FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Organización de Datos (75.06)

Análisis Exploratorio

Primer Cuatrimestre 2020
Grupo:
Integrantes:

Alumno	Padron	Email
Kevin Mendoza	98038	kmendoza@fi.uba.ar
Sebastián Agustín Rinaldi	94290	srinaldi@fi.uba.ar
Francisco Xavier Gauna	100563	fgauna@fi.uba.ar
Joaquin Parodi	100752	jparodi@fi.uba.ar

Link del Github:
https://github.com/FranciscoGauna/TP1-Datos

Índice

1. Introducción

En el presente informe nos proponemos analizar los datos provistos por el set de datos sobre tweets que se puede descargar desde https://www.kaggle.com/c/nlp-getting-started .
El objetivo principal es poder ver que información interesante podemos descubrir sobre estos tweets, ya sea que esa información que esté relacionada con la veracidad de los mismos o no. Para poder llevar a cabo esto, que es un informe sobre el análisis exploratorio de los datos, nos basaremos en un proceso iterativo de preguntas y respuestas.

2. Información de los datos

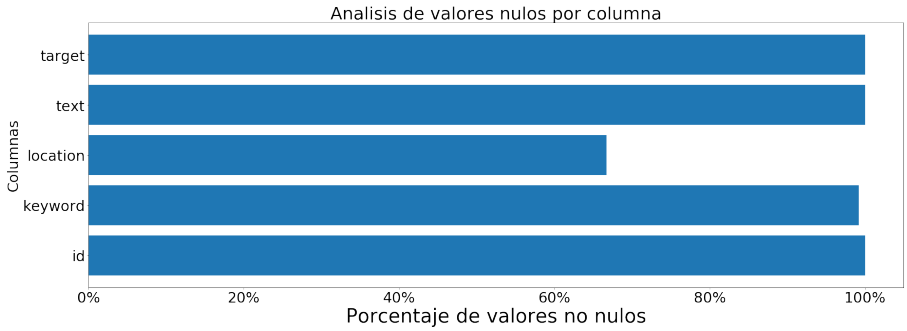
Con 7613 registros de 5 atributos cada uno, tenemos registros con valores nulos.
Los atributos son los siguientes:

- id: Identificador del tweet es único.
- text: El texto del tweet.
- location: La ubicación desde donde fue enviado.
- keyword: La palabra clave para el tweet
- target: Indica si un tweet habla de un desastre real(1) o no(0)

En el presente trabajo si decimos que un tweet es veraz, nos estaremos refiriendo a que se refiere a un desastre real. En caso contrario, nos referiremos a que no es un desastre real.

2.1. Limpieza de datos

Antes de comenzar a analizar los datos tenemos que ver si los registros que tenemos en el set de datos están completos. Una de las cosas a tener en cuenta es ver la cantidad de valores nulos que tiene el set.



Lo que se observa es que solo la columna **location** tiene un faltante de datos importante, por lo que a la hora de hacer análisis relacionados con los valores de esta columna, va a haber que tener en cuenta este dato. Por otro lado se puede ver que la en la columna **keyword** a pesar de no ser todos valores no nulos, el porcentaje de los nulos es muy muy bajo, por lo que según el análisis que se quiera hacer se determinará que hacer con esos registros.

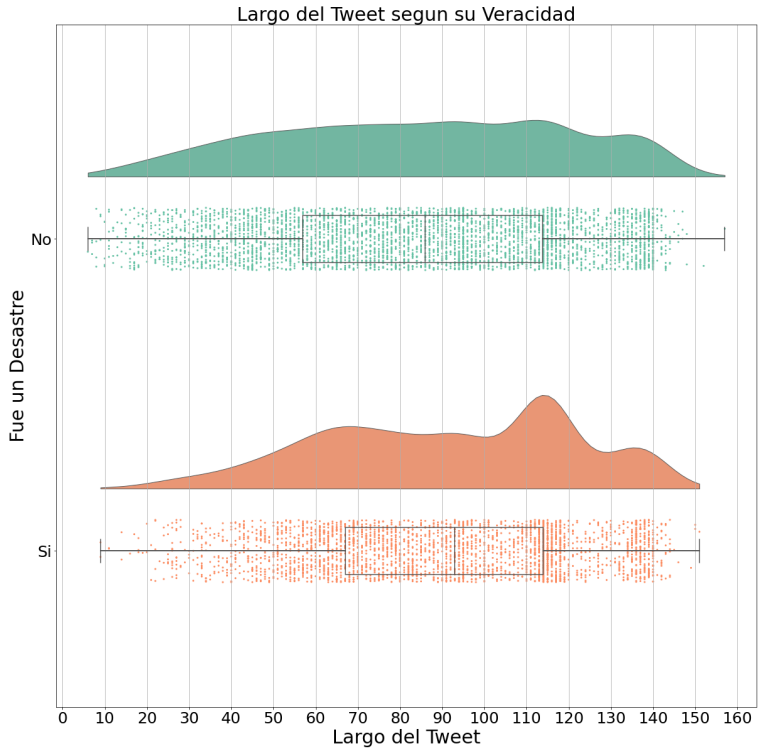
Justamente como algunos datos no nos llegaron completos, para poder subsanar esto tenemos diferentes métodos que podría ser los siguientes:

- Reemplazar el valor nulo por un cero o el valor medio de los datos.
- Eliminar la fila, columna donde se encuentra este dato.

Los métodos mencionados no son los únicos, pero son los que se utilizaran en este trabajo practico.

3. Análisis del Texto

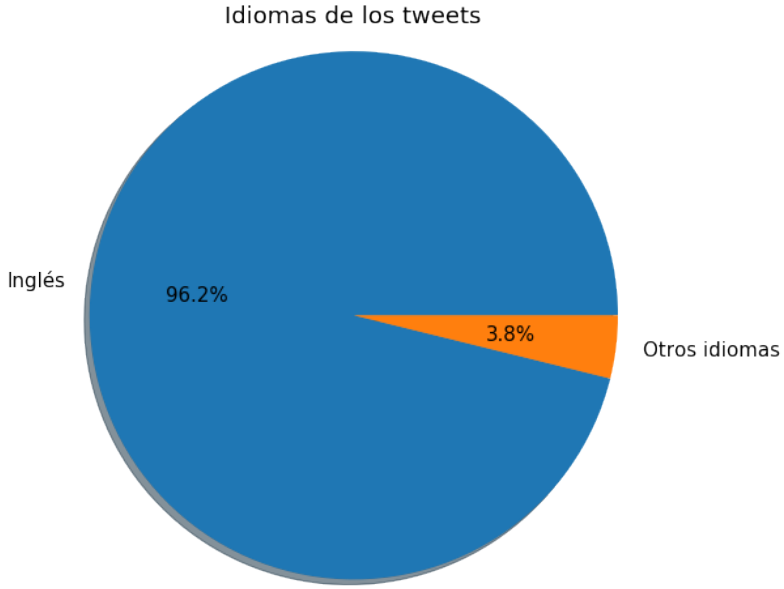
3.1. Largo del tweet según su veracidad



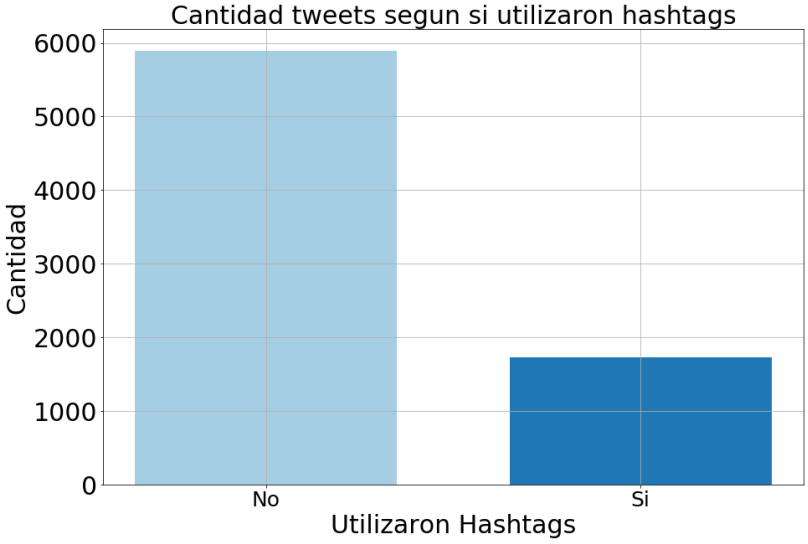
Observando el gráfico, los tweets que son desastre presentan una mayor longitud pero no parece ser una diferencia lo suficientemente significativa.

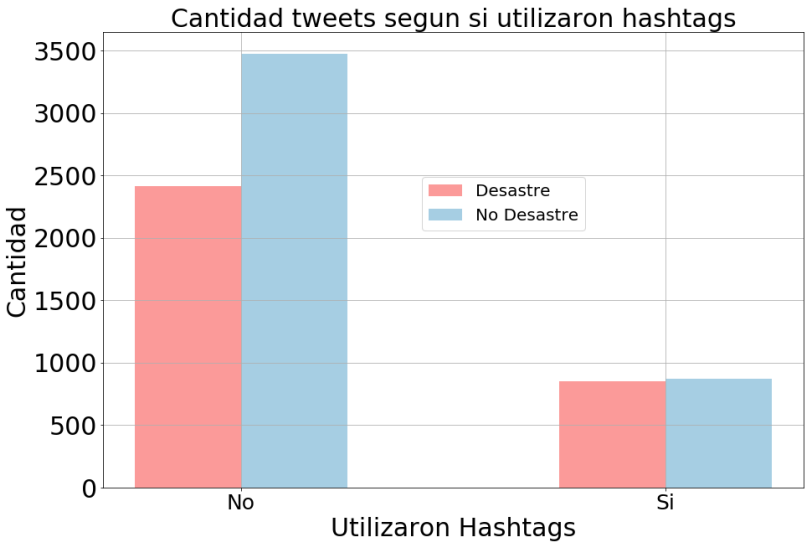
3.2. Idiomas de los tweets

Antes de empezar a analizar el texto de los tweets, nos gustaría saber cuantos tweets por idioma hay, si hay alguno que predomine, etc...



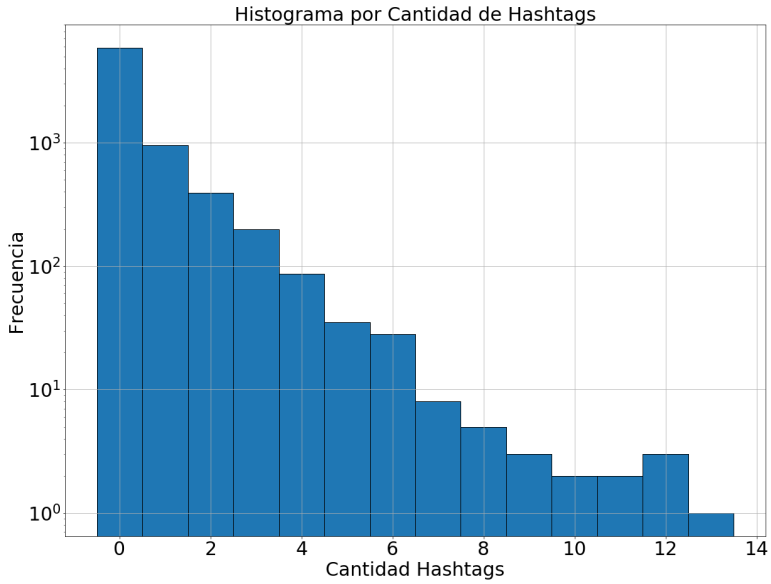
3.3. Cantidad de tweets que utilizan hashtags

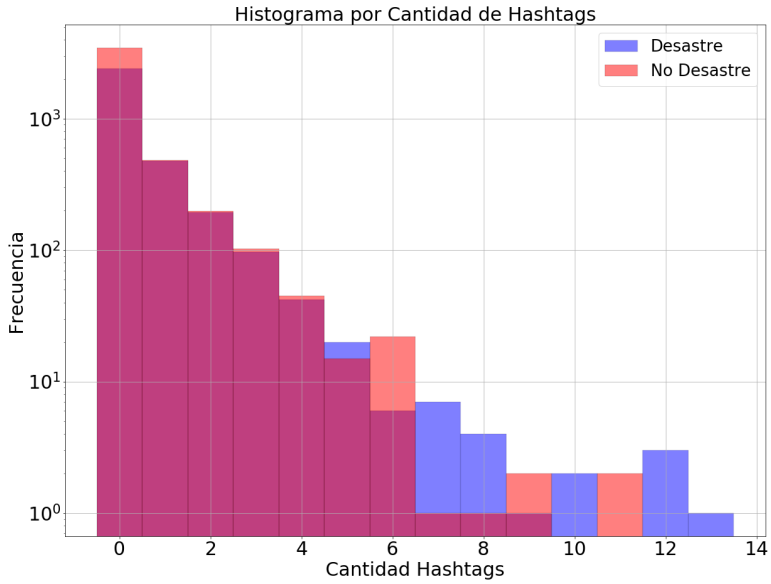




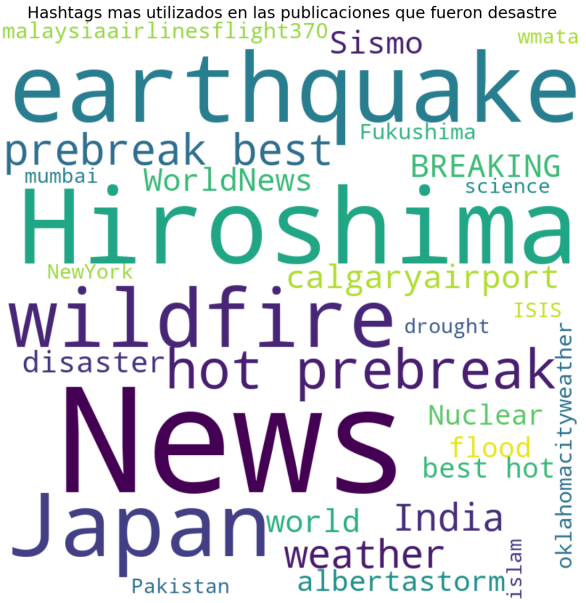
Sabemos que los hashtags se usan para indexar palabras claves que clasifican los contenidos publicados. Esto permite que haya mayor interacción entre el contenido y otros usuarios interesados en el tema. Siguiendo esto mas adelante veremos si en los hashtags tenemos la palabra clave que identifica el desastre.

3.4. Distribución de la cantidad de hashtags utilizados en tweets





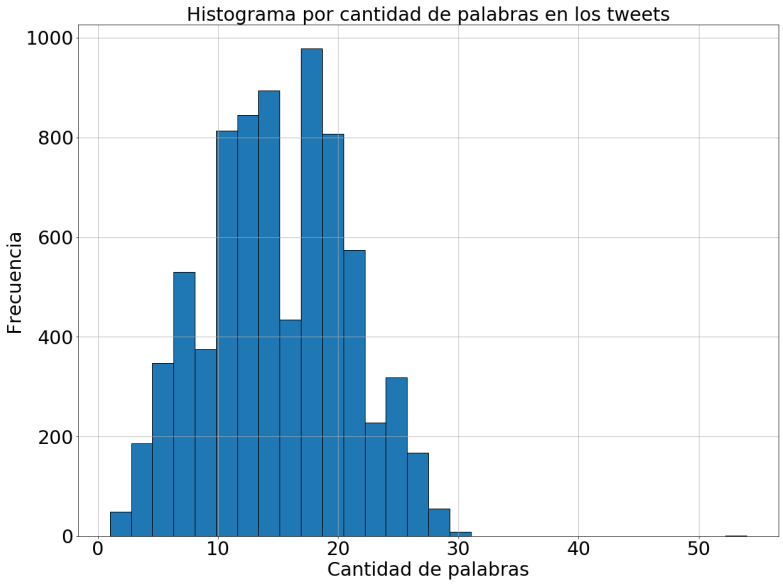
3.5. Hashtags mas populares

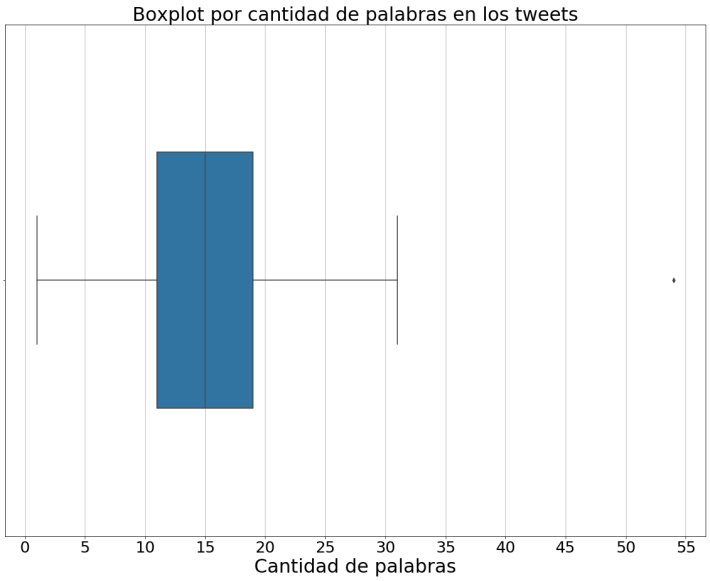


Hashtags mas utilizados en las publicaciones que no fueron desastre

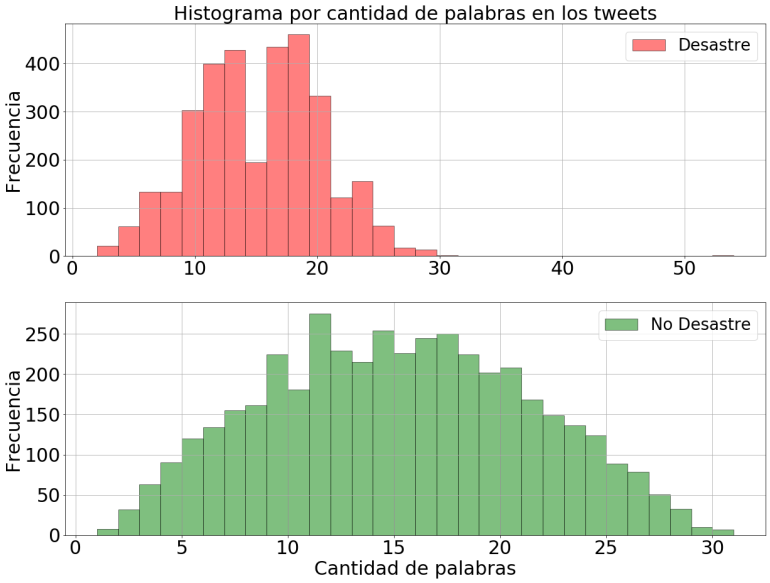
- Vemos que es los tweets que no fueron desastres hablan la mayoría de musica.
- En los tweets que fueron desastre hablan sobre Hiroshima,Japón seguro son publicaciones recordando aquellos hechos.
- Lo que es raro es que en tweets desastre o no están presentes hashtag como hot,prebreak,best.

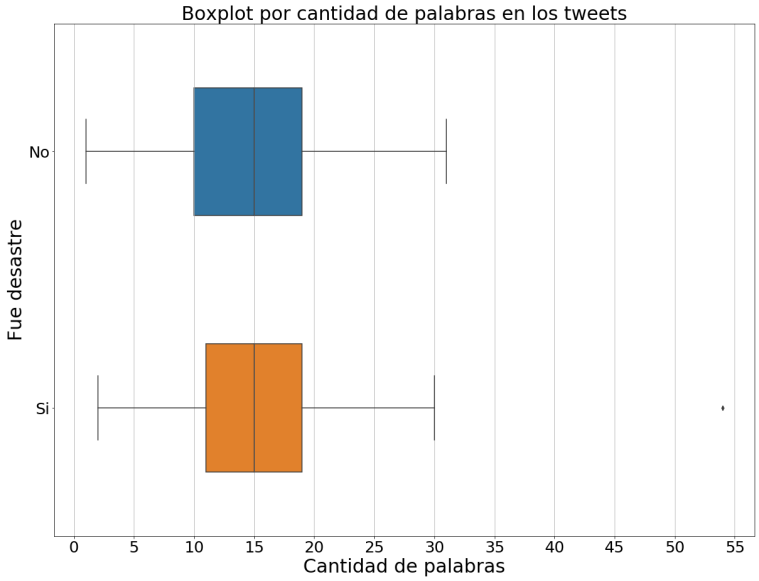
3.6. Distribución de la cantidad de palabras en los tweets





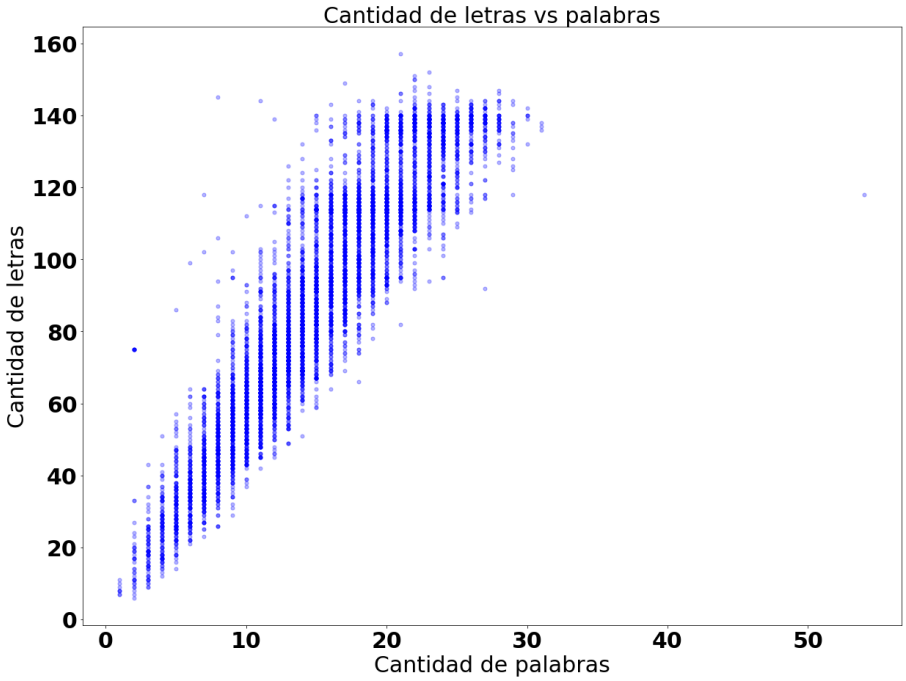
La cantidad media de palabras utilizadas en los tweets es de 15 observando el segundo gráfico (Boxplot) vemos como tenemos un dato nada común alguien publico un tweet con mas de 50 palabras.

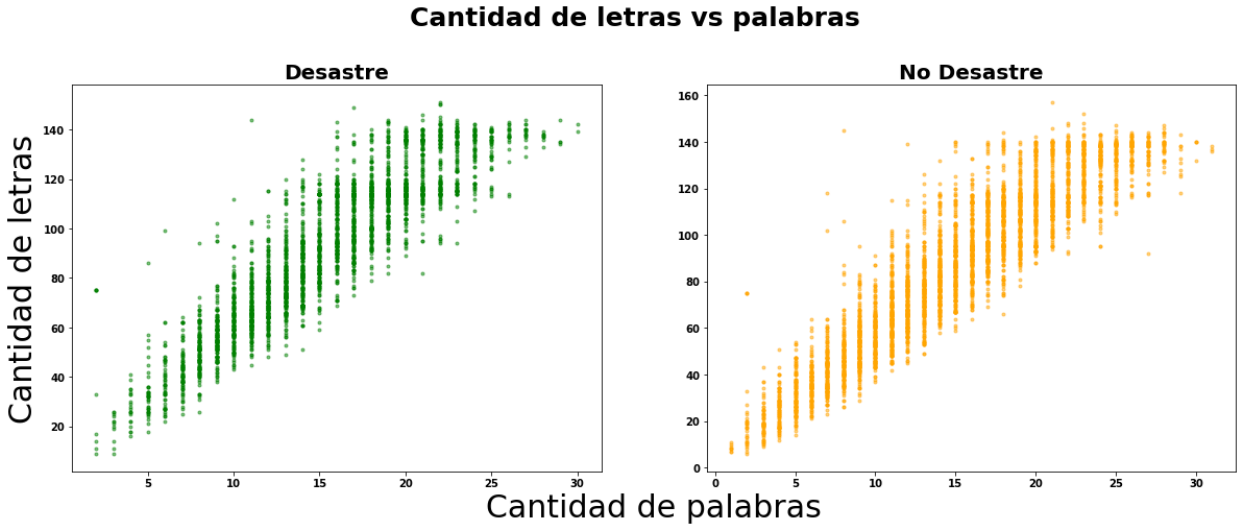




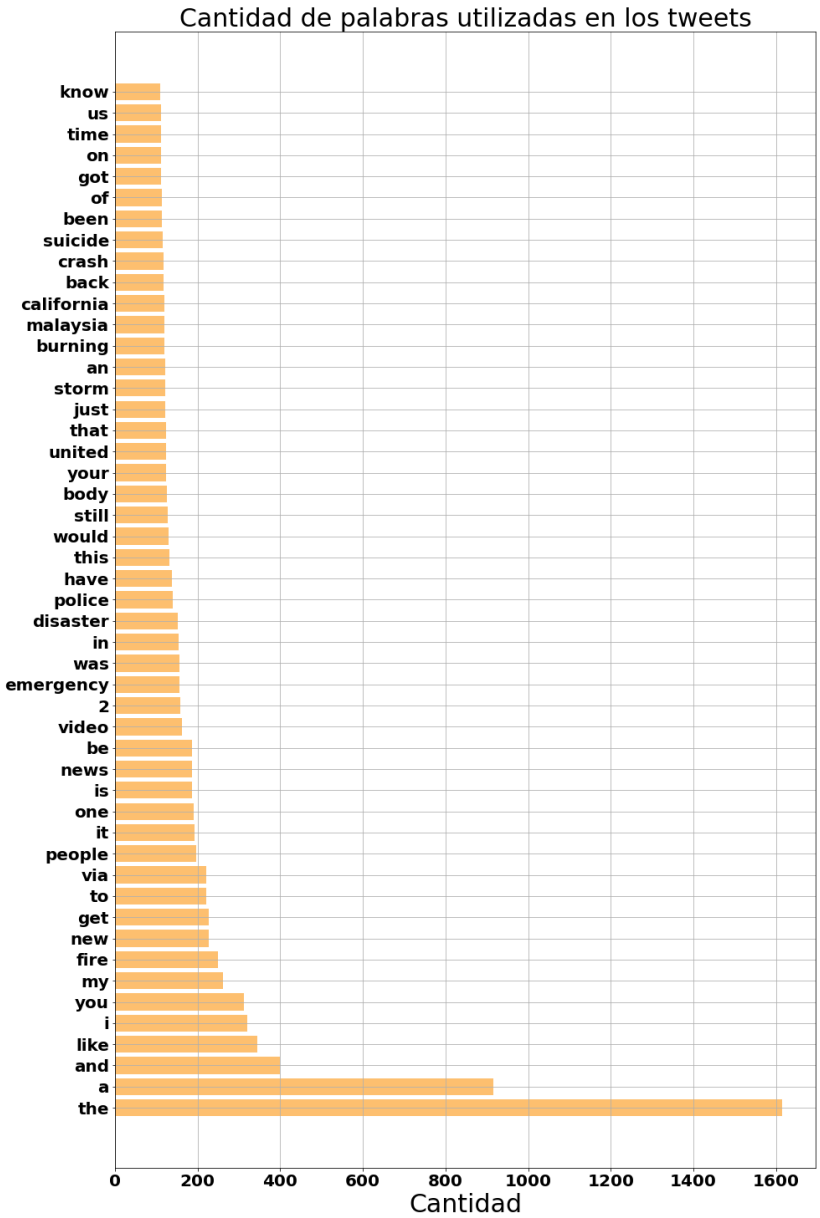
Separando los datos tenemos, para los tweets que no son desastre una forma normal y para los tweets que son desastre podemos verlo como 2 normales solapadas entonces tenemos 2 grupos de personas los que publican tweets con mas de 15 palabras y otras con menos de 15 palabras.

3.7. Relación del largo de la palabra y la cantidad de palabras en el texto



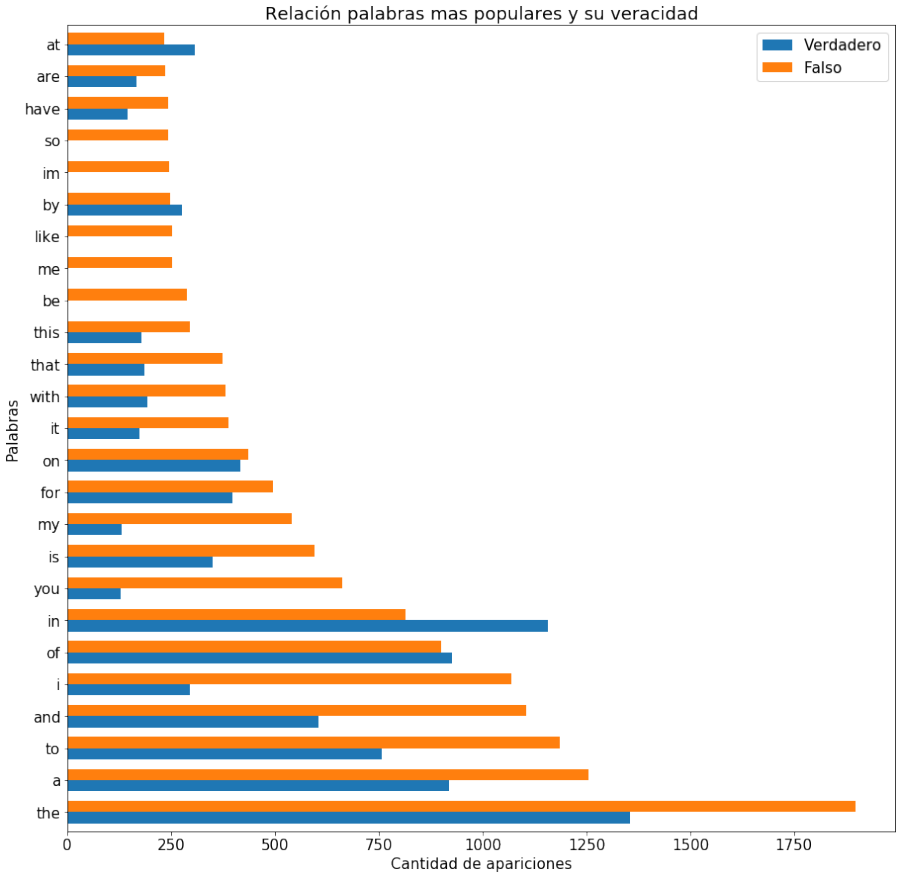


3.8. Las palabras mas populares en los textos



3.9. Comparación palabras mas populares en los tweets

La idea es poder ver si las palabras mas populares en los tweets verdaderos y falsos tienen alguna relación con la veracidad de los mismos

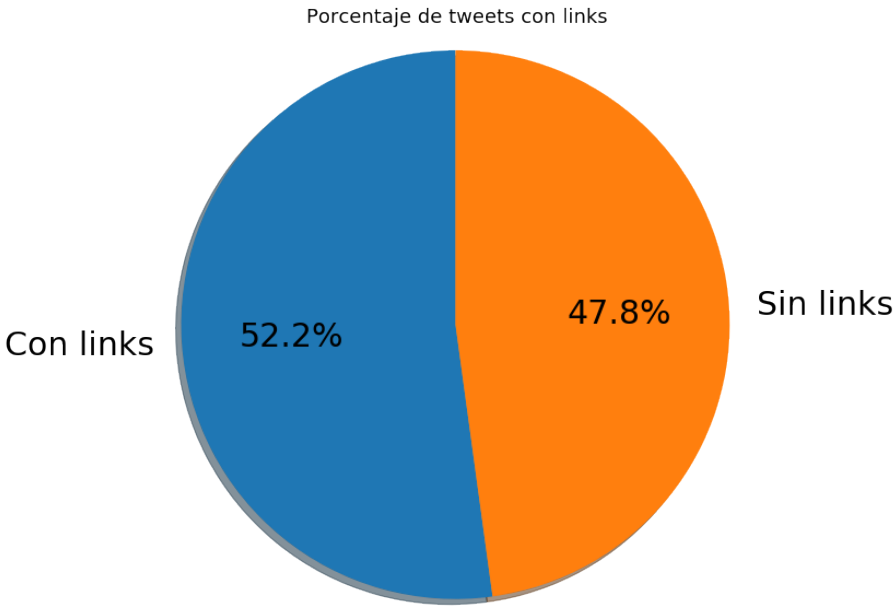


Se puede ver que entre las palabras mas populares de los tweets verdaderos y falsos, solo las palabras "be, me, like, im, so"solo están entre las palabras mas populares de los tweets falsos, y no entre las palabras mas populares de los tweets verdaderos. Sin embargo que estas palabras no estén entre las más populares de los tweets verdaderos no significa que no estén en los tweets verdaderos. Además no son palabras para nada significativas a la hora de identificar si un tweet es verdadero o falso.

3.10. Porcentaje de tuits que tienen links

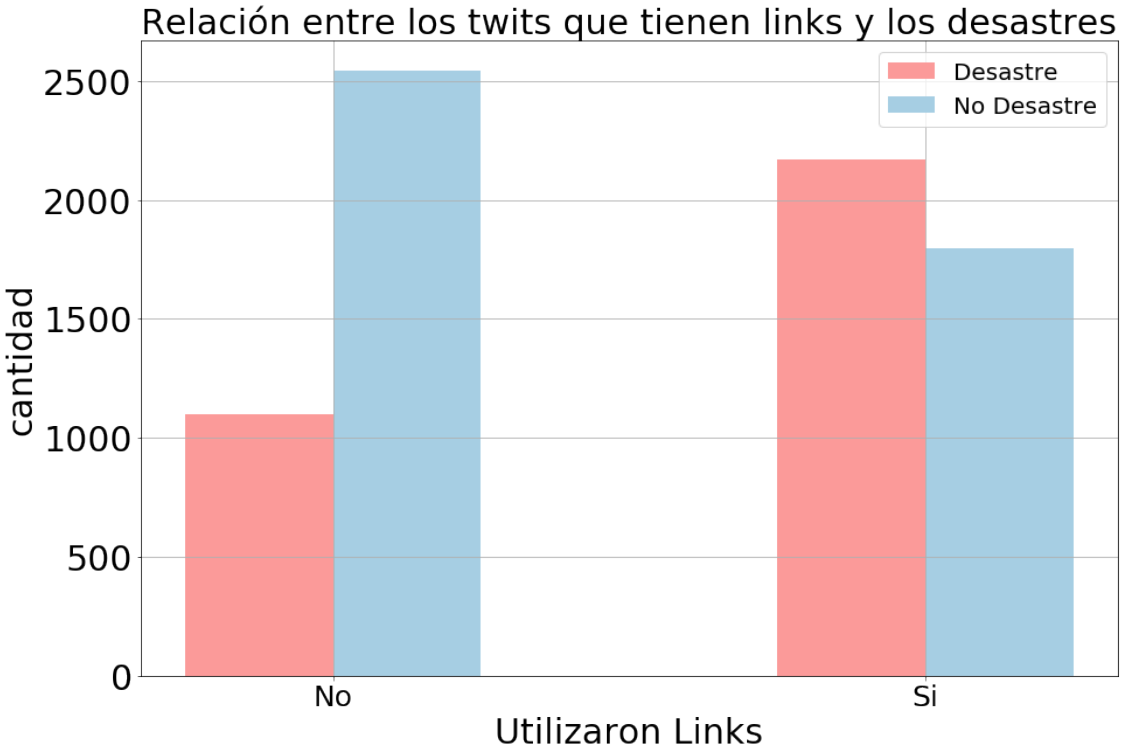
La idea es poder saber si se acostumbra a poner links en los tuits, en ese caso, analizar los sitios a los que se suele apuntar y determinar posteriormente si tiene alguna relación entre la veracidad o falsedad de

los tuits .



Se puede apreciar que más el porcentaje de tweets con links es alto, por lo que vale la pena analizar a donde apuntan esos links y si hay alguna relación entre la veracidad de los tweets y que tengan links o no

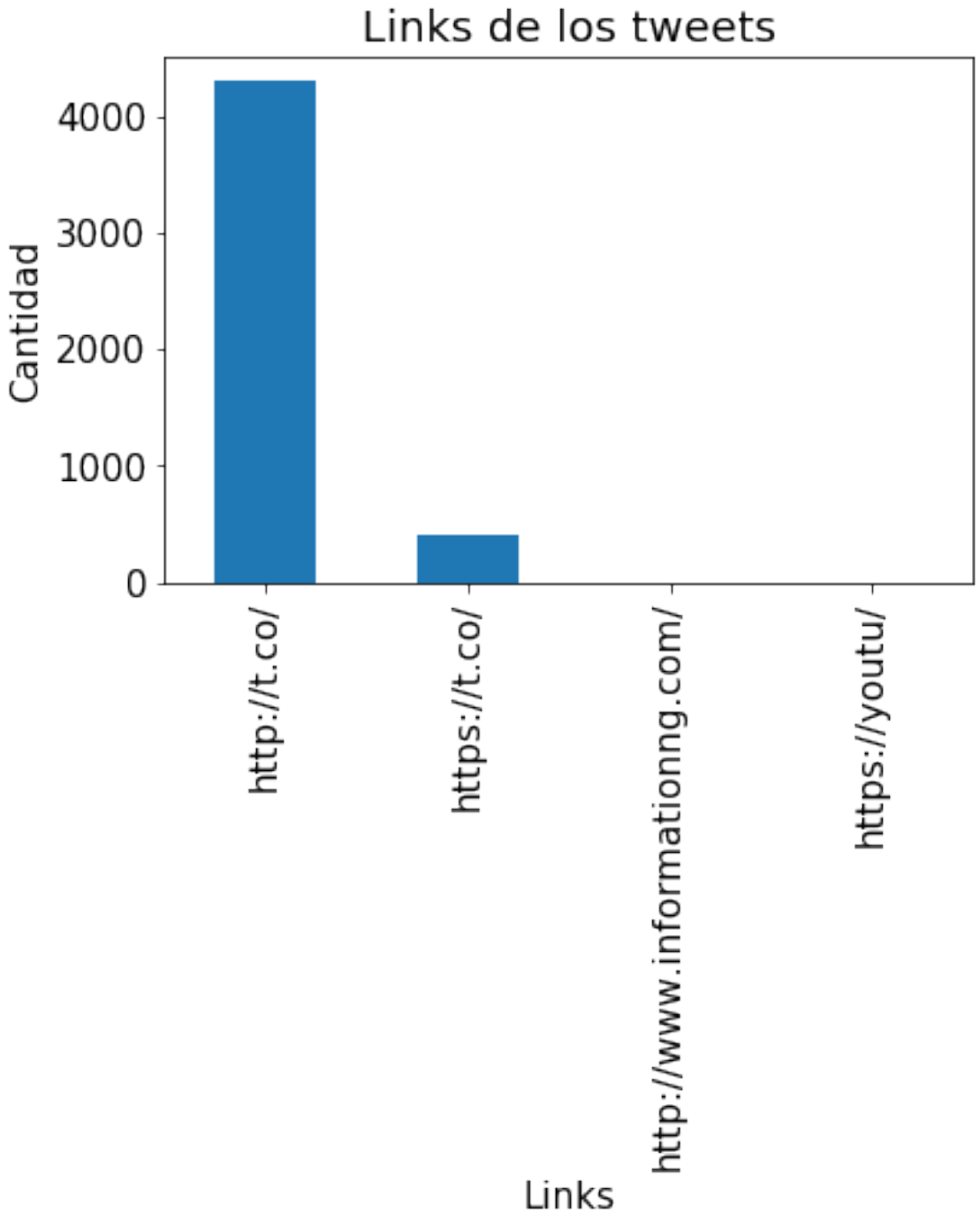
3.11. Relación entre los twits que tienen links y los desastres



Este gráfico muestra claramente que los twits que no tienen links tienden a ser no desastres (por mucha diferencia). En cuanto a los que tienen links tienden a ser desastres reales, pero no por tanta diferencia. Esta información resulta ser muy útil ya que si el twit no tiene links es muy probable de que no se trate de un desastre y si lo tiene, es probable de que informe acerca de un desastre.

3.12. Links mas comunes

Lo que nos interesa ver es cuales son los links más comunes, para luego poder determinar si hay alguna relación por ejemplo entre tuits que tienen links a sitios de noticias y la veracidad de los mismos.

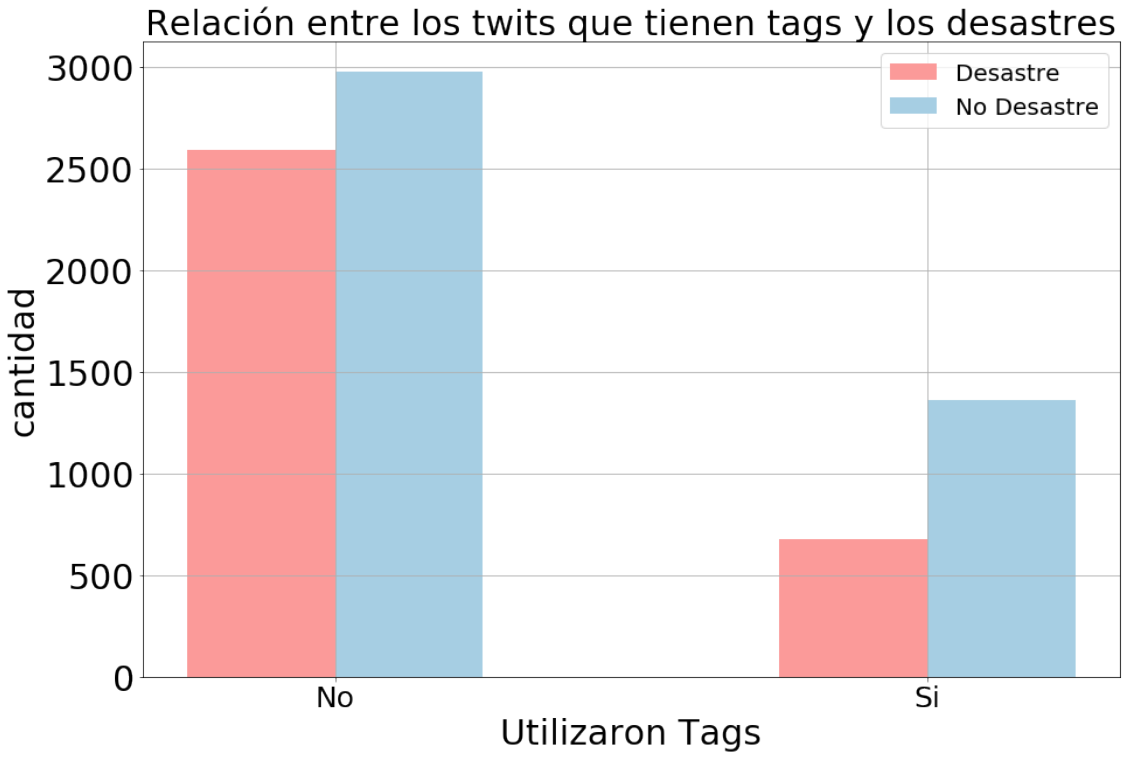


Lo que podemos ver es que la mayoría de los links son acortados por twitter, por lo que no podemos saber a que sitio apuntan de forma eficiente (tendriamos que usar un API por ejemplo para averiguar a donde apuntan esos links, o hacer una petición y ver a donde redirige).

Investigando un poco se obtiene más información de por que prácti-

camente todos links son acortados en este enlace

3.13. Relación entre los twits que tienen etiquetas y los desastres

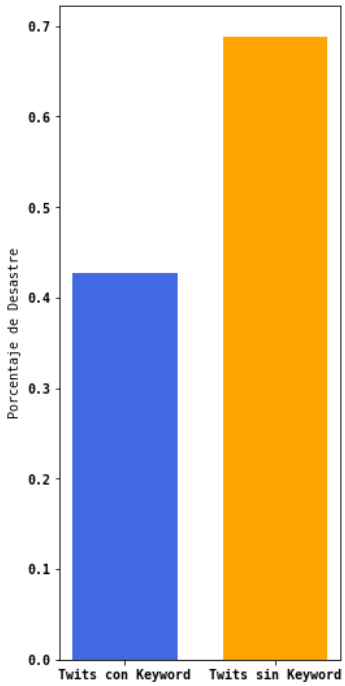


Se puede concluir a través de este gráfico que si el tweet tiene al menos una etiqueta a otra persona, tiene bastante más chances de no ser un desastre. Si no tiene etiquetas, no se puede concluir nada ya que es muy poca la diferencia.

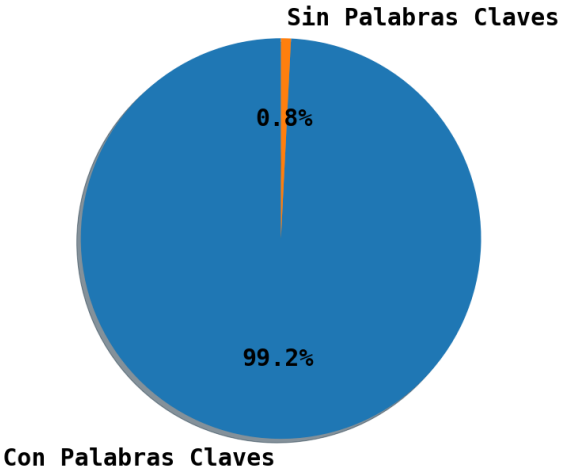
4. Análisis de la palabra clave

4.1. Presencia de las Palabras Claves con los Desastres en Promedio

Dividiendo los tweets entre los que tienen o no tienen palabras claves y calculando cuántos tienen un desastre podemos comparar el porcentaje. Esto nos permite analizar si la existencia de esa categoría nos indica algo sobre la probabilidad de que el tweet se refiera a un desastre real.



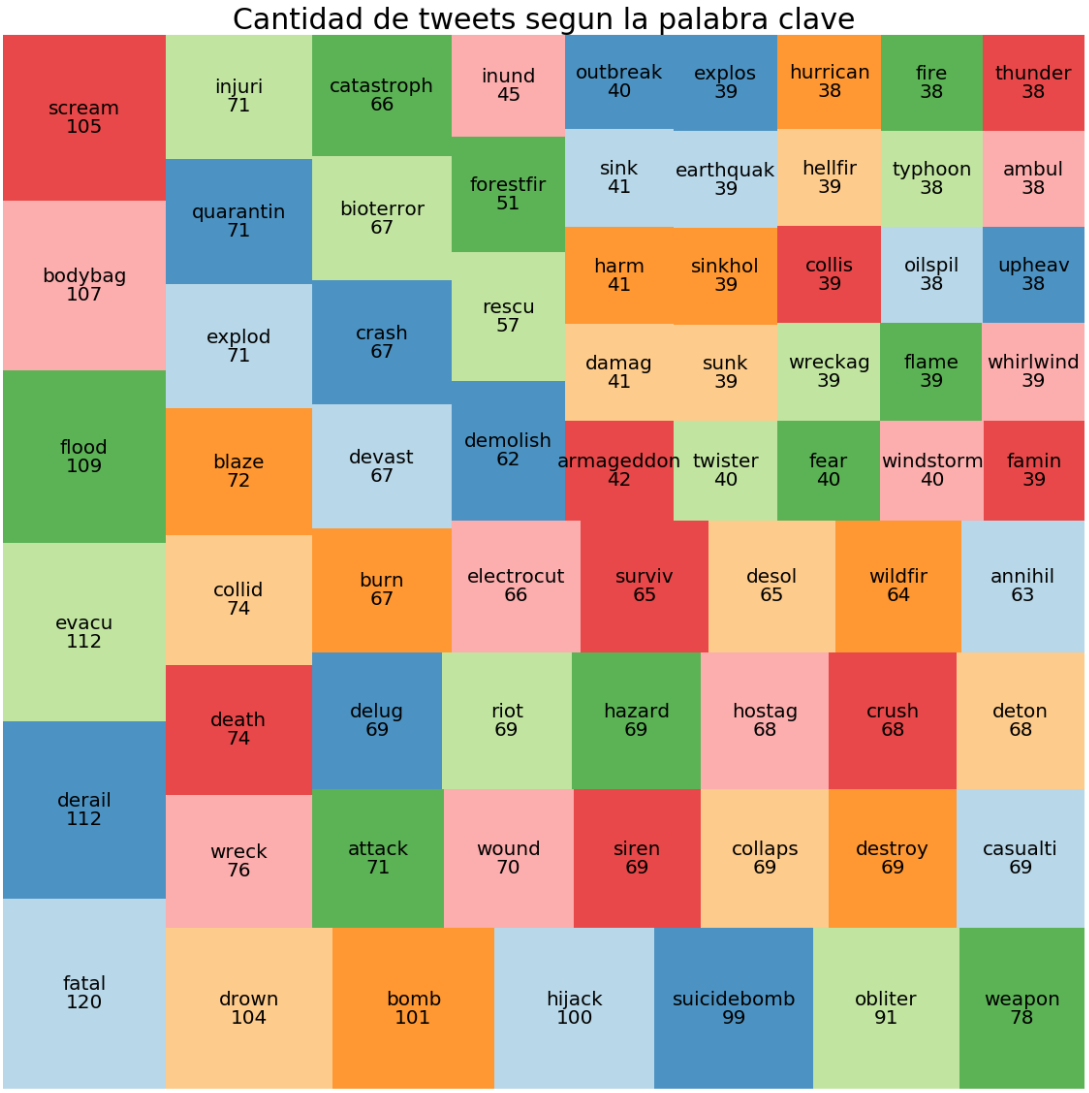
Podemos ver en este gráfico como los tweets sin palabras claves tienen en promedio más desastres. Pero este dato es poco confiable debido a lo a la baja cantidad de tweets que tienen una palabra clave. 0.8 % de los tweets en nuestra base de datos no tiene una palabra clave. Por lo tanto tendríamos que no usar o usar con cuidado esta información para predecir si el tweet se refiere a un desastre real.

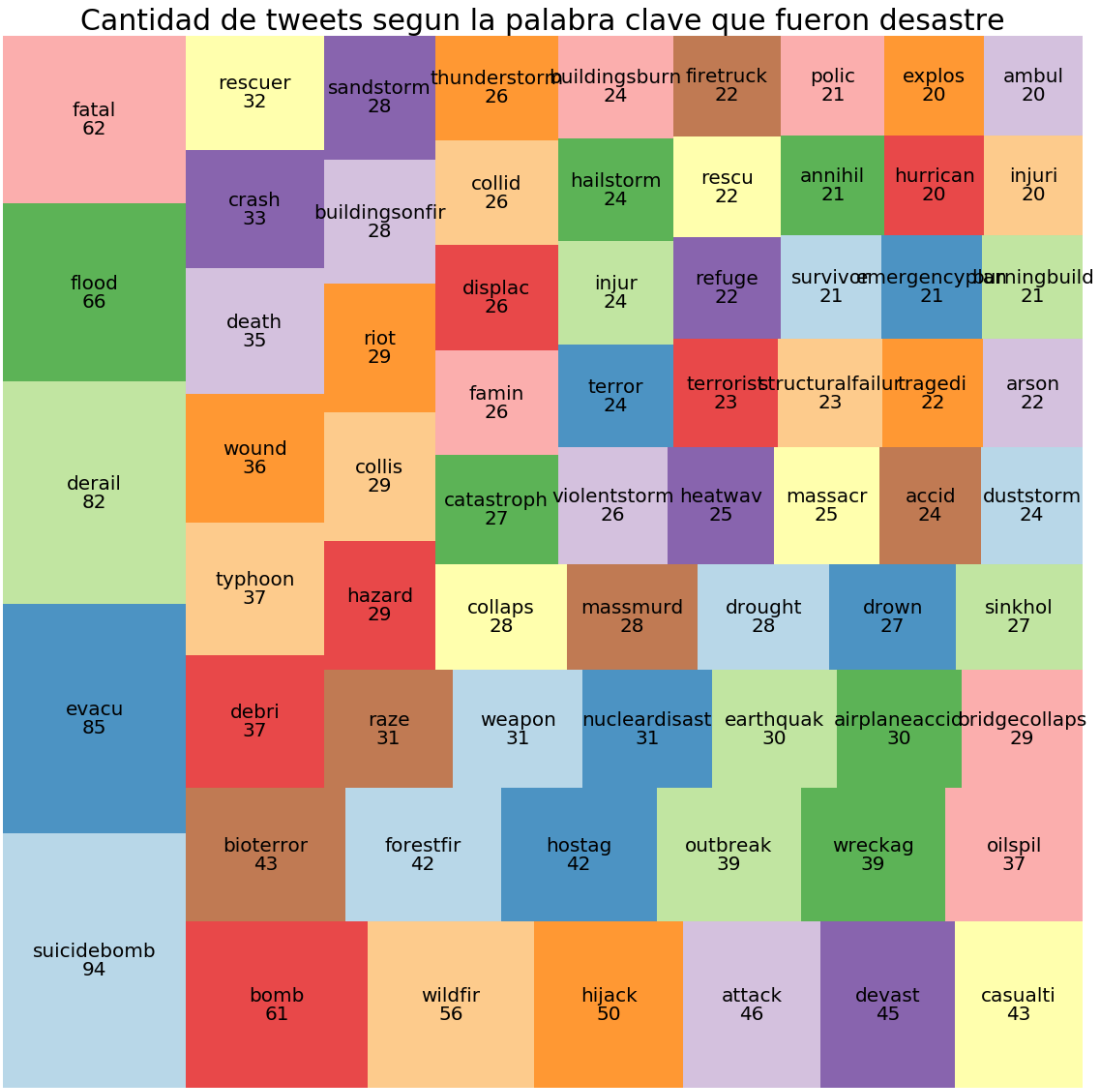


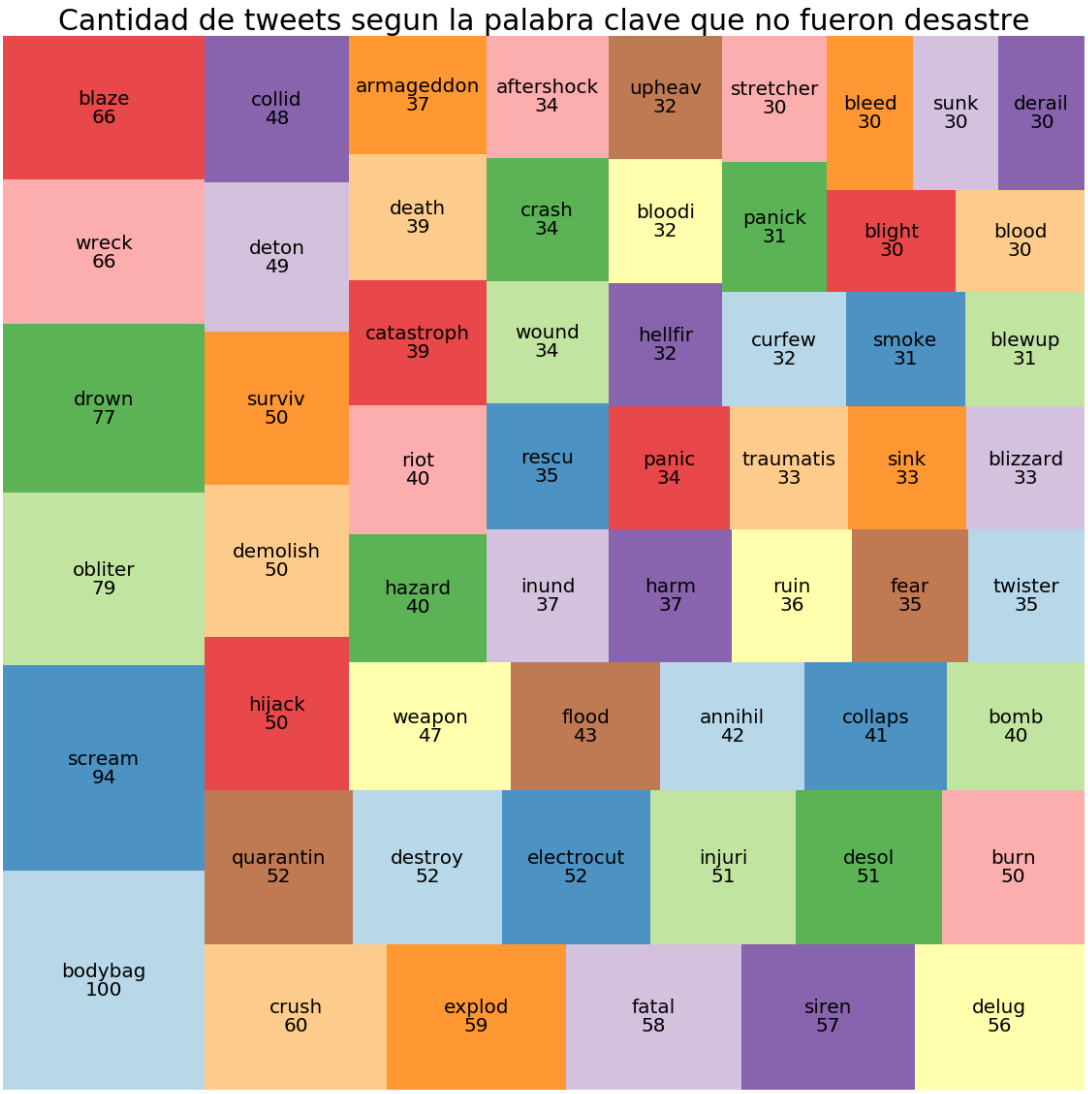
4.2. Palabras Claves con mas y menos Desastres en Total

Queremos saber cuales palabras aparecen mas comúnmente en el contexto de desastre, sean verdad o no. Para eso agrupamos las palabras

claves de acuerdo a su raíz y hacemos un treemap de todas las palabras que aparecen por lo menos 38 veces.




Cantidad de tweets segun la palabra clave que fueron desastre
fatal 62
rescuer 32
sandstorm 28
thunderstorm 26
buildingsburn 24
firetruck 22
polic 21
explos 20
ambul 20
crash 33
buildingsonfir 28
collid 26
hailstorm 24
rescu 22
annihil 21
hurrican 20
injuri 20
flood 66
death 35
riot 29
displac 26
injur 24
refuge 22
survivo 21
emergencyplan 21
burningbuild 21
famin 26
terror 24
terrorist 23
structuralfailur 23
tragedi 22
arson 22
derail 82
wound 36
collis 29
catastroph 27
violentstorm 26
heatwav 25
massacr 25
accid 24
duststorm 24
typhoon 37
hazard 29
collaps 28
massmurd 28
drought 28
drown 27
sinkhol 27
evacu 85
debri 37
raze 31
weapon 31
nucleardisast 31
earthquak 30
airplaneaccid 30
bridgecollaps 29
bioterror 43
forestfir 42
hostag 42
outbreak 39
wreckag 39
oilspil 37
suicidebomb 94
bomb 61
wildfir 56
hijack 50
attack 46
devast 45
casualti 43



Este gráfico nos permite ver algunas raíces mas comunes en el contexto del desastre. En particular vemos que palabras claves que empiezan con evacu- y derail- son indicadores de desastre mientras que bodybag- y scream- son ejemplos de palabras que aparecen mucho mas en casos donde no era un desastre.

4.3. Palabras Claves con mas y menos Desastres en Promedio

Para analizar cuales palabras claves tienen mayor porcentaje de desastres reales agrupando la base de datos y promediando los desastre. Con esto sacamos tablas de las palabras claves que tienen mayor y menor porcentaje de desastre que ocurrieron realmente. Además, agrego una columna con la muestra, para verificar que no sea un punto con muestra

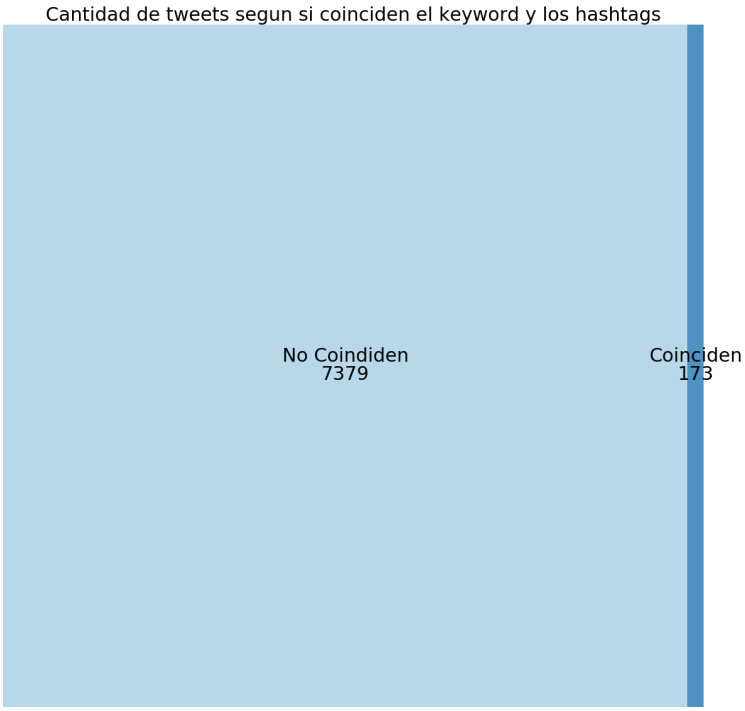
demasiada bajo.

Palabras Claves	Porcentaje de Desastre	Muestra
aftershock	0.0%	34
body bags	2.4%	41
ruin	2.7%	37
blazing	2.9%	34
body bag	3.0%	33

Palabras Claves	Porcentaje de Desastre	Muestra
debris	100.0%	37
wreckage	100.0%	39
derailment	100.0%	39
outbreak	97.5%	40
oil spill	97.4%	38

Podemos ver que palabras como 'body bag' y 'afterschock' son usados rara vez en casos de desastres real. Mientras tant palabras como 'debris' 'wreckage' y 'oil spill' tienden a referirse a desastres reales. Este tipo de palabras nos da un indicio de que deberíamos investigar si el nivel técnico de las palabras tiene una correlación con la probabilidad de que el desastre haya ocurrido.

4.4. Cantidad de tweets donde la palabra clave coincide con los hashtags utilizados



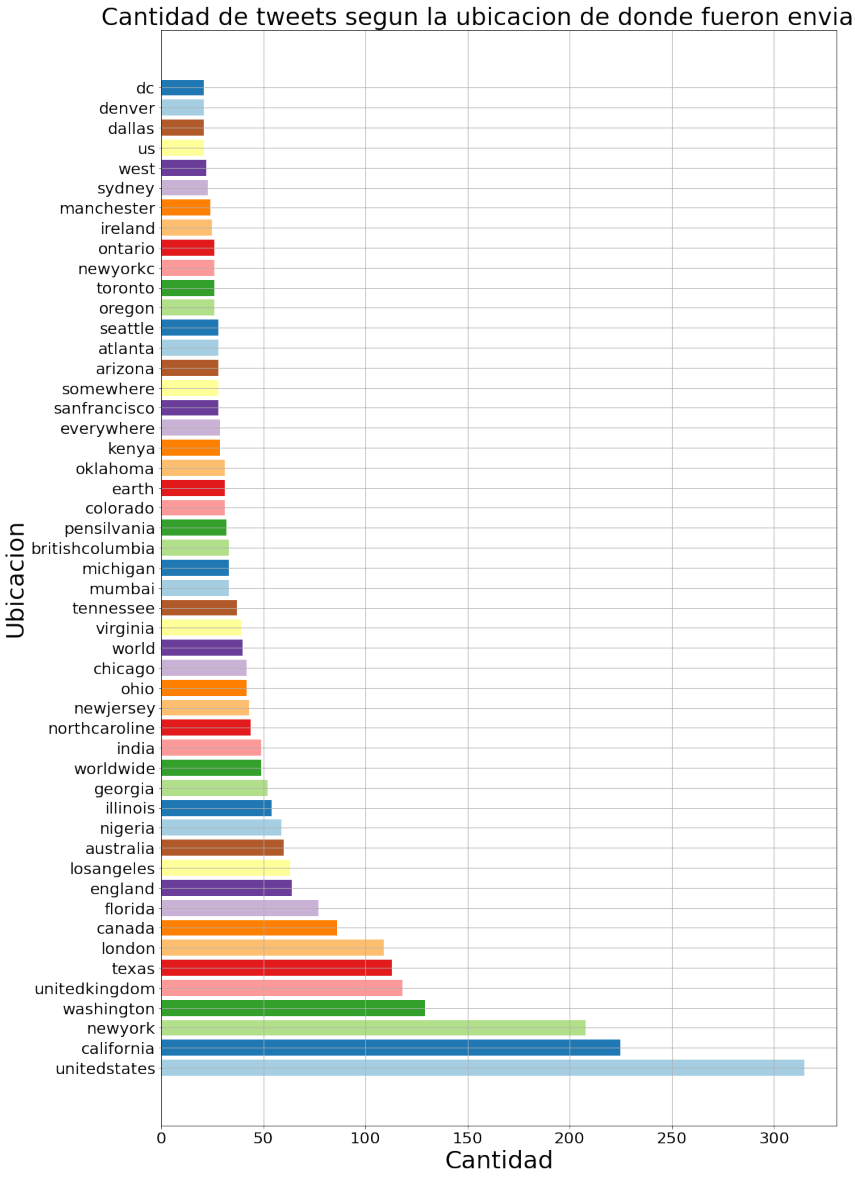
	keyword_coincide_hashtags	cantidad
0	False	7379
1	True	173

		cantidad
keyword_coincide_hashtags	target	
False	0	4232
	1	3147
True	0	91
	1	82

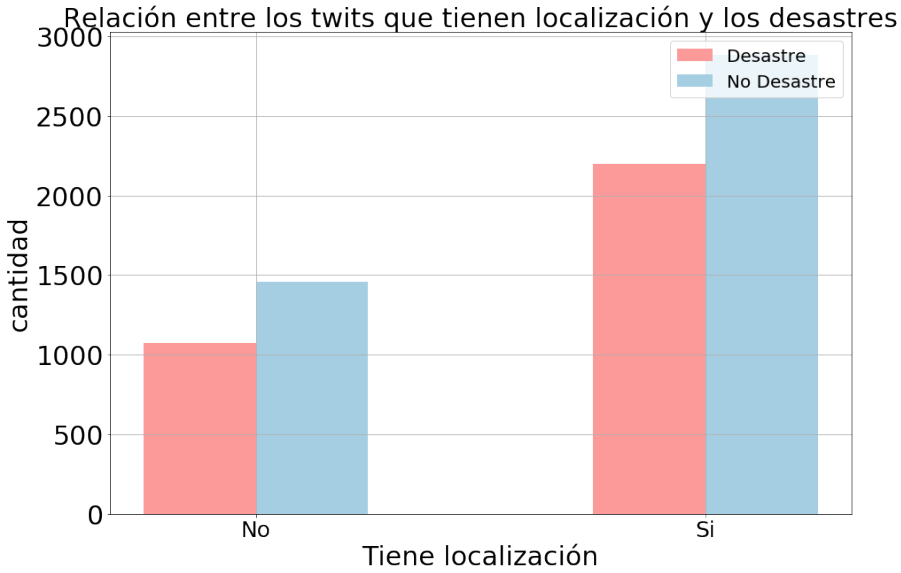
Vemos que la mayoría de los hashtags utilizados no coinciden con la palabra clave,entonces con los hashtags no podemos sacar información sobre el tipo de desastre.

5. Análisis de la ubicación

5.1. Cantidad de tweets según el lugar donde se envían



5.2. Relación entre los twits que tienen localización y los desastres



Esta visualización muestra que la relación entre los twits con o sin localización y los desastres no están para nada vinculados.

6. Conclusiones

- Realmente no se ve relación sustancial entre la longitud de los tweets y su veracidad.
- En el set de datos hay tweets en varios idiomas, pero más del 96 % de los mismos están en inglés.
- La cantidad de tweets que no utilizan hashtags es el triple de la cantidad de tweets que usan hashtags.
- En los tweets que no utilizan hashtags, hay una diferencia sustancial entre los que son veraces y los que no.
- En los tweets que fueron desastre los hashtags mas populares son: Japan, Hiroshima, Sismo, wildfird, WorldNews.
- La cantidad de palabras de los tweets se concentra alrededor de 15 palabras, a medida que usan menos palabras, o mas palabras, hay menos tweets. Sin embargo tiene un compartimiento atípico, porque para los tweets con una cantidad de palabras cercanos a la media (15 palabras) hay un decremento fuerte de tweets (se puede observar en la visualización).
- Las palabras mas populares de los textos son banales: "the, a, and, like, I, you, my...".
- Comparando las palabras mas populares de los tweets verdaderos y falsos podemos ver que en los tweets falsos hay algunas palabras que son más populares que en los tweets falsos "be, me, like, im, so"pero es dificil explicar el por qué de esto.
- La cantidad de tweets que tienen links y las que no tienen es prácticamente la misma.
- Los links que aparecen en los tweets son en su mayoría acortados por tweeter, lo que hace dificil poder ver si hay alguna relación entre los links apuntados y la veracidad de los tweets.
- Las keywords que más identifican a los tweets que representan un desastre son: "fatal, flood, derail, evacu, suicidebomb, bomb, bioterror".
- Las ubicaciones mas comunes en los tweets son "nitedstated, california, newwork, washington, edkingdom, texas, london, canada, florida, england, losangeles, australia, nigeria".
- Que un link tenga localización o no, no parece estar relacionado con si representa un desastre o no.