FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico 1

Análisis Exploratorio de Datos

CuarenData 2.0

Primer Cuatrimestre 2020

Fecha de Entrega: 21/05/2020

Integrantes:

Nombre y Apellido	Padrón
Anarella Nicoletta	94.551
Pablo Prieto	91.561
Sherly Porras	91.076
Zoraida Flores	87.039

Link al repositorio: https://github.com/ZoraidaF/Cuarendata

Introducción

El objetivo del presente trabajo práctico es realizar un análisis exploratorio sobre un set de datos puntual. En esta ocasión se parte de un archivo CSV, que contiene tweets (publicaciones dentro de la red social Twitter) e información relacionada a los mismos. Este set de datos fue recopilado y publicado por Kaggle en el marco de una competencia de Machine Learning, en la que figura como set de entrenamiento para el modelo predictivo.

El objetivo de la competencia es poder predecir si un tweet hace referencia a un hecho catastrófico (o a algún tipo de desastre) real, o no. Los tweets pertenecientes al set dado ya han sido manualmente clasificados en base a su veracidad (al referir a un suceso desafortunado).

El archivo se llama ‘train.csv’ , y contiene las siguientes 5 columnas:

- **id:** identificador único de cada tweet
- **keyword:** palabra clave para el tweet (podría estar en blanco)
- **location:** ubicación desde la que fue enviada el tweet (podría estar en blanco)
- **text:** texto del tweet
- **target:** indica si se trata de un desastre real (1) o no (0)

En lo que respecta a volumen de datos, el archivo cuenta con 7.613 tweets.

Como herramienta para realizar el análisis exploratorio hemos decidido utilizar la herramienta Python Pandas.

Data Cleansing

Pudimos detectar que en el set de datos hay 61 tweets que tienen el campo keyword en blanco. Considerando que representan menos del 1% de la cantidad total de registros, hemos decidido descartarlos al momento de generar las categorías de los keywords.

Para analizar las regiones desde donde fueron enviados los tweets, también se realizó una limpieza del dataset. Se descartaron aquellos tweets que tenían el campo location en blanco como primer paso.

Lo siguiente que hicimos fue filtrar también los tweets que tenían como ubicación información que no correspondía a una región, por ejemplo WorldWide.

Para las regiones de los tweets que tenían similitudes, lo que hicimos fue unificar la clave de la región para un mejor análisis, donde por ejemplo las regiones New York, NYC, New York, NY, New York City, new york, NY, ny se unificaron en una clave New York.

Además, para las ubicaciones que corresponden a un país, consideramos tomar como región la capital de dicho país para que el análisis tenga un mejor enfoque. Por ejemplo, USA, United States, US les asignamos la clave Washington, D.C.

Luego, armamos un nuevo dataset, con las claves de las regiones analizadas y con información de latitud y longitud, para poder generar un archivo kml con el cual, gracias a la herramienta de ***Google My Maps***, obtener una geo-visualización que nos ayude a comprender mejor los datos.

El set de datos final se vio reducido a 5080 registros, ya que el 33% de los tweets original no informaba locación o tenía basura en dicho campo.

Análisis Exploratorio

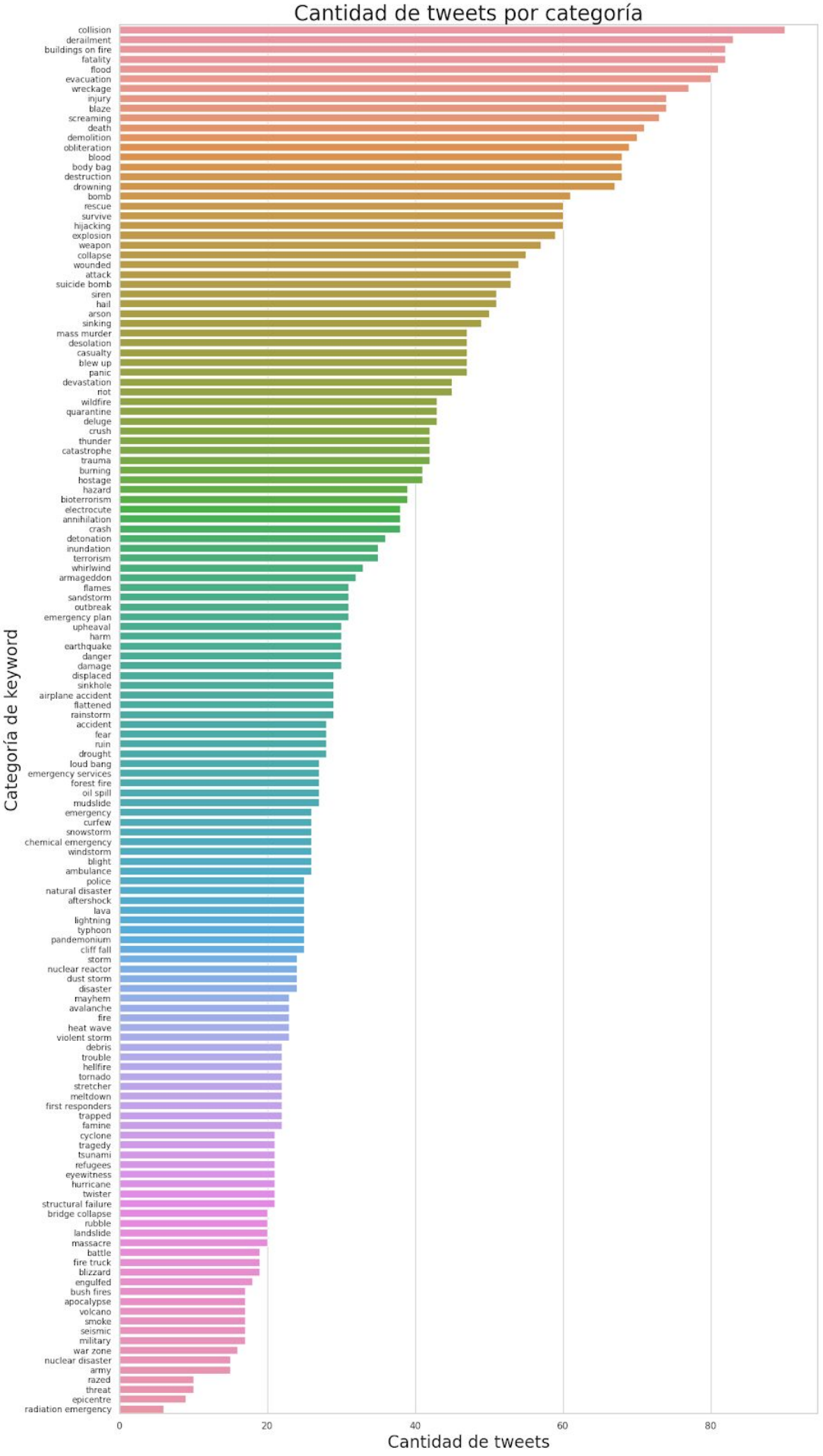
¿Cómo son los keywords de los tweets?

Observamos que algunos de los keywords eran muy similares entre sí, o hacían referencia al mismo evento. Por ejemplo, los términos *ablaze*, *blaze* y *blazing*, pueden asociarse ya que son distintas formas de referirse a un incendio (o llamaradas). De esta forma, podemos decir que pertenecen a la misma categoría *blaze*. Siguiendo este razonamiento generamos una nueva columna en el dataframe, llamada ***keyword_category***, con la categoría asociada a cada tweet.

Cabe destacar que nos limitamos a agrupar aquellas palabras que tenían no solo significado en común, sino que también compartían la misma raíz. El análisis se realizó en forma manual, partiendo de la totalidad de keywords recibidos.

Inicialmente, en el dataset contábamos con 221 keywords diferentes. Luego de clasificar a los keywords, llegamos a un conjunto de 142 categorías de keywords.

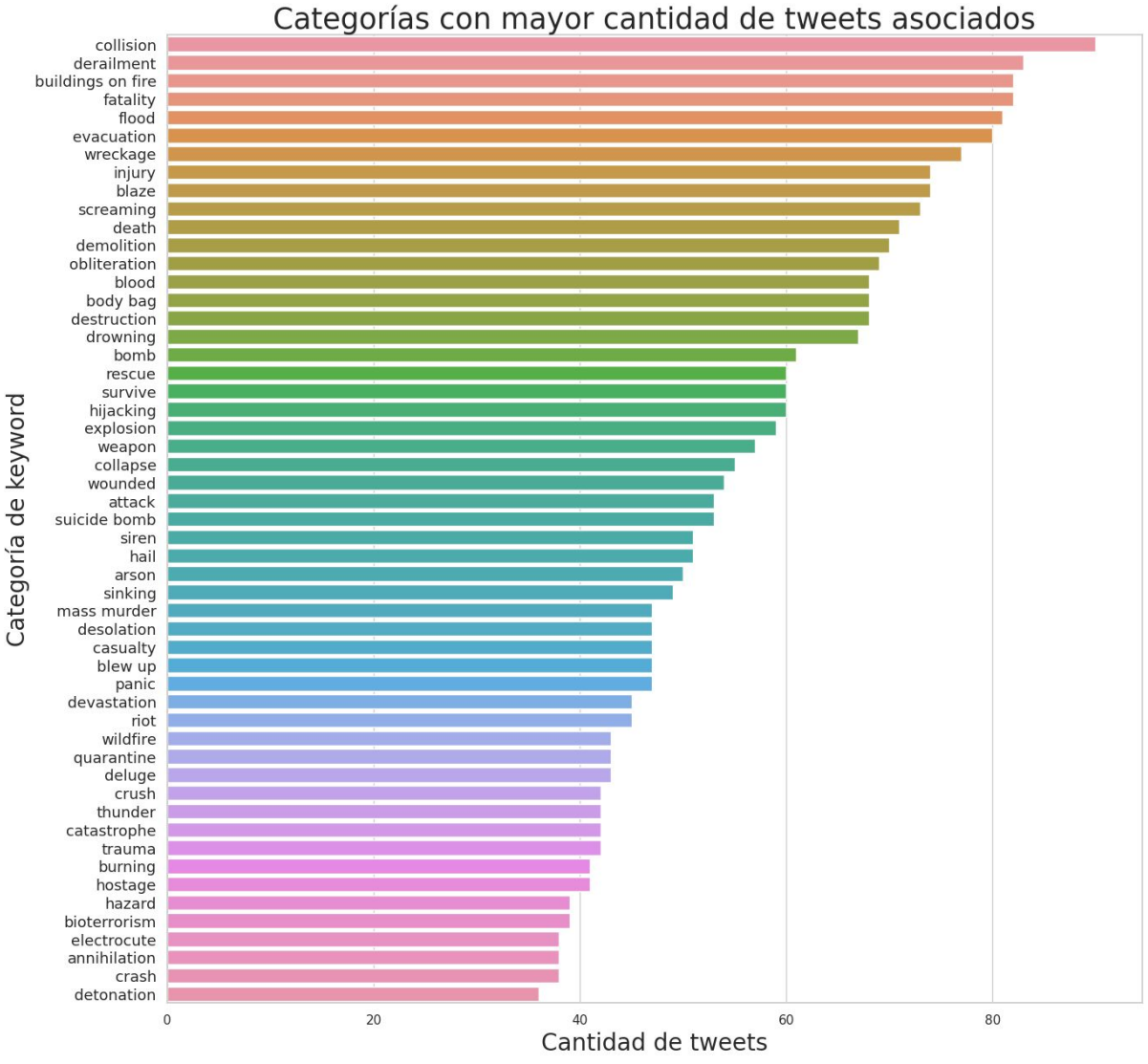
El siguiente gráfico muestra la distribución de todos los tweets por categorías. Sin embargo, dada la gran cantidad de las mismas, no se llega a apreciar la información de la mejor manera. Solamente se observan agrupaciones de barras de longitud muy similar.


Cantidad de tweets por categoría
Categoría de keyword
collision
derailment
buildings on fire
fatality
flood
evacuation
wreckage
injury
blaze
screaming
death
demolition
obliteration
blood
body bag
destruction
drowning
bomb
rescue
survive
hijacking
explosion
weapon
collapse
wounded
attack
suicide bomb
siren
hail
arson
sinking
mass murder
desolation
casualty
blew up
panic
devastation
riot
wildfire
quarantine
deluge
crush
thunder
catastrophe
trauma
burning
hostage
hazard
bioterrorism
electrocute
annihilation
crash
detonation
inundation
terrorism
whirlwind
armageddon
flames
sandstorm
outbreak
emergency plan
upheaval
harm
earthquake
danger
damage
displaced
sinkhole
airplane accident
flattened
rainstorm
accident
fear
ruin
drought
loud bang
emergency services
forest fire
oil spill
mudslide
emergency
curfew
snowstorm
chemical emergency
windstorm
blight
ambulance
police
natural disaster
aftershock
lava
lightning
typhoon
pandemonium
cliff fall
storm
nuclear reactor
dust storm
disaster
mayhem
avalanche
fire
heat wave
violent storm
debris
trouble
hellfire
tornado
stretcher
meltdown
first responders
trapped
famine
cyclone
tragedy
tsunami
refugees
eyewitness
hurricane
twister
structural failure
bridge collapse
rubble
landslide
massacre
battle
fire truck
blizzard
engulfed
bush fires
apocalypse
volcano
smoke
seismic
military
war zone
nuclear disaster
army
razed
threat
epicentre
radiation emergency
0
20
40
60
80
Cantidad de tweets

¿Cuáles son las categorías de keywords con más tweets asociados?

Para esta sección, se utilizó la categoría del keyword generada antes. Además, decidimos filtrar el dataframe considerando únicamente aquellas categorías para las que la cantidad de tweets supera el promedio general de tweets por categoría (que es de casi 36 tweets).

Las categoría más frecuente es *collision*, que engloba además *collide* y *collided*. Así mismo las restantes del gráfico son, en todo caso, agrupaciones de más de un término (keyword). Se ve que la forma del gráfico es muy similar a la de aquel obtenido en el ítem anterior, dado que las distribuciones de keywords son bastante parejas.

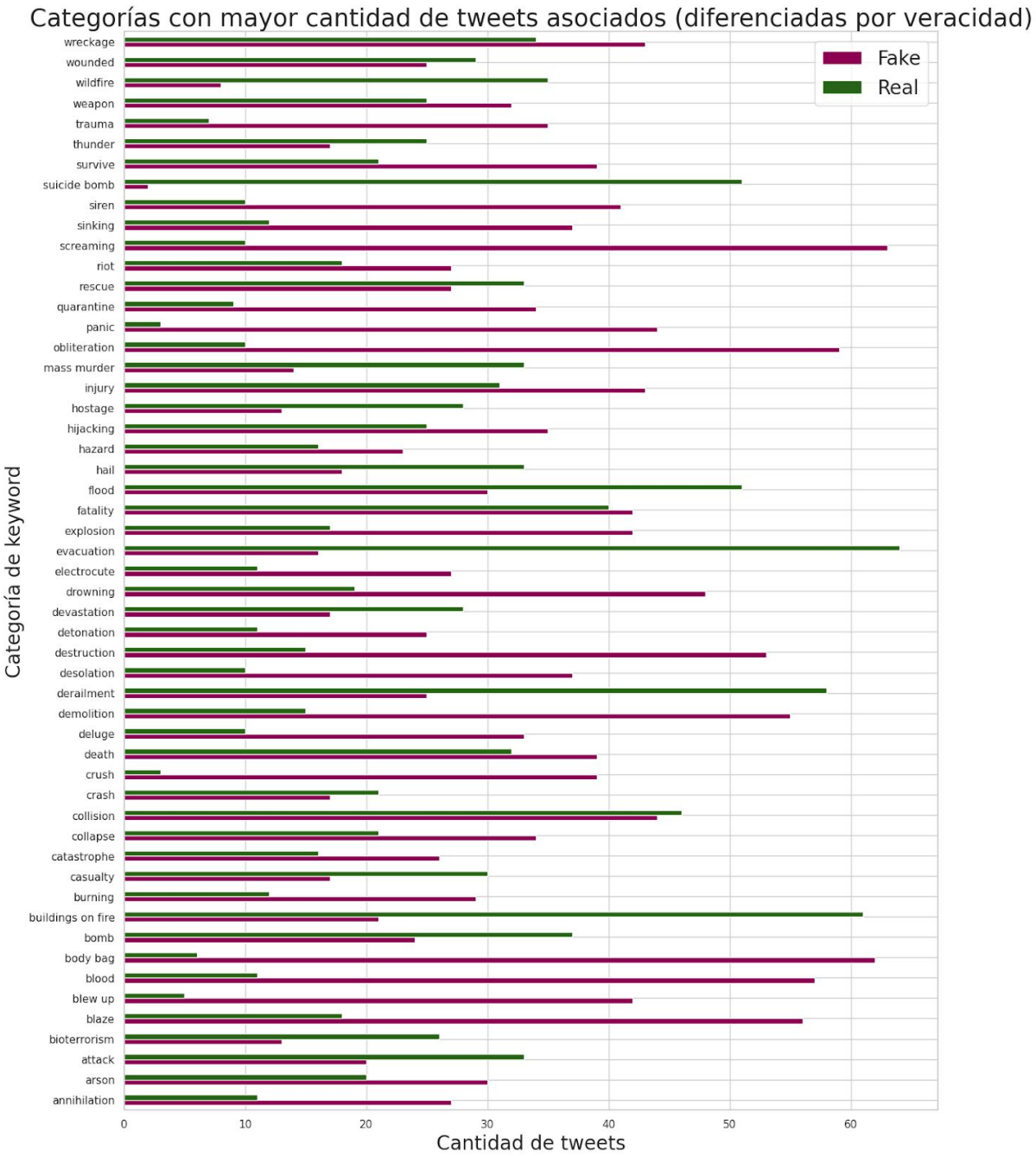


¿Cómo se distribuyen los tweets verdaderos y falsos dentro de estas categorías?

Aquí podemos encontrar escenarios de los más diversos. Por un lado, hay categorías con distribuciones muy parejas, como en el caso de *collision* y *fatality*, que tienen casi la misma cantidad de tweets verdaderos y falsos.

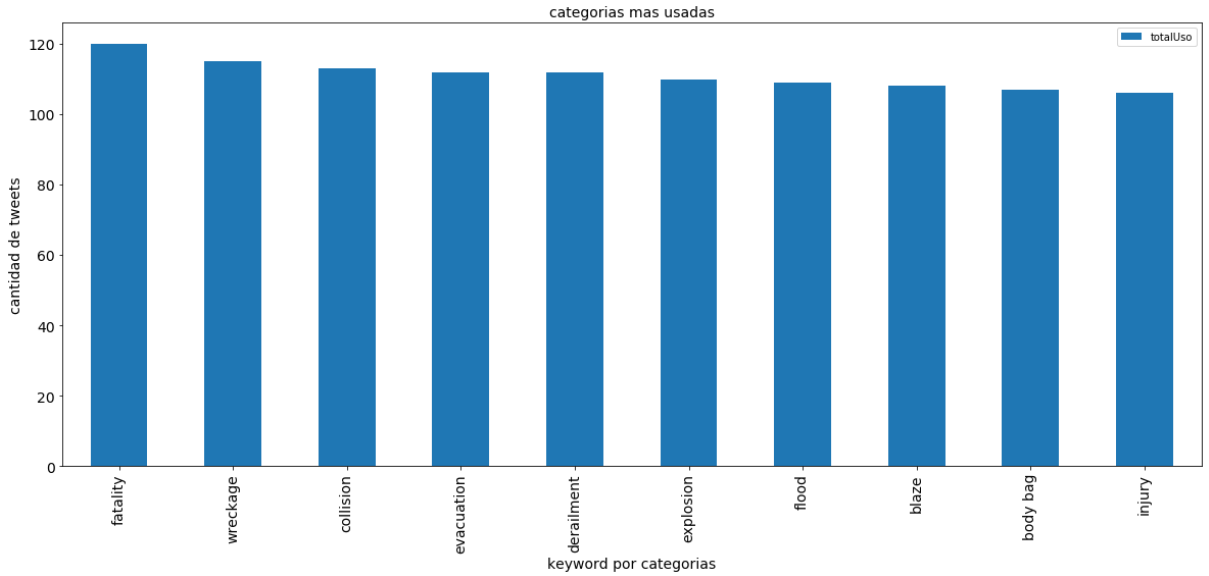
Si nos concentramos en la diferencia entre tweets verdaderos y falsos, se destaca *suicide bomb*, donde prácticamente todos los tweets publicados hacen referencia a un suceso desastroso real.

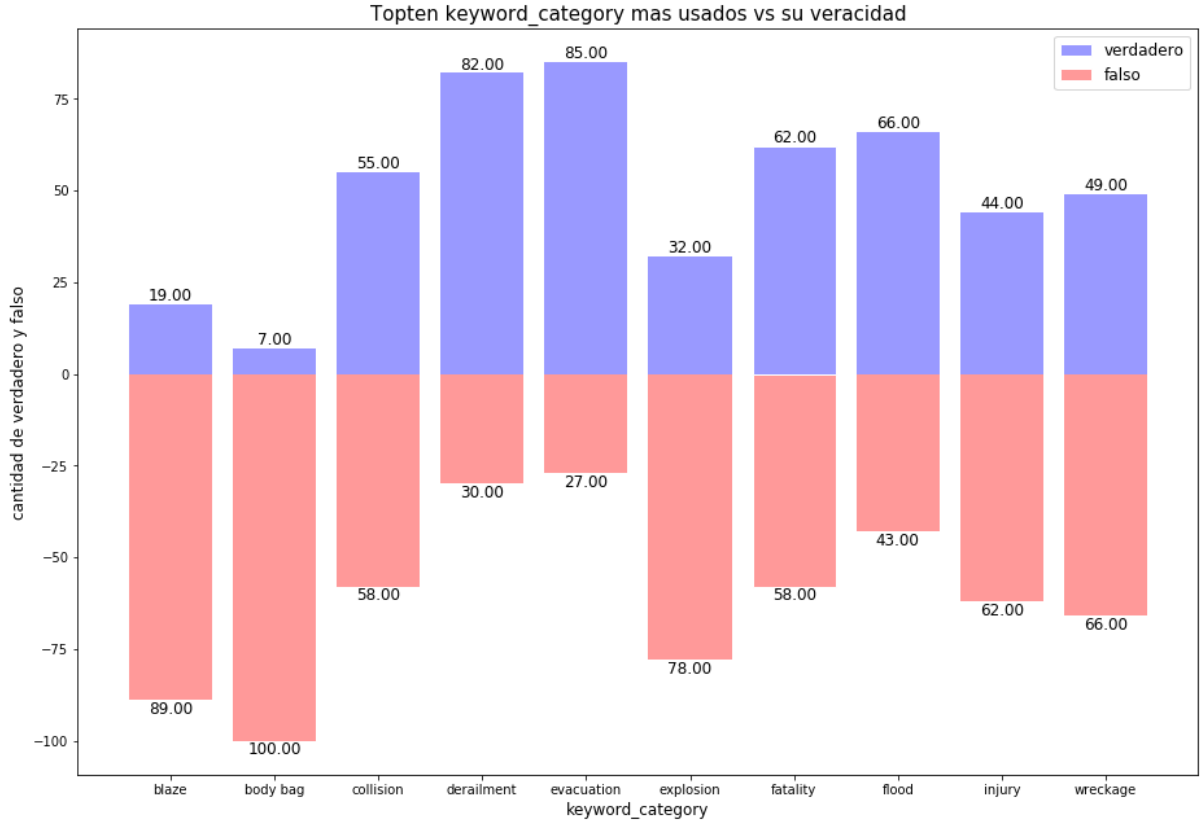
Al contrario sucede con los términos *panic, screaming* y *crush*, donde ésta relación se invierte y predominan los fake tweets. Es interesante que *crush* también puede significar 'enamoramiento' o 'amor platónico', y éste término rara vez se usa al reportar la ocurrencia de una catástrofe.



¿Cuáles son las 10 categorías de keywords más usadas, con su respectiva veracidad?

Al filtrar por categoría tenemos a categoría Fatality como con más keyword por lo que tendremos más tweets clasificados.

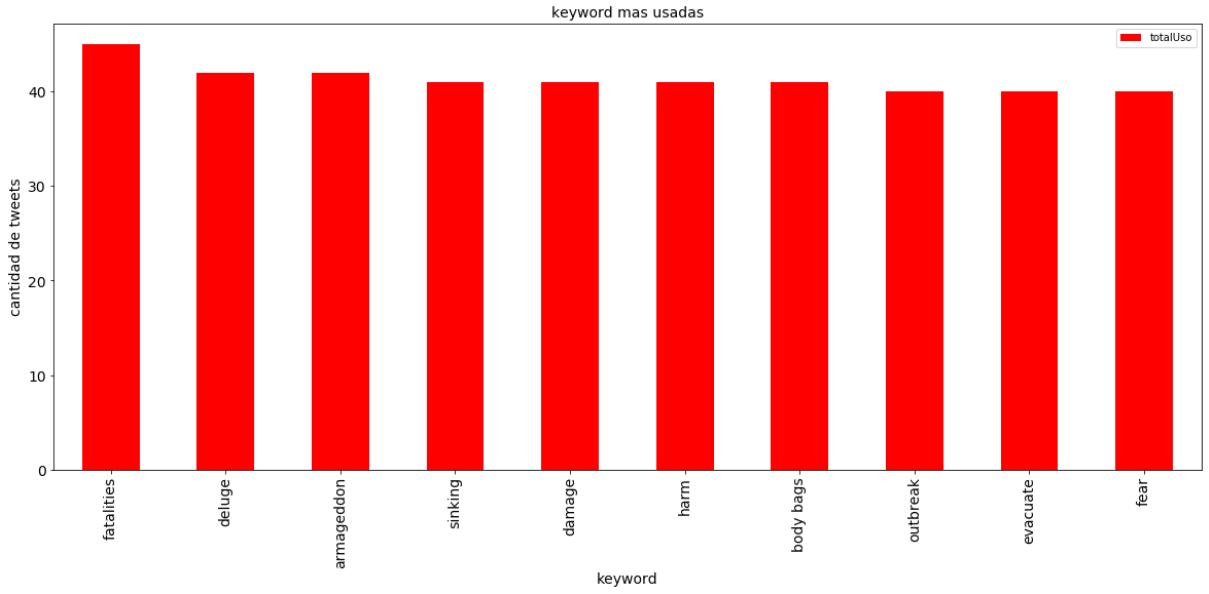


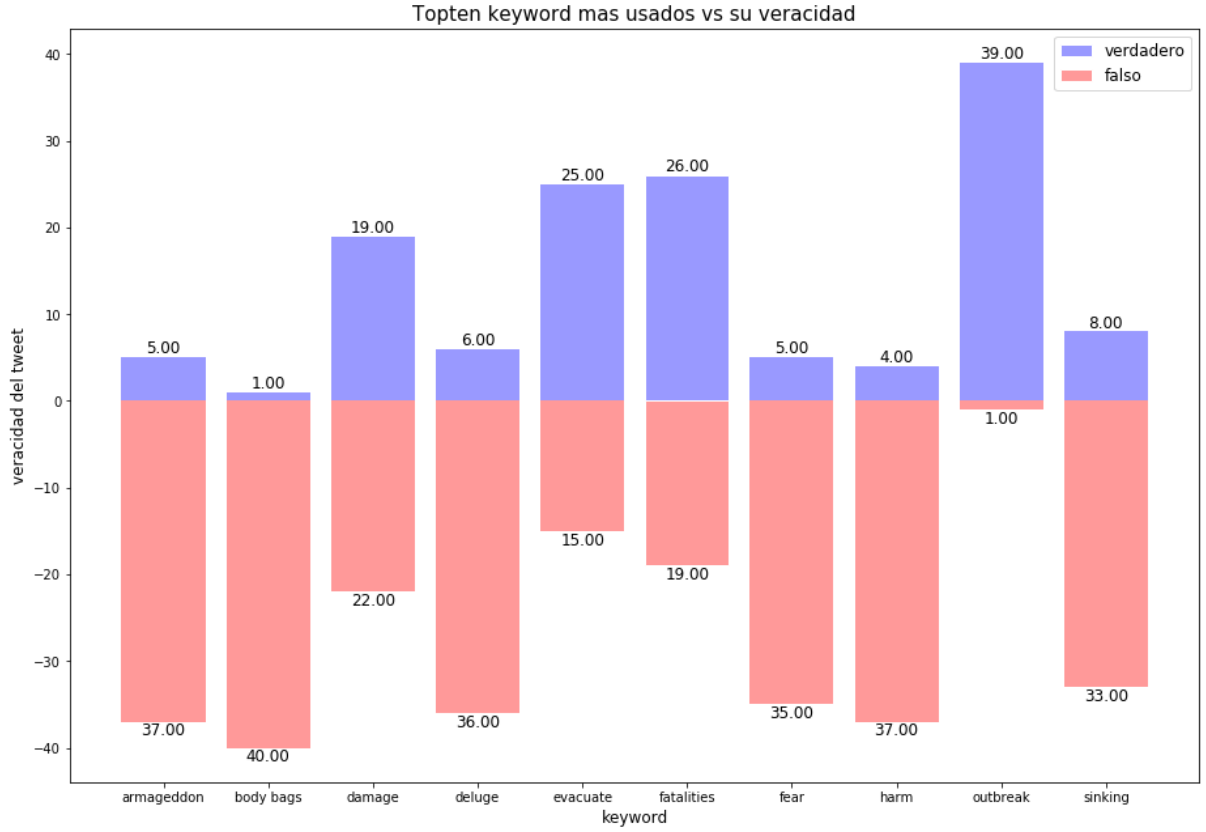


¿Cuáles son los 10 keywords más usados, con su respectiva veracidad?

Vemos a los 10 keywords más usados, que referencian a los diferentes tweets, y con ello a la cantidad de verdadero o falsos.

Además podemos observar que el keyword que es más usado es el fatalities.

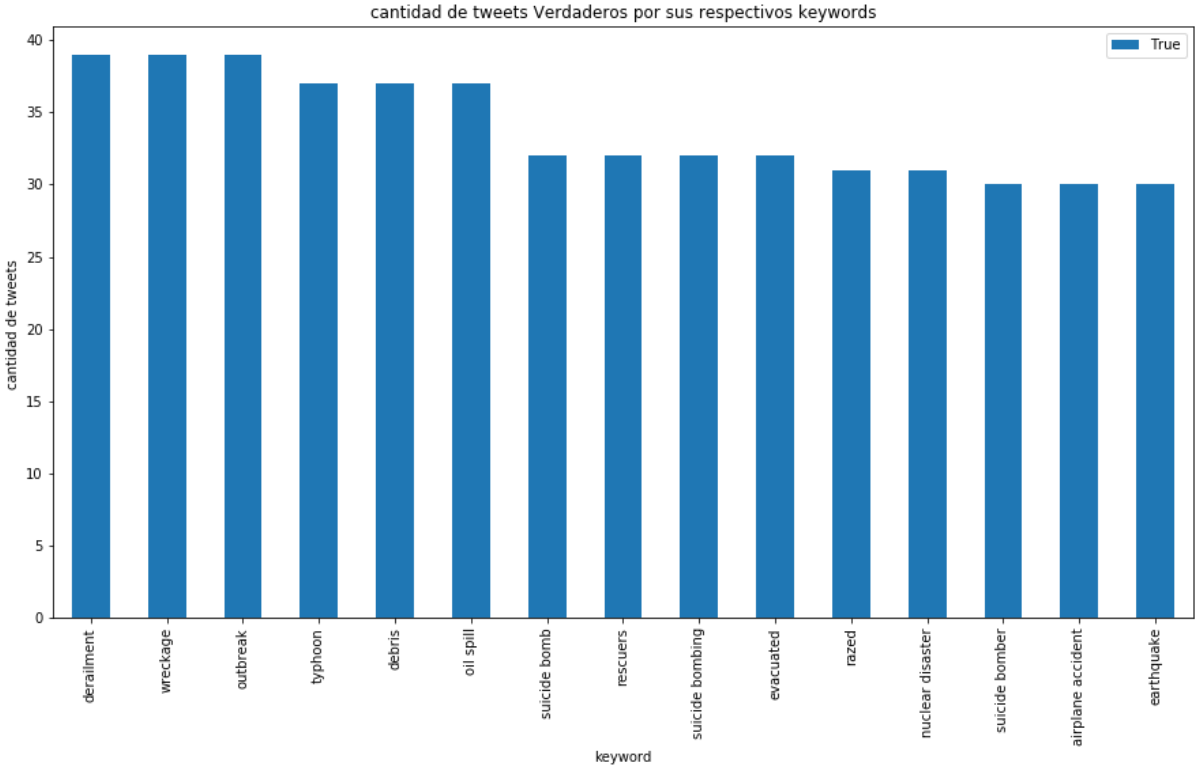




Son los Keywords mas usados sin incluir los nans

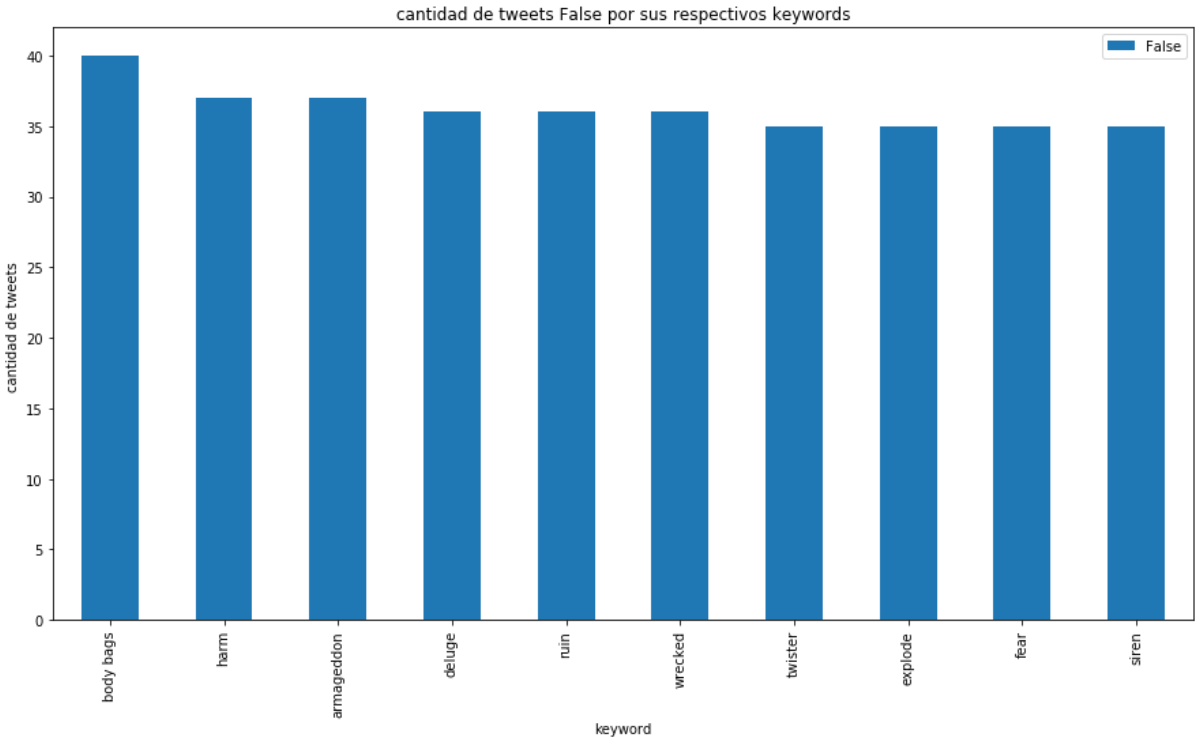
¿Cuáles son los 10 keywords que tienen mayor cantidad de tweets verdaderos?

A continuación vemos los keywords que están asociados a tweets que representan verdaderos desastres.

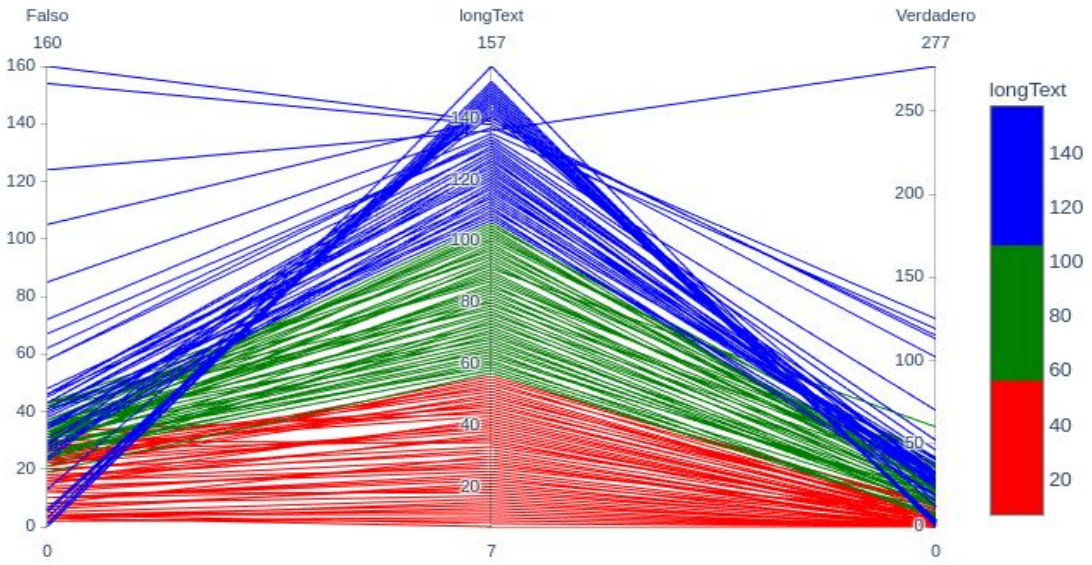


¿Cuáles son los 10 keywords que tienen mayor cantidad de tweets falsos?

El siguiente gráfico ilustra a aquellos keywords que están asociados a una gran cantidad de tweets que son falsos, es decir, que no refieren a un desastre.



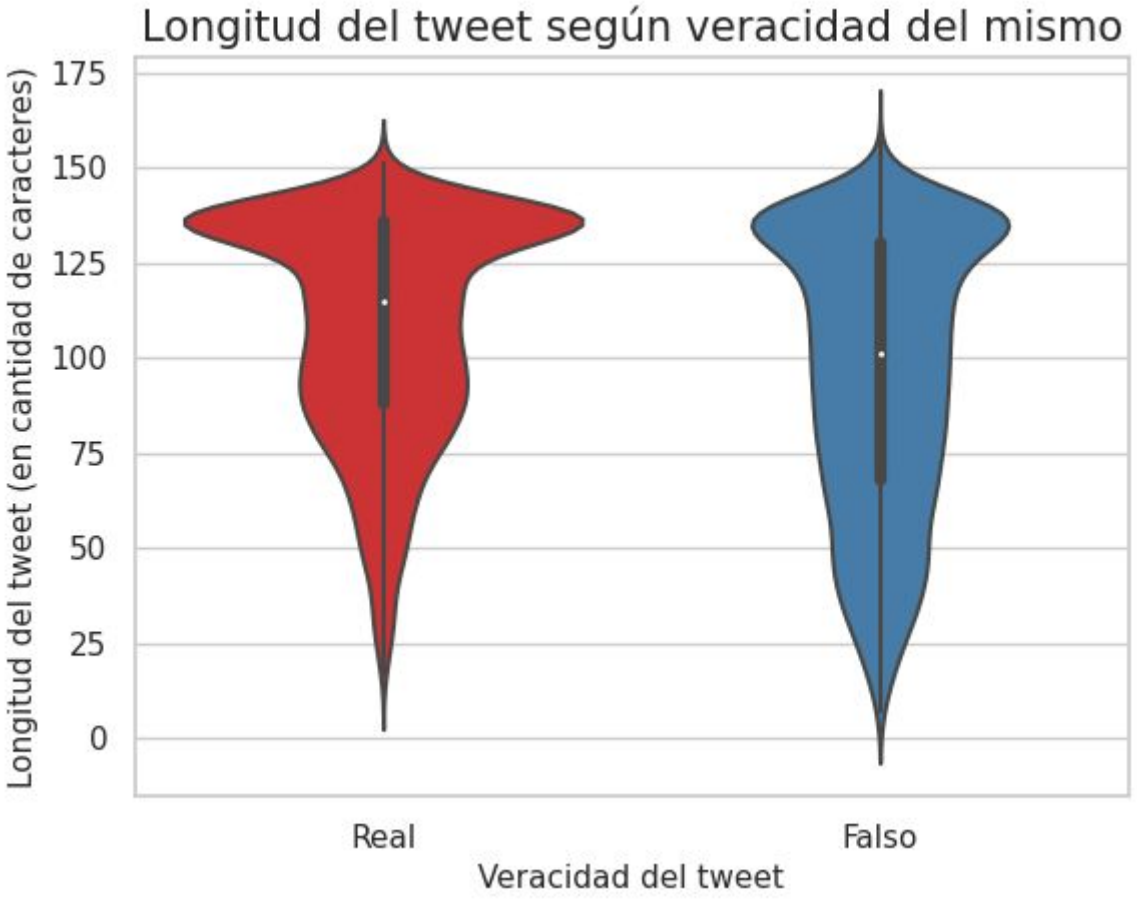
¿Cuántos verdaderos y falsos existen para cada longitud de texto?

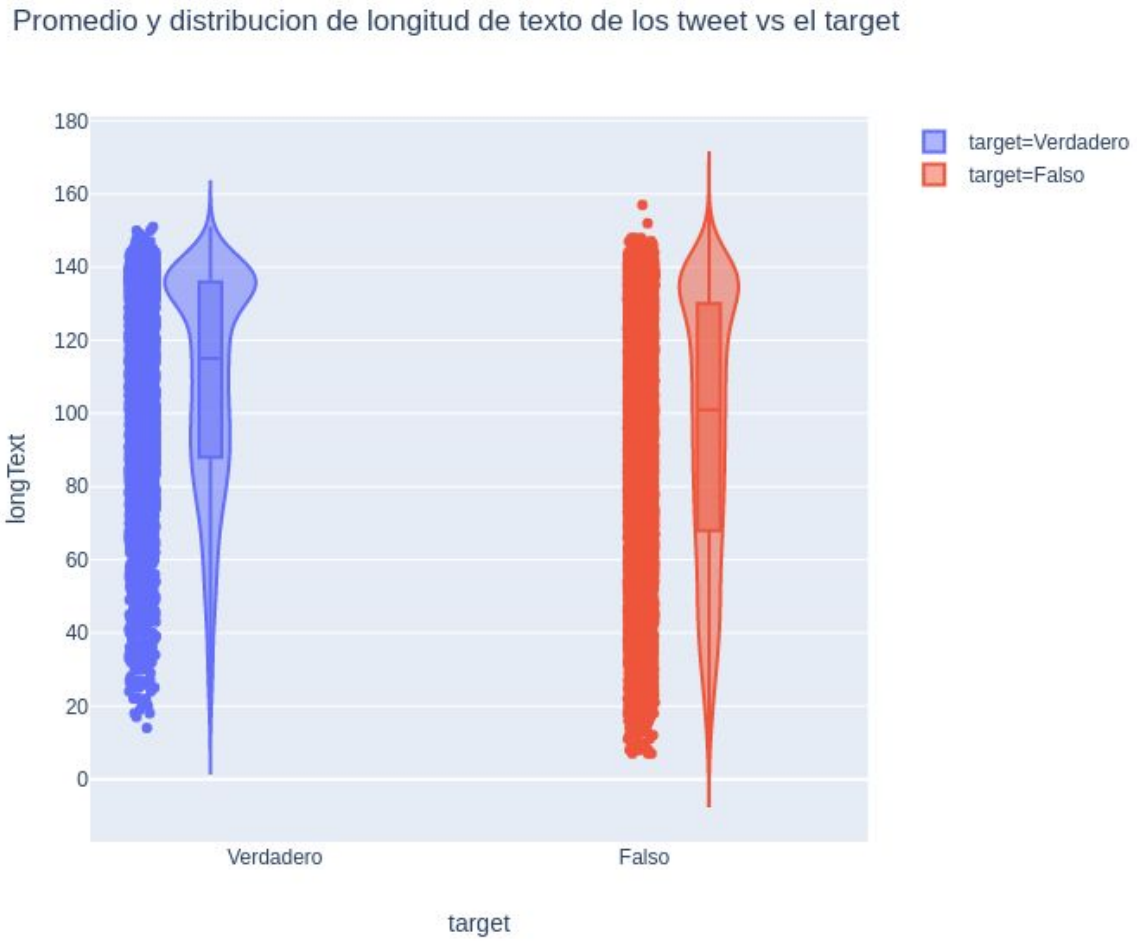


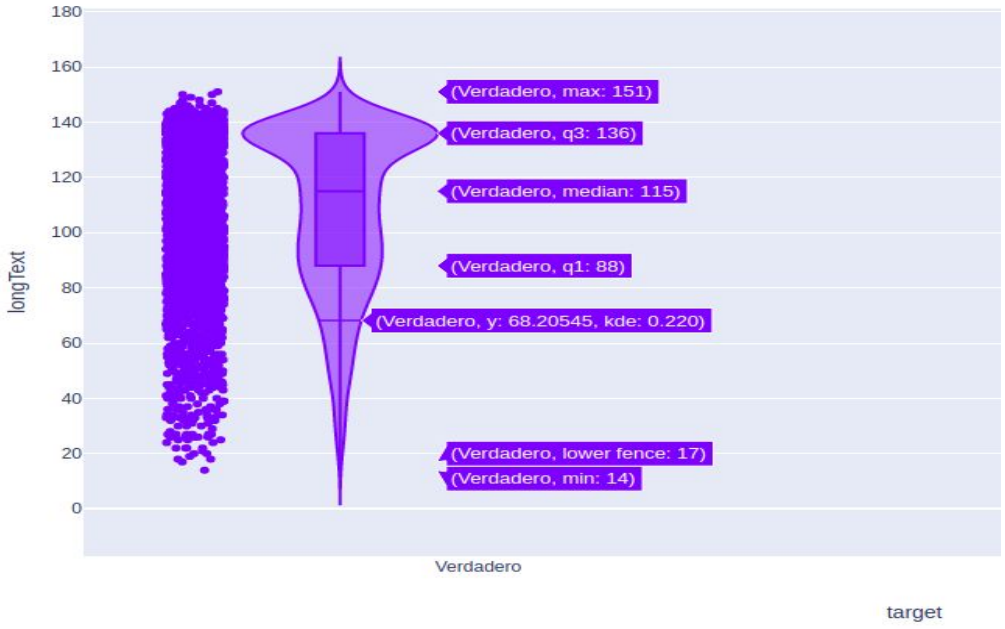
¿Cuál es la longitud del tweet según la veracidad del mismo?

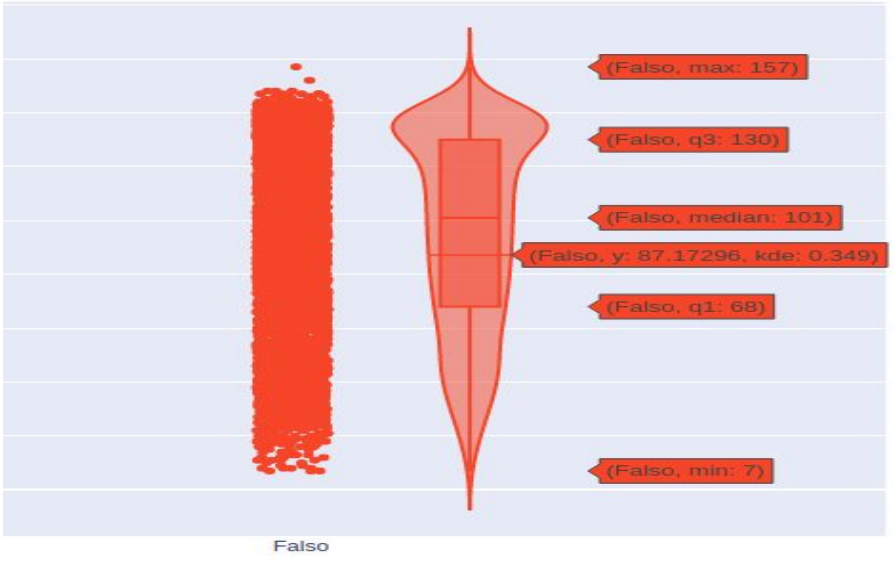
Como se observa en la siguiente visualización, no existe una relación directa entre la longitud del tweet y la veracidad del mismo. Sí podemos afirmar que los tweets reales tienden a ser más extensos que los tweets falsos, y que hay muchos más tweets cortos falsos, que tweets breves verdaderos. También es interesante que la mayoría de los tweets verdaderos tienen una longitud de entre 90 y 130 caracteres.

La distribución puntual por veracidad puede observarse en los últimos tres gráficos de esta sección.


Longitud del tweet según veracidad del mismo
Longitud del tweet (en cantidad de caracteres)
175
150
125
100
75
50
25
0
Real
Falso
Veracidad del tweet


Promedio y distribucion de longitud de texto de los tweet vs el target
target=Verdadero
target=Falso
longText
180
160
140
120
100
80
60
40
20
0
Verdadero
Falso
target





¿Cuáles son las palabras más relevantes de los tweets?

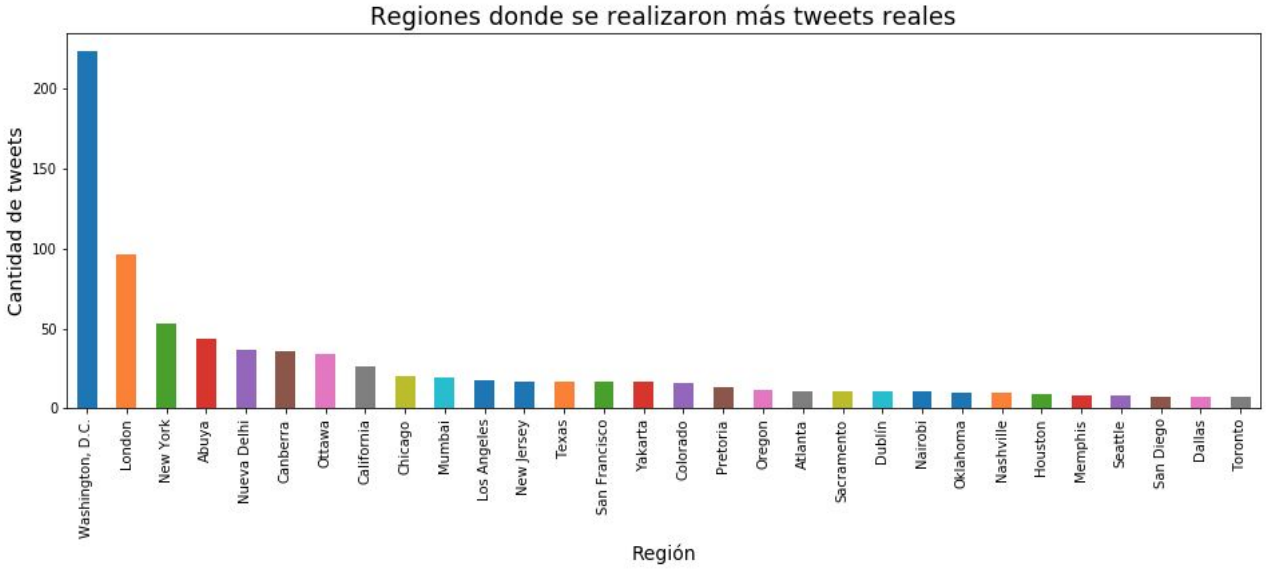
Para esta sección, se utilizó el campo ***text***, considerando irrelevantes a los signos de puntuación, y convirtiendo todo el texto a minúsculas. También se realizó un filtrado manual de palabras irrelevantes.

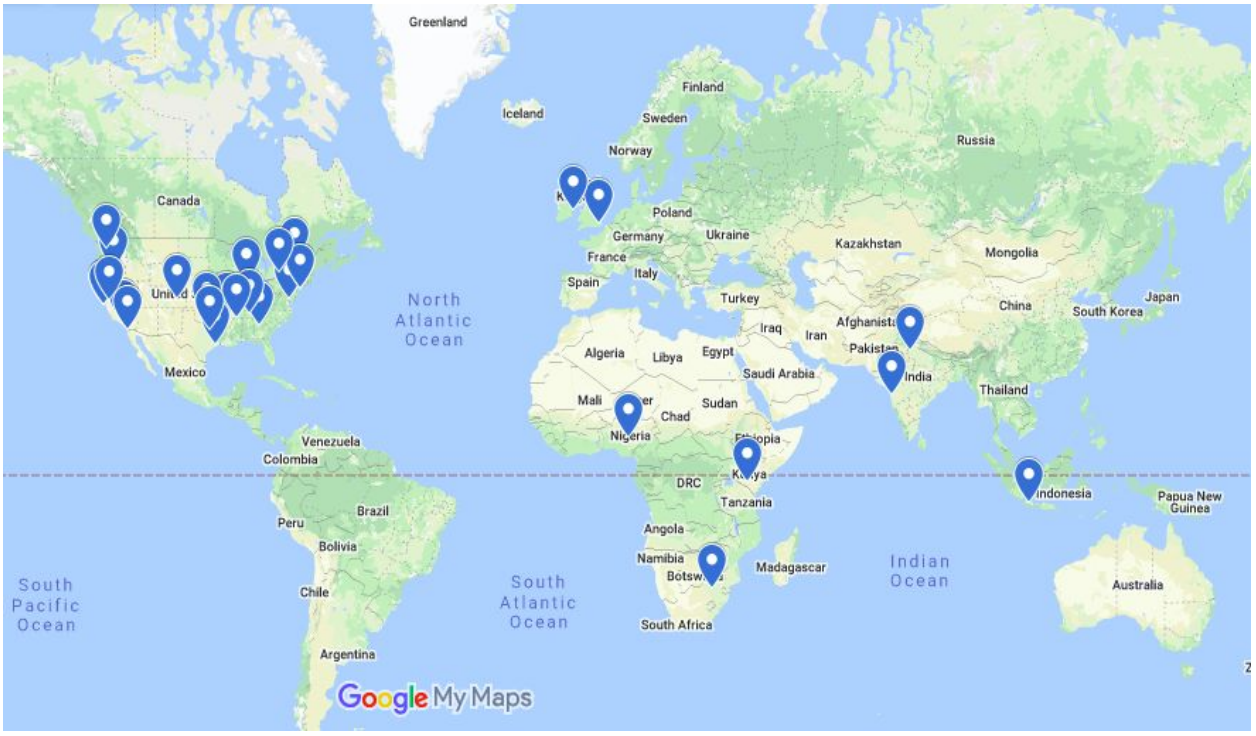
Con el resultado hemos generado el siguiente wordcloud.

¿Cuáles son las regiones donde se realizaron más tweets reales?

Se puede observar que la mayor concentración de publicaciones de los tweets son de Estados Unidos. El resto de las publicaciones está distribuida en los continentes de Europa, África y Asia.

Con una clara ventaja se aprecia que Washington, D.C. lidera el ranking.

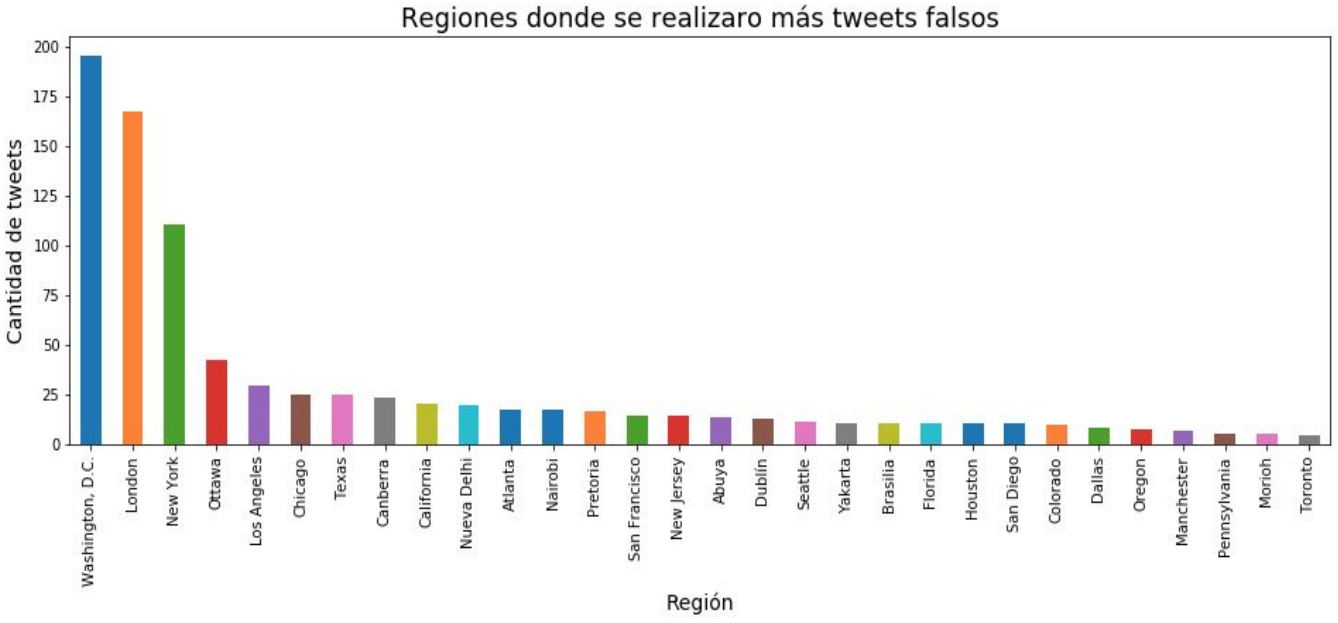


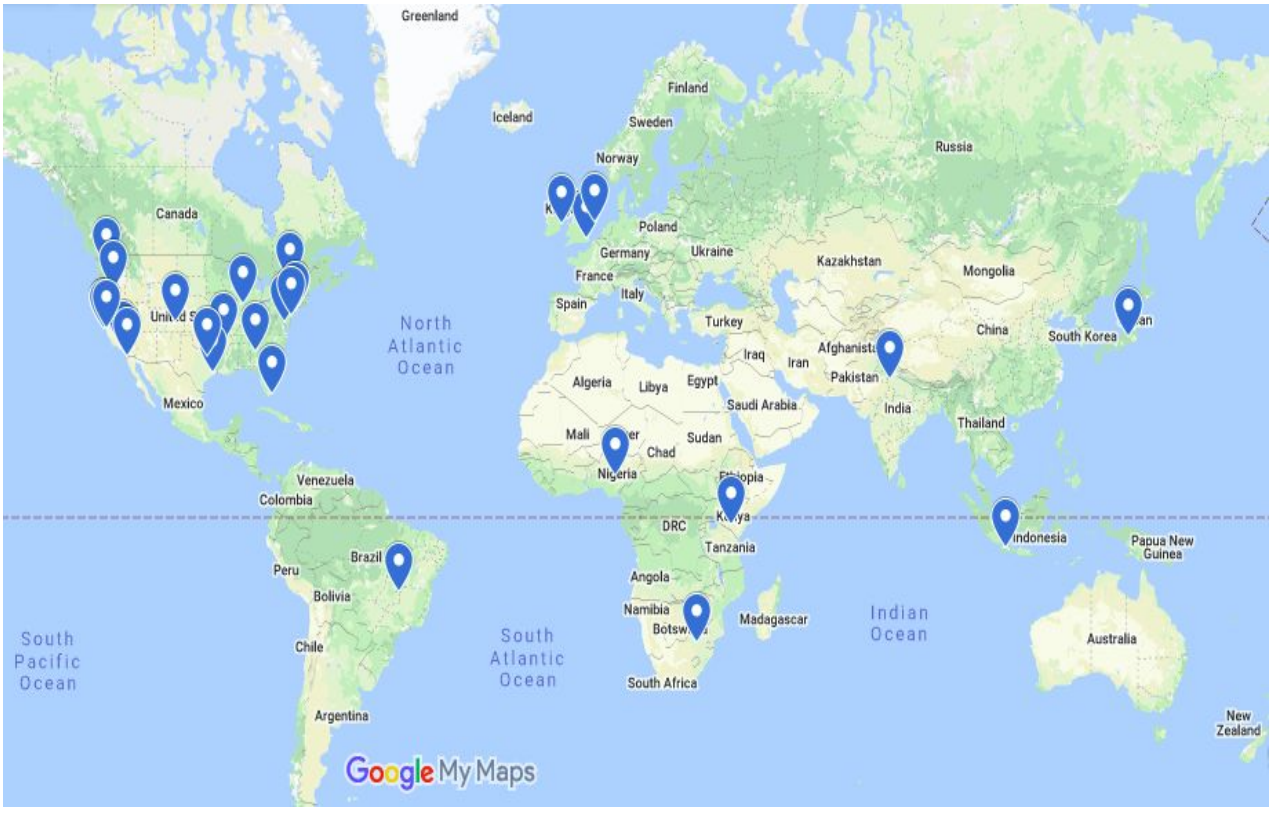


¿Cuáles son las regiones donde se realizaron más tweets falsos?

En este caso, se observa un ranking parejo entre Washington, D.C. y London, aunque Washington, D.C. mantiene el liderazgo de publicaciones.

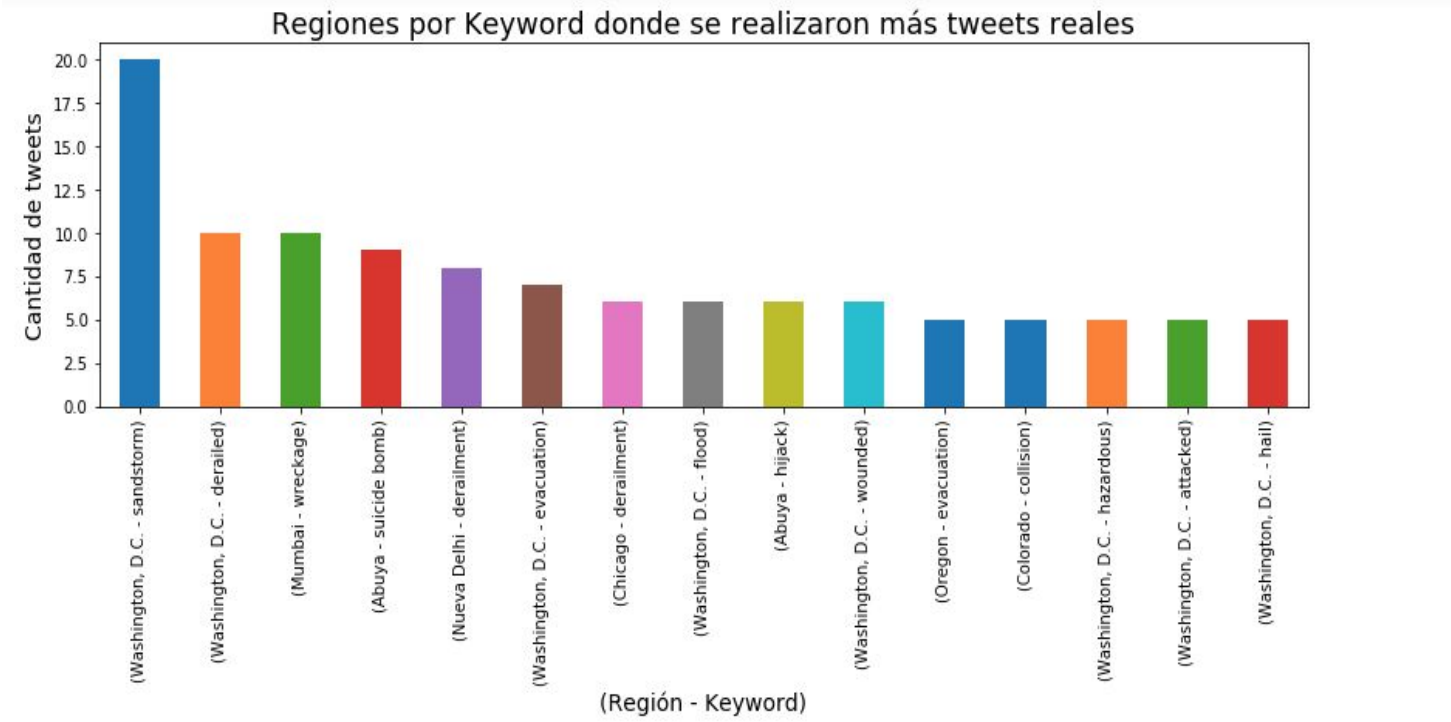
La distribución de publicaciones entre los tweets reales y falsos es muy parecida.





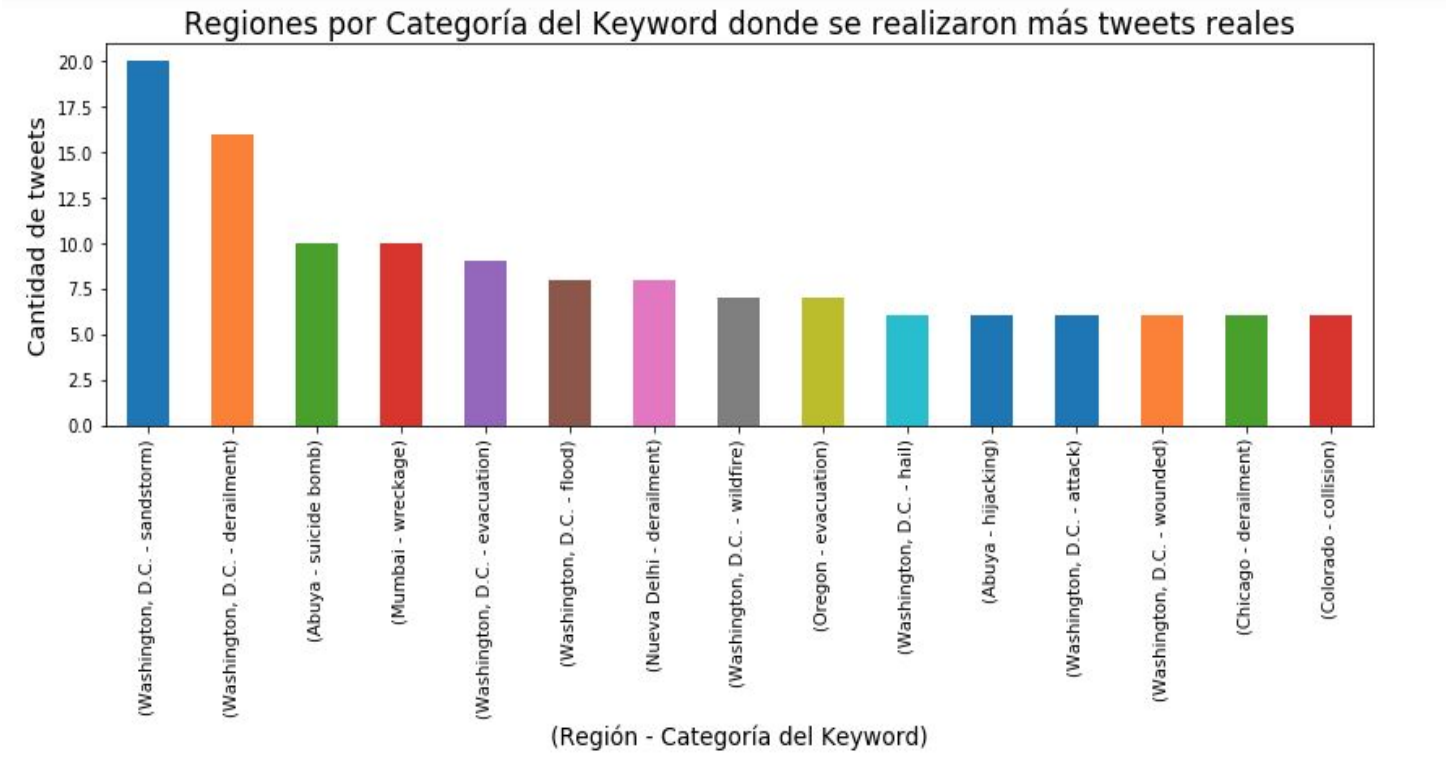
¿Cuáles son las regiones por keyword donde se realizaron más tweets reales?

Observamos que la regiones que predominan son de los países de Estados Unidos, Nigeria, Kenia y la India.
Donde sigue siendo Washington, D.C. el que tiene más publicaciones del tipo “sandstorm” como principal keyword.



¿Y por categoría del keyword?

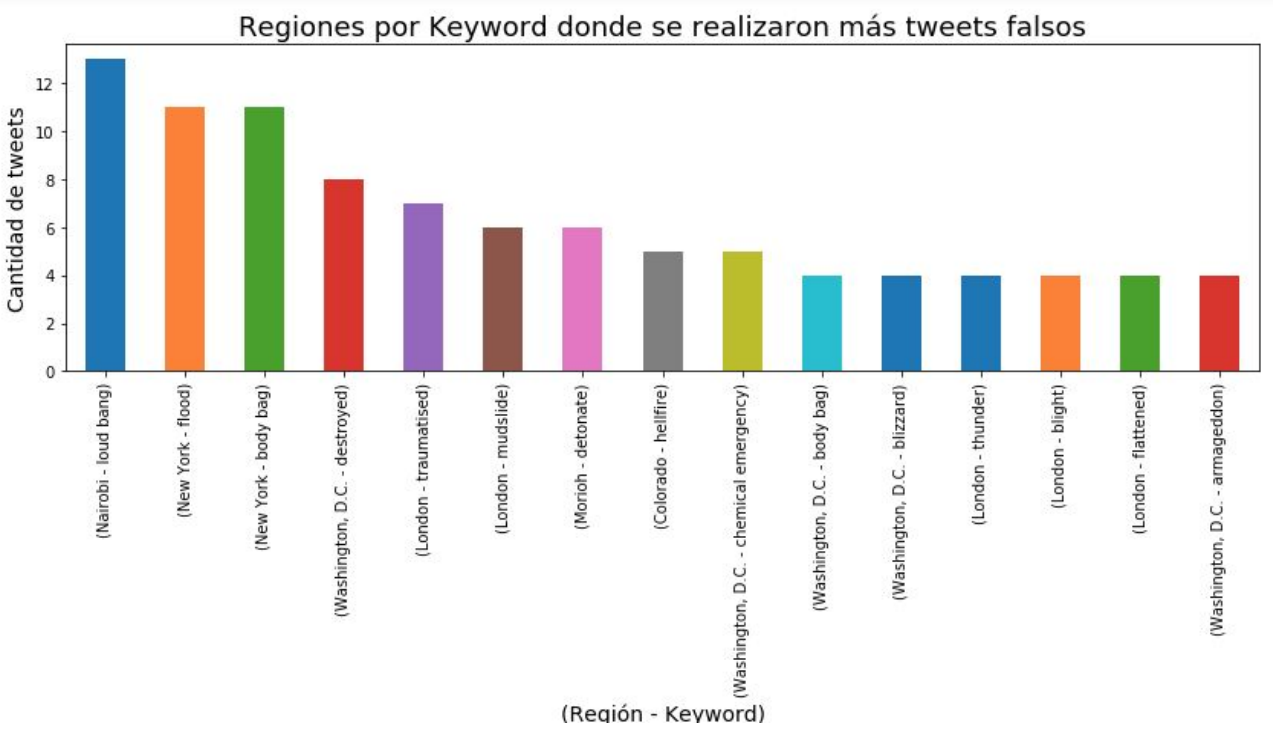
Se mantiene casi la misma distribución de publicaciones.



¿Cuáles son las regiones por keyword donde se realizaron más tweets falsos?

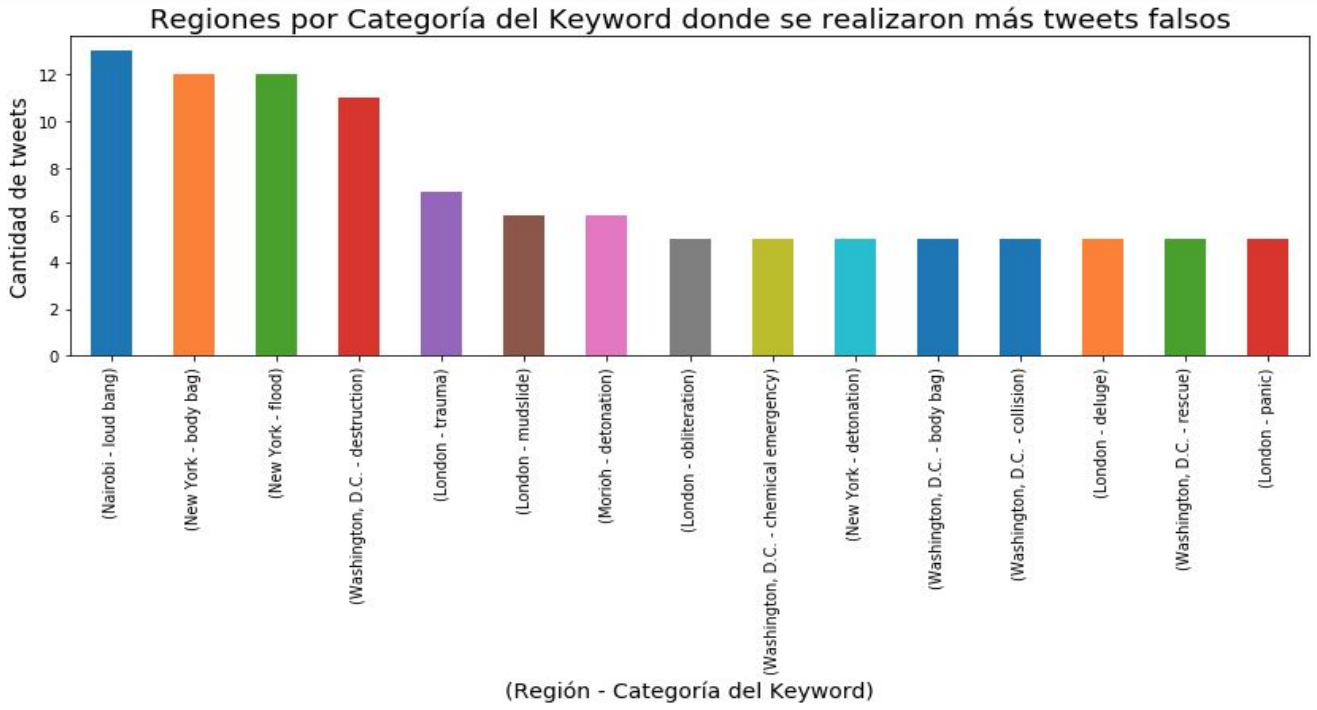
Observamos que predominan los mismo países.

En este caso, el que lidera el ranking es Nairobi, con más publicaciones del tipo “loud bang” como principal keyword.



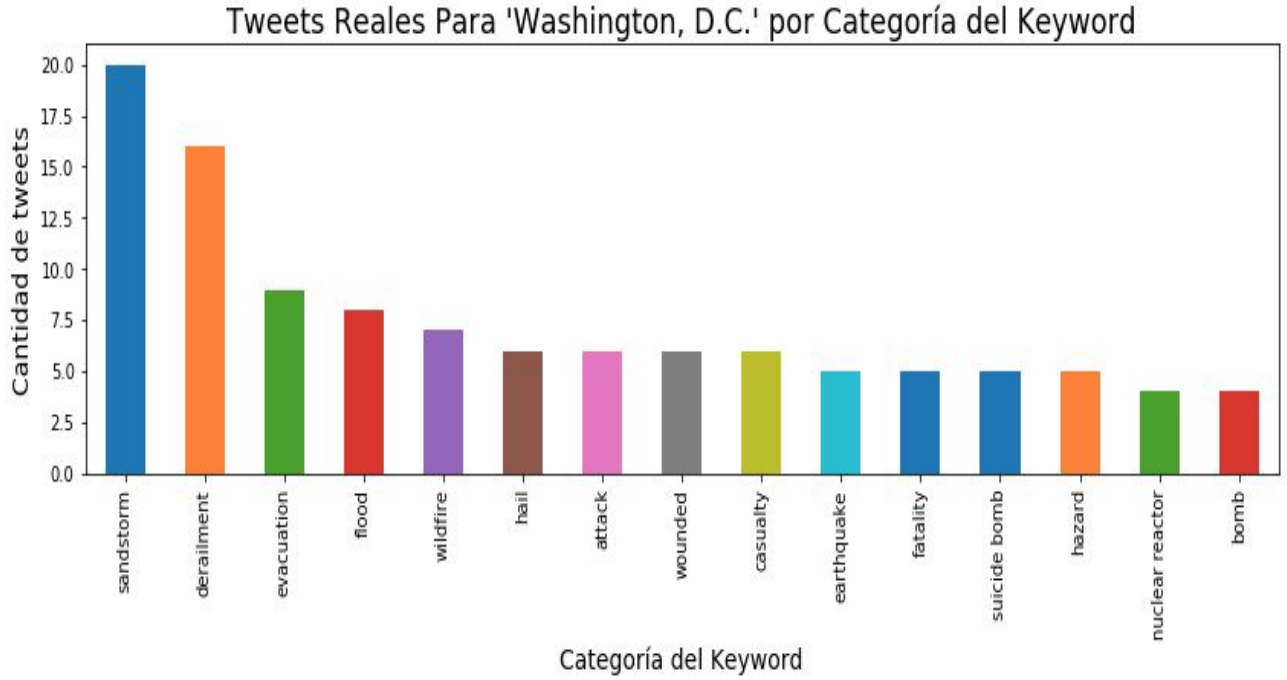
¿Y por categoría del keyword?

Se mantiene casi la misma distribución de publicaciones.



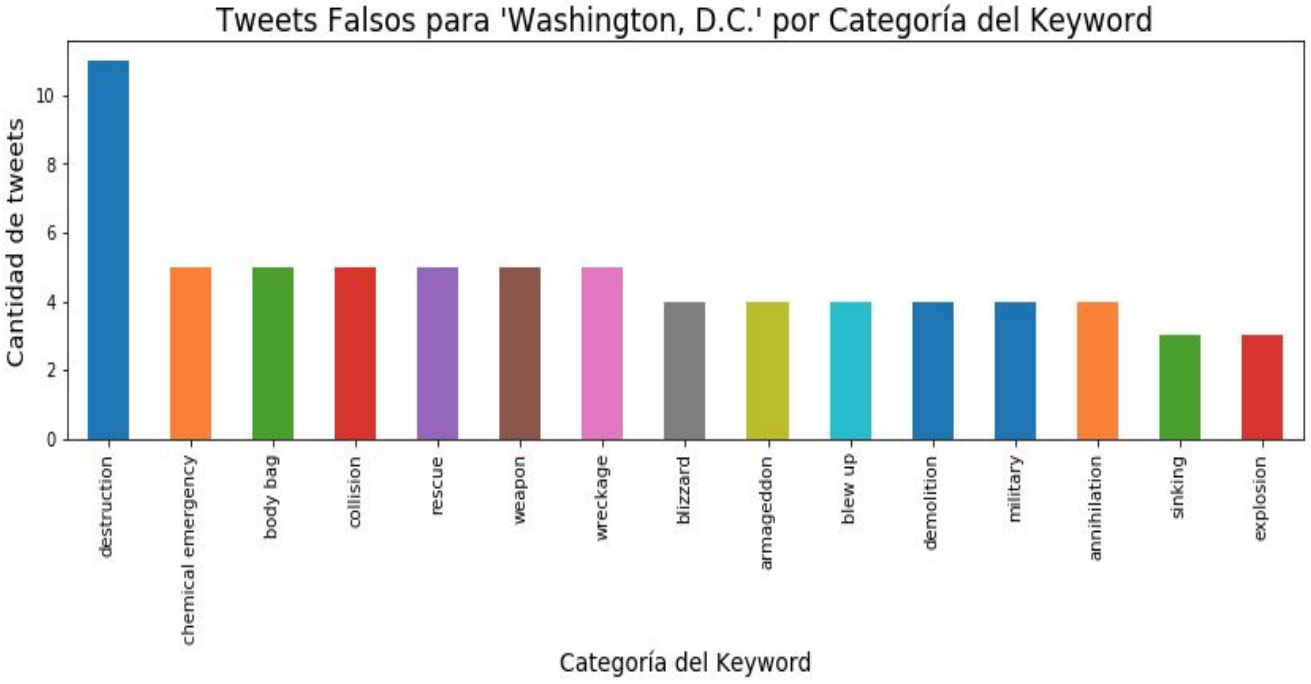
¿Cómo se distribuyen las categorías del keyword para la región donde se realizaron más tweets Reales (Washington D.C.) ?

Como se observó anteriormente, “sandstorm” es la categoría que predomina para Washington D.C.

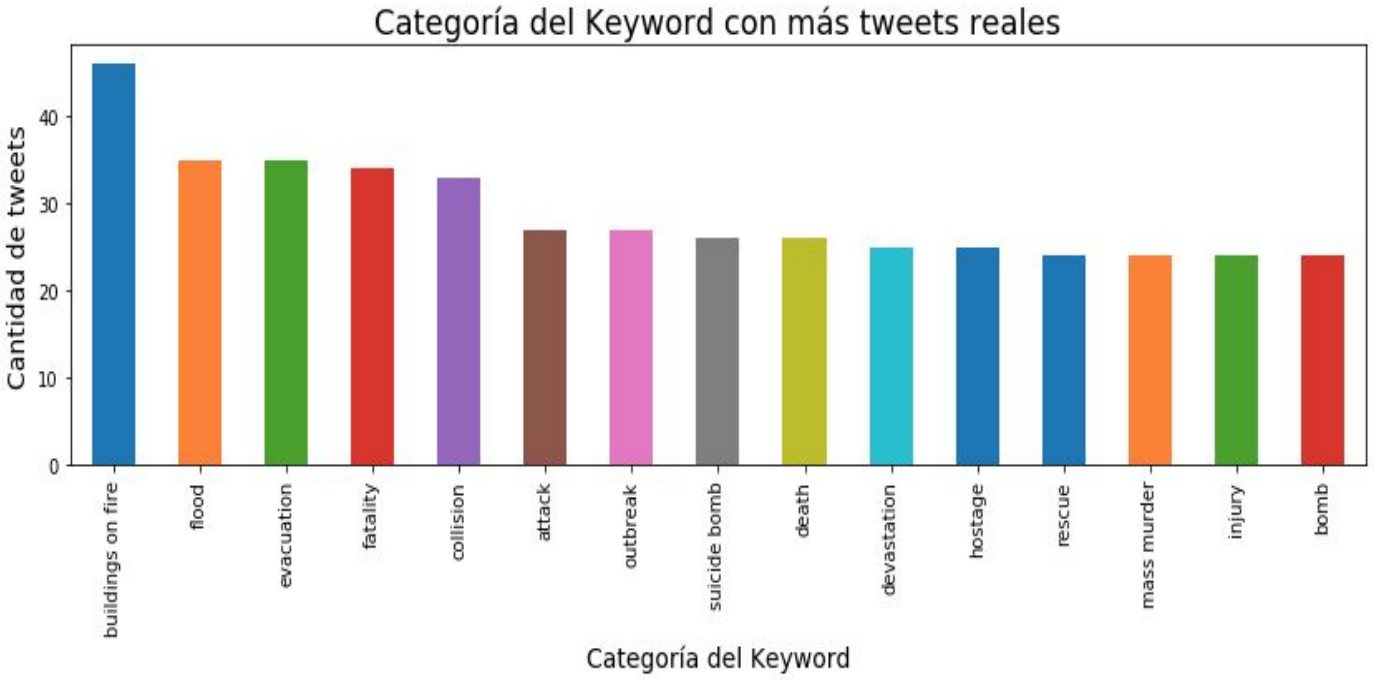


¿Cómo se distribuyen las categorías del keyword para la región donde se realizaron más tweets Falsos (Washington D.C.) ?

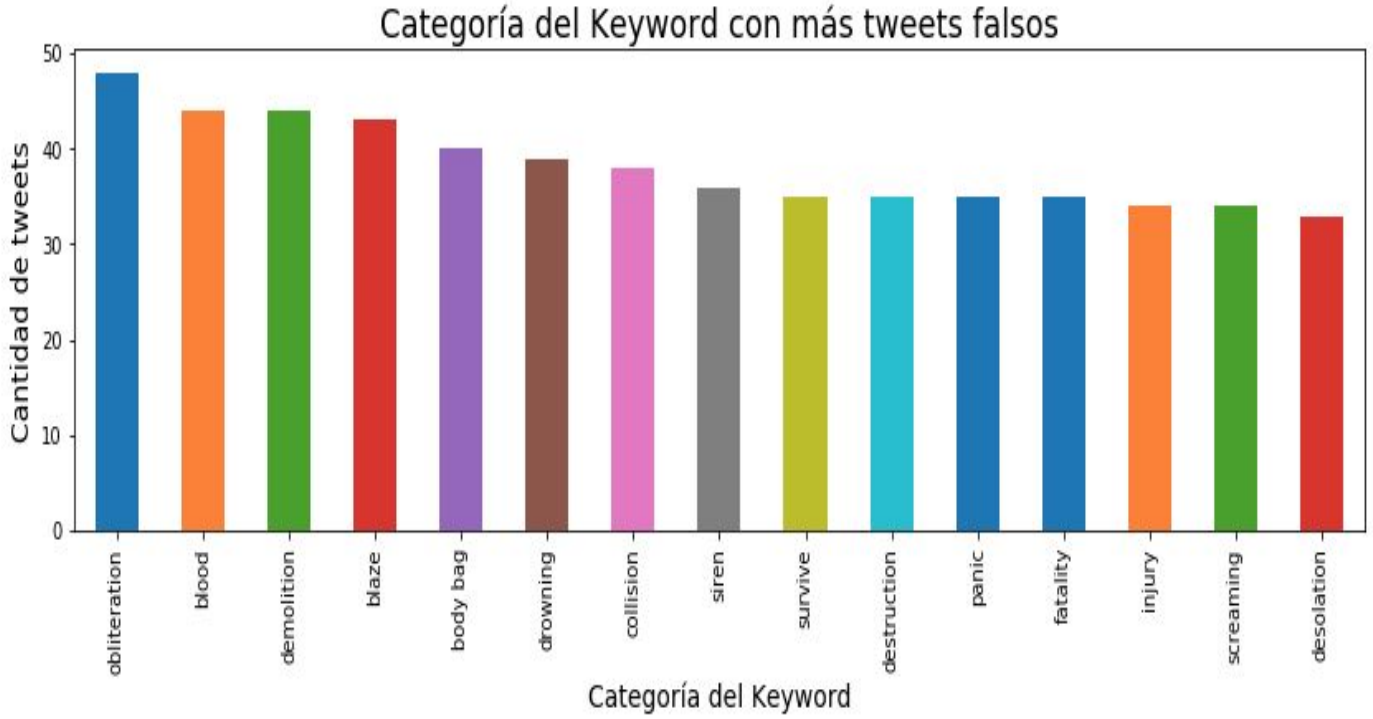
En este caso se observa, que la categoría predominante es “destruction” y “sandstorm” queda fuera del ranking.



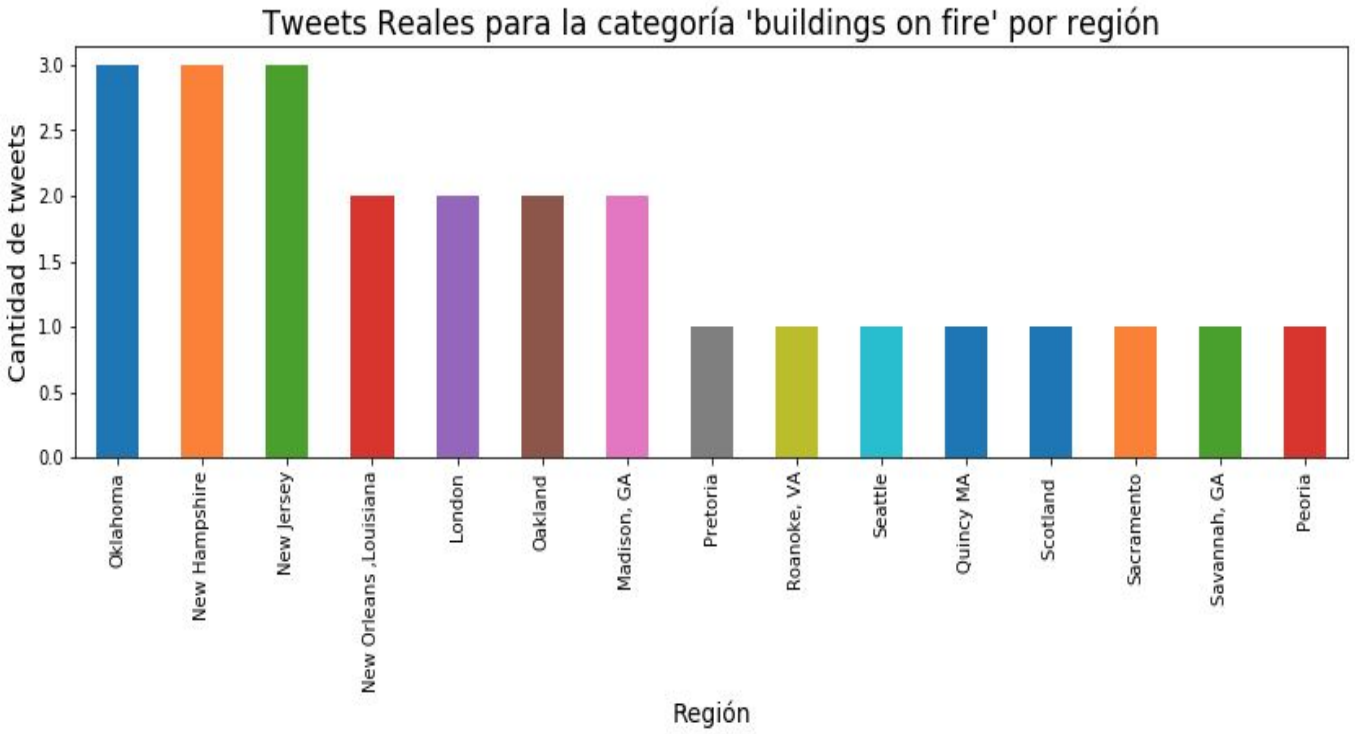
¿Cuáles son las categorías del keyword con más tweets reales para todas las regiones?

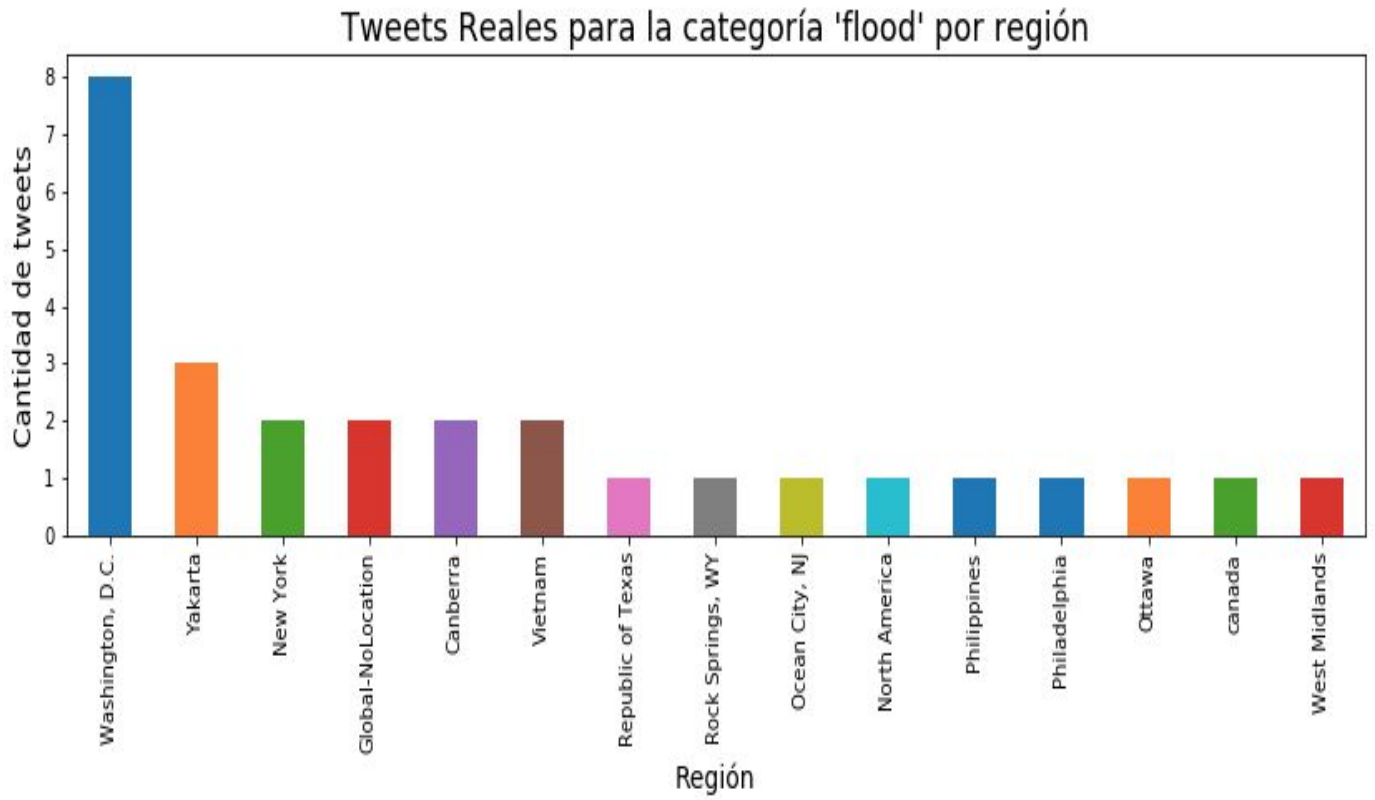


¿Cuáles son los categorías del keyword con más tweets falsos para todas las regiones?

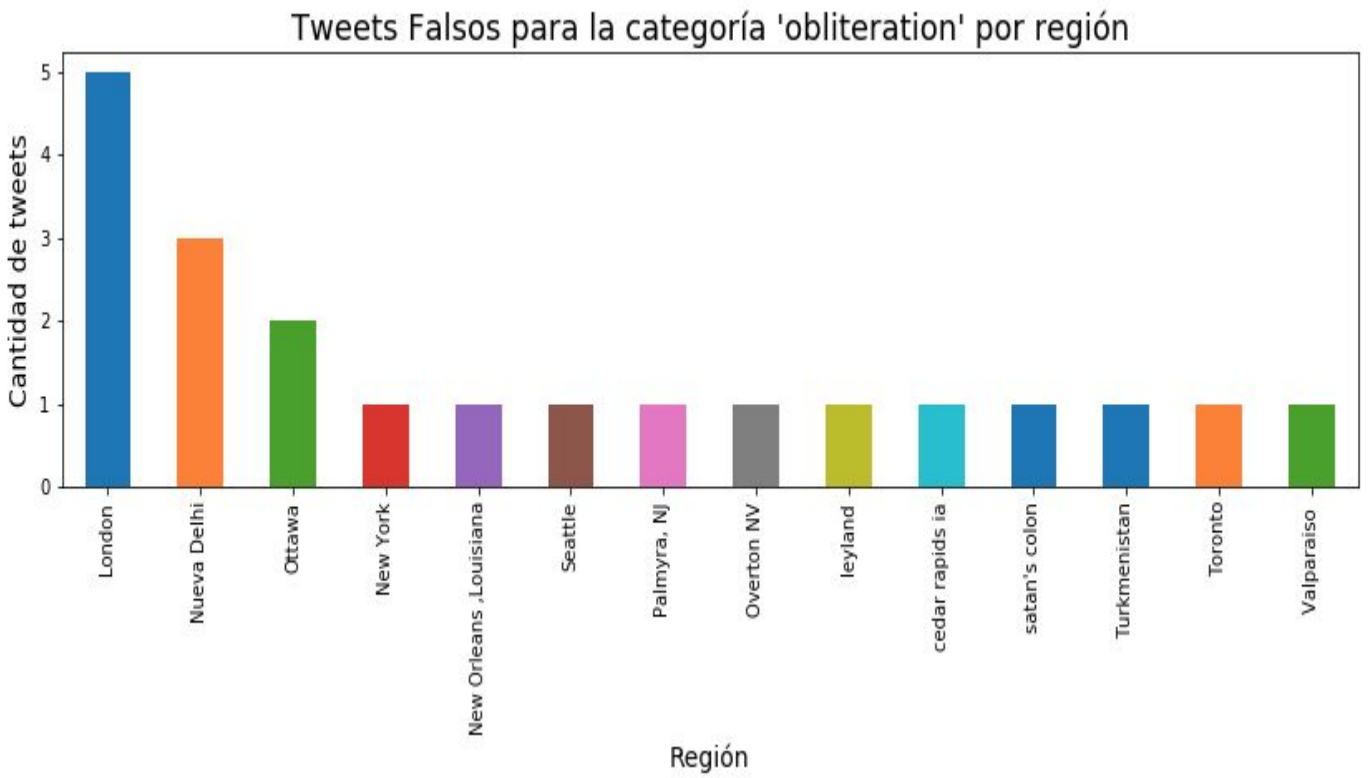


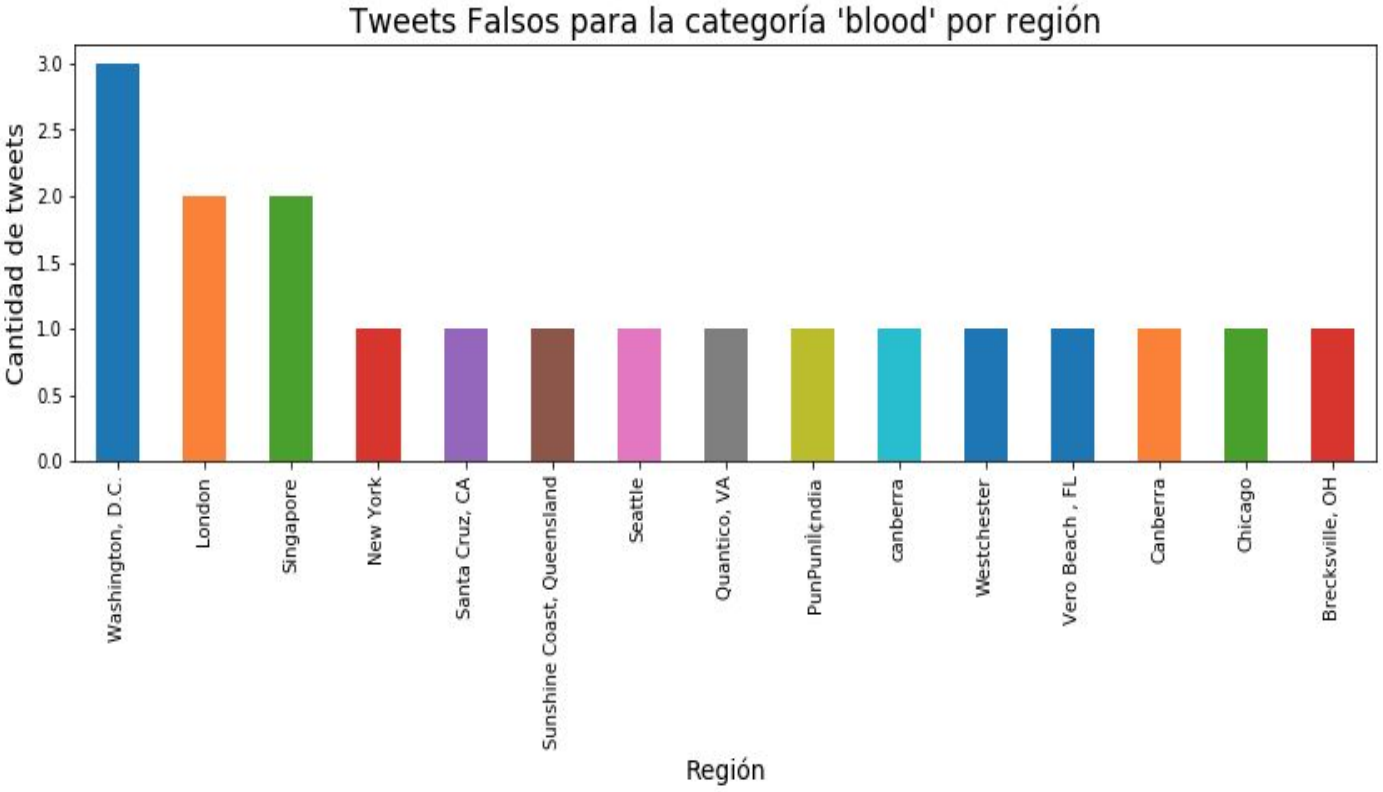
¿Cómo se distribuyen las regiones para las categorías del keyword donde se realizaron más tweets Reales ('buildings on fire' y 'flood')?





¿Cómo se distribuyen las regiones para las categorías del keyword donde se realizaron más tweets Falsos ('obliteration' y 'blood')?


Tweets Falsos para la categoría 'obliteration' por región
Cantidad de tweets
0
1
2
3
4
5
London
Nueva Delhi
Ottawa
New York
New Orleans ,Louisiana
Seattle
Palmyra, NJ
Overton NV
leyland
cedar rapids ia
satan's colon
Turkmenistan
Toronto
Valparaiso
Región


Tweets Falsos para la categoría 'blood' por región
Cantidad de tweets
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Washington, D.C.
London
Singapore
New York
Santa Cruz, CA
Sunshine Coast, Queensland
Seattle
Quantico, VA
PunPunil¢ndia
canberra
Westchester
Vero Beach , FL
Canberra
Chicago
Brecksville, OH
Región

Conclusiones

Basándonos en las geo-visualizaciones, podemos observar que la mayor concentración de tweets, tanto reales como falsos, está en Estados Unidos. No existe una gran diferencia entre las cantidades de tweets verdaderos y falsos. El resto de las publicaciones está distribuida en los continentes de Europa, África y Asia. En cuanto al continente americano, podemos suponer que si el tweet proviene de América del Norte, su veracidad puede ser tanto real como falsa, pero si es de América del Sur, es muy probable que sea falso.

Quedándonos con un Top 10, tenemos para los tweets reales las ciudades de: Washington, D.C., London, New York, Abuya, Nueva Delhi, Canberra, Ottawa, California, Chicago y Mumbai. Y para los tweets falsos: Washington, D.C., London, New York, Ottawa, Los Angeles, Chicago, Texas, Canberra, California, Nueva Delhi. Ambos en sus correspondiente orden.

Según el análisis de las regiones con los distintos tipos de keyword con su categoría correspondiente, observamos que la regiones donde se realizan más tweets reales predominan los países de Estados Unidos, Nigeria, Kenia y la India.
Donde la mayor concentración sigue estando en Estados Unidos, más particularmente en Washington D.C. (que lidera el ranking) con la categoría-keyword que corresponde a “sandstorm” para los tweets reales, y “destruction” para los tweets falsos. La zona de Eastern Washington es una zona propensa a sufrir tormentas de viento muy fuertes durante el otoño. En cuanto a la categoría del keyword en los tweets falsos lidera Nairobi con “loud bang”.

Sin embargo, si consideramos solo la categoría del keyword con más tweets reales corresponde a “buildings on fire”, esto es asi por que se considera para todas las regiones. Si analizamos qué región pertenece podemos visualizar que predomina más en la región Oklahoma. Para la categoría-keyword con más tweets falsos corresponde a “obliteration”, nuevamente esto es así porque se considera todas las regiones. Si analizamos qué región tiene más casos de dicha categoría, corresponde a London.

Cuando analizamos los keywords más usados en los tweets, encontramos que corresponde “fatalities”. Pero si ahondamos y analizamos su veracidad, este queda excluido, ya que el keyword con mayor veracidad corresponde a “evacuation”, y el keyword con menor veracidad corresponde a “blaze”. Por otro lado, es esperable cuando analizamos los keywords por categoría, dichas keyword estén dentro del ranking.