizar un conteo para saber cuáles son las más ocurrentes, por lo que se necesitaba una visualización que resalte justamente las ocurrencias. Por esta razón, se decidió utilizar **Wordclouds**.

3.3.2. ¿Por qué Wordclouds?

Debido a que se podrían expresar las diferentes *keywords* en base a su ocurrencia (cantidad de apariciones), tanto para los tweets verdaderos como para los falsos. Este tipo de visualización se especializa en enfatizar, aumentando su tamaño, aquellas palabras que tienen mayor aparición para mezclarlas entre las demás, para asi obtener una noción de importancia entre sus pares.

3.3.3. Wordcloud de keywords más utilizadas en tweets verdaderos

En esta visualización se destacan, principalmente, keywords como *wreckage, derailment y outbreak* como las más ocurrentes entre los tweets verdaderos. También, se puede observar que, en menor medida, que aparecen *keywords* de hechos trágicos un poco menos frecuentes (debido al carácter del mismo), tales como *suicide bombing*, *airplane accident*.

Wordcloud de *keywords* más ocurrentes para Tweets verdaderos

3.3.4. Wordcloud de keywords más utilizadas en tweets falsos

Esta vez, las *keywords* que más se observan son *body bags, armageddon, harm*, entre otros. Se pueden observar otras poco descriptivas, como *ruin, fear, harm, panic*, entre otras.

Wordcloud de *keywords* más ocurrentes para Tweets falsos

3.3.5. Conclusión Wordclouds

En base a los resultados obtenidos en ambos Wordclouds, se puede observar que las *keywords* con más ocurrencias en los tweets verdaderos suelen ser palabras con un aspecto más realista, es decir, tragedias descriptas por *keywords* precisas. También, existen excepciones a la regla, tales como *debris, wreckage*. Como ejemplo de keywords precisas podemos mencionar *typhoon, oil spill o derailment*, esta última haciendo referencia a tragedias de descarrilamiento de trenes. Éstos son hechos reales y concretos que pueden suceder en cualquier parte del mundo. Por su contraparte, en los tweets falsos se observan *keywords* de índole más abstracta, ambigüas o "fantasiosas", por ejemplo *armageddon, hellfire* o *fear*. Si se quisiera dar un ejemplo más concreto, se podría seleccionar un tweet del set de datos que esté categorizado dentro de "fear", simplemente porque ocurrió un hecho que puede generar miedo en alguien, pero no necesariamente tuvo que ocurrir una tragedia para que se generara tal sentimiento. Sin embargo, eso no quita que haya casos donde haya ocurrido realmente una tragedia, y estén categorizadas bajo un *keyword* abstracto.

3.3.6. Ocurrencias de keywords de tweets verdaderos versus tweets falsos

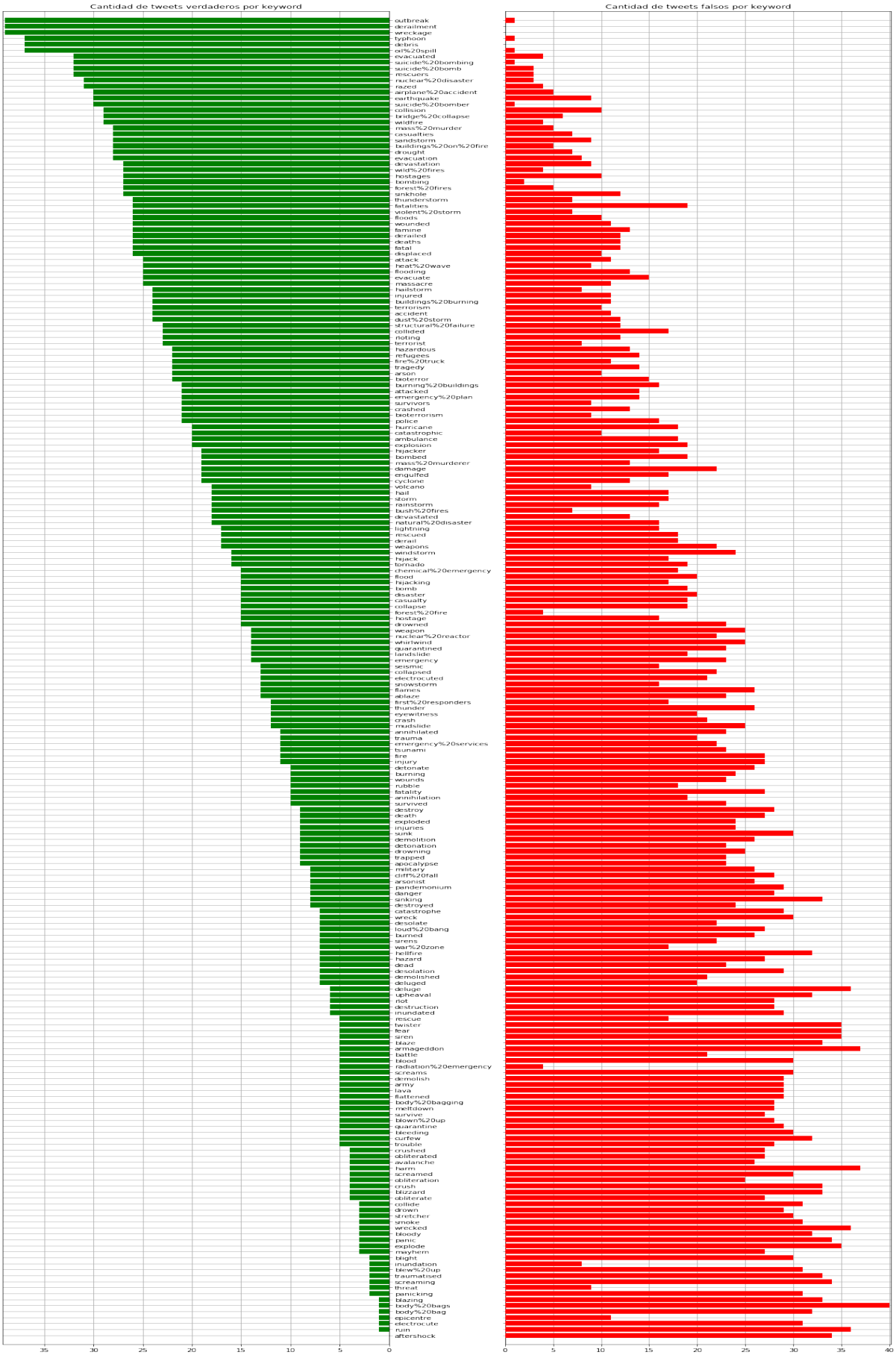
Una vez analizado los wordclouds, surgió otra interrogante: '¿Qué pasa si comparamos las ocurrencias de cada keyword para los tweets verdaderos con los tweets falsos? ¿Tendrán alguna relación?'. Se tenían las ocurrencias de cada *keyword* por separado ya que fue lo utilizado para graficar los *Wordclouds*, por lo que era necesario agruparla en una sola estructura. Una vez realizado, había que graficar esa estructura, lo que llevó a considerar utilizar un **Pyramid Barchart**.

3.3.7. ¿Por qué Pyramid Barchart?

Con el propósito de presentar un "versus"de las ocurrencias de las *keywords* entre los tweets falsos y verdaderos. Surgió la idea de un Pyramid barchart luego de haber descartado los *Stacked Barchart* y *Grouped Barchart*, debido a que la cantidad de *keywords* presentes eran demasiadas como para lograr una visualización limpia.

3.3.8. Visualización Pyramid Barchart

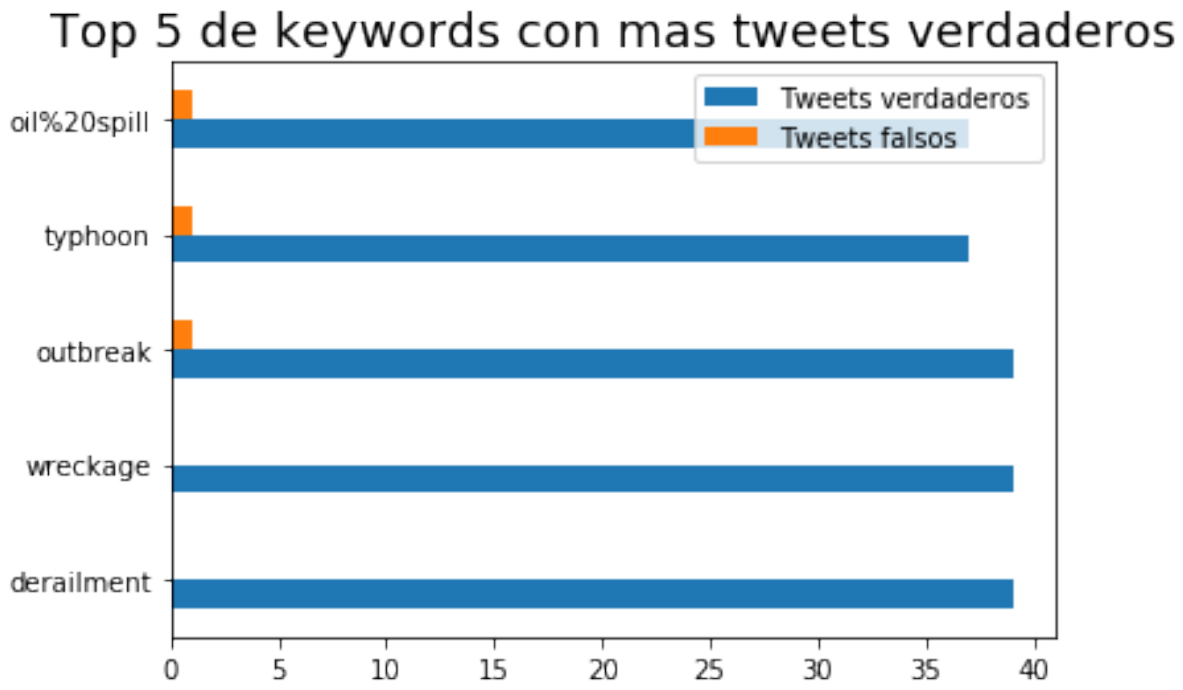
Si bien el gráfico presenta muchísima información, tomamos la decisión de presentarlo de todas formas debido a que queremos enfatizar el **crecimiento gradual** de las ocurrencias falsas cuando las verdaderas decrecen, y viceversa.



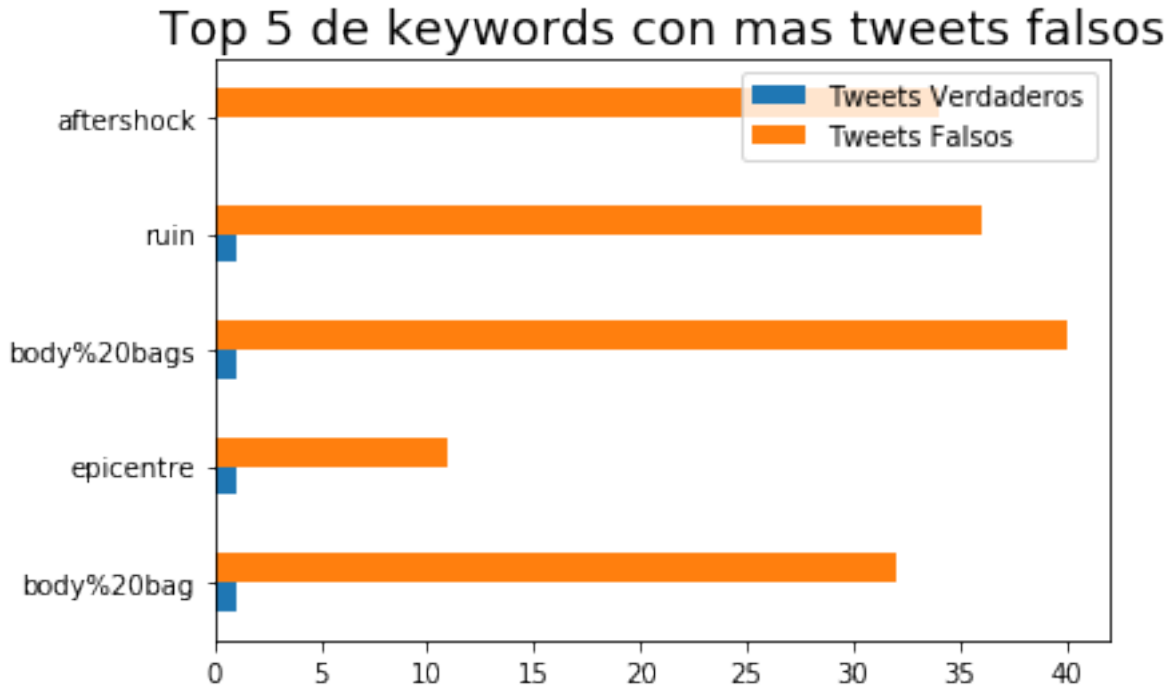
Pyramid Barchart: Ocurrencias de keywords verdaderos vs Ocurrencias de keywords falsos

3.3.9. Observaciones adicionales

Se agregaron dos *Grouped Barcharts* para enfatizar la diferencia de ocurrencias en los extremos de la tabla anterior.



Grouped barchart: Top 5 *keywords* con mayor ocurrencias verdaderas



Grouped barchart: Últimas 5 *keywords* con menor ocurrencias verdaderas

3.3.10. Conclusiones Pyramid Barchart

Como conclusión de esta visualización, como bien se mencionó, es posible observar que durante el decrecimiento de las ocurrencias verdaderas de las *keywords*, las ocurrencias falsas presentan un crecimiento (existiendo excepciones). Esto puede suceder debido a lo mencionado en las conclusiones de *Wordclouds*: Las *keywords* con mayor ocurrencia verdadera son más descriptivas y reales, y a medida que la cantidad de ocurrencias verdaderas disminuyen, las *keywords* pasan a ser de carácter más abstracto. Aumentan las ocurrencias falsas, ya que en éstas es donde residen la mayor cantidad de tweets falsos.

3.4. Análisis del criterio de selección de keyword

En el momento que se realizaron los analisis, una interrogante más que se planteó fue ¿Existira un criterio a la hora de asignar las *keyword* a los tweets?¿Habra sido automatizado?. Para esto se consideró analizar si los tweets que tienen una *keyword* asignada, la contienen como parte de su texto.

3.4.1. Datos para el analisis

Una vez analizado lo planteado, se observó que ninguna *keyword* está contenida en todas las ocurrencias de los tweets que tiene asociados. El caso opuesto sí se puede observar para algunas *keywords*: Existen keywords que no estan contenidas en ninguno de sus tweets asociados. Se trató de hacer una visualización de estos datos pero no fue posible lograrlo, debido a que se requería de una explicación clarificadora , pero de todas formas se pensó interesante mostrar los datos sin visualizacion. Se hizo un recuento por *keyword* de cuantos tweets la contenian y cuantos no, y se llego a los siguientes datos:

- Hay 143 *keywords* que, en mas de la mitad de sus tweets asociados, estas estan contenidas dentro del texto
- Hay 75 *keywords* que, en menos de la mitad de sus tweets asociados, estas estan contenidas dentro del texto
- Hay 3 *keywords* que estan contenidas dentro del texto de sus correspondientes tweets en exactamente la mitad de ellos

3.4.2. Conclusión del criterio de seleccion de keywords

Luego de analizar los datos se llegó a la conclusion de que los tweets que tienen *keyword* pero no se menciona en el mismo tweet no pudieron ser etiquetados con la *keyword* por un programa automatizado, ya que esos tweets se refieren a cosas en relacion con la *keyword* y no debe ser posible automatizar ese proceso.

3.5. Análisis de tweets duplicados

Realizando el análisis anterior, por accidente se llegó a una *keyword* que contiene muchos tweets basicamente iguales. Esta es la *keyword* 'sandstorm', que a diferencia de un link en el tweet el resto era identico. Esto sembró la duda sobre si los duplicados afectan las estadísticas. Se logró liberar el *dataframe* de 87 tweets duplicados con un 'drop-duplicates()', pero al revisar la *keyword* 'sandstorm' nuevamente se llegó a la conclusión de que lo que impidió que se liberarán esos duplicados era la dirección del link compartido en el tweet.

3.5.1. Conclusión sobre tweets duplicados

Si bien los tweets analizados no eran exactamente iguales, debido al link desde el punto de vista de comparación caracter a caracter, cabe destacar que el caso 'sandstorm' es uno particular que genera algunas incognitas. ¿Sucederá esto con otras keywords? ¿Existirá alguna forma de eliminar los duplicados en estos casos de una forma automatizada?

3.6. Datos de interés

Tras realizar los análisis precedentes, se han podido hallar los siguientes datos de color:

- La única *keyword* que no tiene tweets verdaderos es *'Aftershock'*.
- Existen 3 keywords que no tienen tweets falsos: *'debris'*, *'wreckage'* y *'derailment'*.
- Si un tweet contiene una *location* distinta de *Null*, también tendrá una *keyword* distinta de *Null*.
- Existen tweets verdaderos un poco 'cuestionables'. Un ejemplo de estos es *'Bloody insomnia again! Grrrr!! #Insomnia'* (Tweet id: 1296).

4. Conclusiones Generales

Como se ha mencionado en las conclusiones de cada visualización, se puede confirmar que existe una relación entre las ocurrencias de tweets verdaderos respecto de los falsos: a mayor cantidad de una de éstas, disminuye la otra. La explicación que hemos encontrado ante este hecho es debido a la índole de las *keywords*. Las *keywords* consideradas mas 'abstractas', o que pueden ser asociadas con emociones (como *'fear'*), daban lugar a mas tweets falsos ya que el usuario podría enviar un tweet categorizado bajo esa *keyword*, sin haber ocurrido una tragedia. Por otro lado, se encuentran las *keywords* que usualmente se atribuyen a hechos más concretos, como *oil spill o derailment*, las cuales uno suele categorizarlas como tragedias.

También, se pudo confirmar, gracias al *boxplot*, que para aquellos tweets que no tienen *location* ni *keyword*, los tweets verdaderos contienen más caracteres que los falsos ya que estos últimos podrían no mencionar ninguna tragedia y ser un simple comentario de un usuario, mientras que los verdaderos utilizan más caracteres para poder describir la tragedia sucedida.

Se ha observado que existen casos en los que el tweet es verdadero y, sin embargo, al leer el contenido del *text* se contempló que no se está hablando de una tragedia sucedida, o bien que el hecho podría no considerarse como tragedia. Eso llevó a cuestionar, ¿Bajo qué criterio se ha definido en el set de datos denominar a un tweet como 'verdadero' (*target* = 1), o 'falso' (*target* = 0)? Cabe mencionar la interrogante propuesta en la sección 3.4: ¿Existirá un criterio a la hora de asignar las *keyword* a los tweets?
Estas fueron algunas de las incógnitas que aún no hemos encontrado una respuesta certera.

Para concluir, se puede considerar que el set de datos analizado contiene ciertas particularidades mencionadas a lo largo del presente informe. Se han planteado interrogantes, se han logrado responder una gran cantidad de éstas, y han surgido otras cuya respuesta aún no ha sido encontrada. Gracias a las visualizaciones se ha podido lograr reflejar la mayor parte de los datos obtenidos y responder a nuestras interrogantes iniciales.