FACULTAD
DE INGENIERIA
Universidad de Buenos Aires

75.06 Organización de datos

Trabajo Práctico 1
Análisis Exploratorio de Datos

Integrantes del grupo:

Nombre	Mail
Germán Sequeira Wolf	gsequeira@fi.uba.ar
Claudio Collado	cjccollado@gmail.com
Celeste Cingolani	mcelestecingolani@gmail.com

Repositorio de github:

https://github.com/tporganizaciondedatos1c2020/TP1

1º cuatrimestre 2020

1. Introducción

El presente trabajo práctico consiste en realizar el análisis exploratorio de datos (EDA) de los tweets del set de datos 'train' de la competencia:
https://www.kaggle.com/c/nlp-getting-started

2. Características generales del Dataset

¿Cual es la estructura general y características principales del set de datos en análisis?

A continuación se describen características relevantes del set de datos en estudio:

A. El set de datos se encuentra formado por 7613 filas
B. El set de datos posee como atributos las siguientes columnas:

    a. **Id** - identificador único para cada tweet
    b. **Keyword** - un keyword para el tweet
    c. **Location** - ubicación desde donde fue enviado
    d. **Text** - el texto del tweet
    e. **Target** indica si se trata de un desastre real (1) o no (0)

C. En la Figura N°1 se indica para cada columna el tipo de dato identificado por Pandas al momento de la lectura del csv

Columna	Tipo de dato
id	int64
keyword	object
location	object
text	object
target	int64

Figura N°1: Tipo de datos por columna

Según se observa se tienen 2 columnas con tipo de datos *int64* y 3 columnas con tipo de dato *object*.

3. Análisis de valores nulos

¿Que cantidad de nulos tiene el set de datos?

En la Figura N°2 se observa un resumen de datos nulos (NaN) para todas las columnas del set de datos.

	Columna	¿Tiene NaN?	Cantidad de NaN	% de NaN
0	id	False	0	0.0
1	keyword	True	61	0.8
2	location	True	2533	33.3
3	text	False	0	0.0
4	target	False	0	0.0

Figura N°2: Nulos por Columna

Según se observa en particular la columna *'location'* tiene un % de NaN considerable y deberá analizarse a posteriori si esto representa un condicionante para análisis a realizar en el TP N°2
Además se puede verificar que no existe ningún registro que contenga el string 'NaN', esto debe tenerse en cuenta para considerar la posibilidad de llenar los registros nulos con el string antes mencionado para evitar incongruencias a la hora de agrupar datos u otras operaciones similares.

4. Análisis de tweets repetidos

¿Existen tweets repetidos?

En la Figura N°3 se observan en forma de tabla todos los tweets que son idénticos junto con su cantidad distribuida según su veracidad. En la Figura Nº4 graficamos la relación entre los keywords de los tweets y la frecuencia con la que aparecen en la tabla.
Finalmente en la Figura Nº5 analizamos de manera gráfica los tweets repetidos según su ubicación (la cual es predominantemente nula) y veracidad.
En total tenemos 97 registros repetidos con un total de 35 textos distintos

En las Figuras Nº6 y N°7 se puede apreciar la distribución tanto de la cantidad de palabras en los tweets como la cantidad de caracteres de estos.

	keyword	location	text	longitud	Falso	Verdadero	Total
75	ablaze	Live On Webcam	Check these out: http://t.co/rOI2NSmEJJ http:/...	114	2	0	2
157	aftershock	Switzerland	320 [IR] ICEMOON [AFTERSHOCK] | http://t.co/TH...	138	2	0	2
159	aftershock	US	320 [IR] ICEMOON [AFTERSHOCK] | http://t.co/vA...	138	2	0	2
179	airplane%20accident	NaN	Experts in France begin examining airplane deb...	136	0	2	2
653	bioterrorism	NaN	To fight bioterrorism sir.	26	2	2	4
2373	demolition	NaN	General News Û¢åÊ'Demolition of houses on wat...	137	2	0	2
2470	derailment	India	Madhya Pradesh Train Derailment: Village Youth...	63	0	2	2
2488	derailment	NaN	Madhya Pradesh Train Derailment: Village Youth...	63	0	2	2
2489	derailment	NaN	Madhya Pradesh Train Derailment: Village Youth...	136	0	3	3
2674	detonate	Morioh, Japan	@TinyJecht Are you another Stand-user? If you ...	99	3	0	3
2675	detonate	Morioh, Japan	@spinningbot Are you another Stand-user? If yo...	101	3	0	3
2838	displaced	NaN	#KCA #VoteJKT48ID 12News: UPDATE: A family of ...	141	0	2	2
2852	displaced	Pedophile hunting ground	#Myanmar Displaced #Rohingya at #Sittwe point...	136	0	2	2
2853	displaced	Pedophile hunting ground	.POTUS #StrategicPatience is a strategy for #G...	134	1	3	4
3262	engulfed	NaN	He came to a land which was engulfed in tribal...	123	4	0	4
3463	exploded	NaN	that exploded & brought about the\r\nbegin...	140	2	0	2
3595	fatal	NaN	11-Year-Old Boy Charged With Manslaughter of T...	136	0	6	6
4002	floods	NaN	Who is bringing the tornadoes and floods. Who ...	139	2	1	3
4226	hazardous	NaN	Caution: breathing may be hazardous to your he...	51	1	1	2
4295	hellfire	NaN	Beware of your temper and a loose tongue! Thes...	120	2	0	2
4296	hellfire	NaN	Hellfire is surrounded by desires so be carefu...	100	2	1	3
4299	hellfire	NaN	Hellfire! We donÛªt even want to think about ...	107	2	0	2
4300	hellfire	NaN	The Prophet (peace be upon him) said 'Save you...	114	3	2	5

Figura N°3: Tweets repetidos

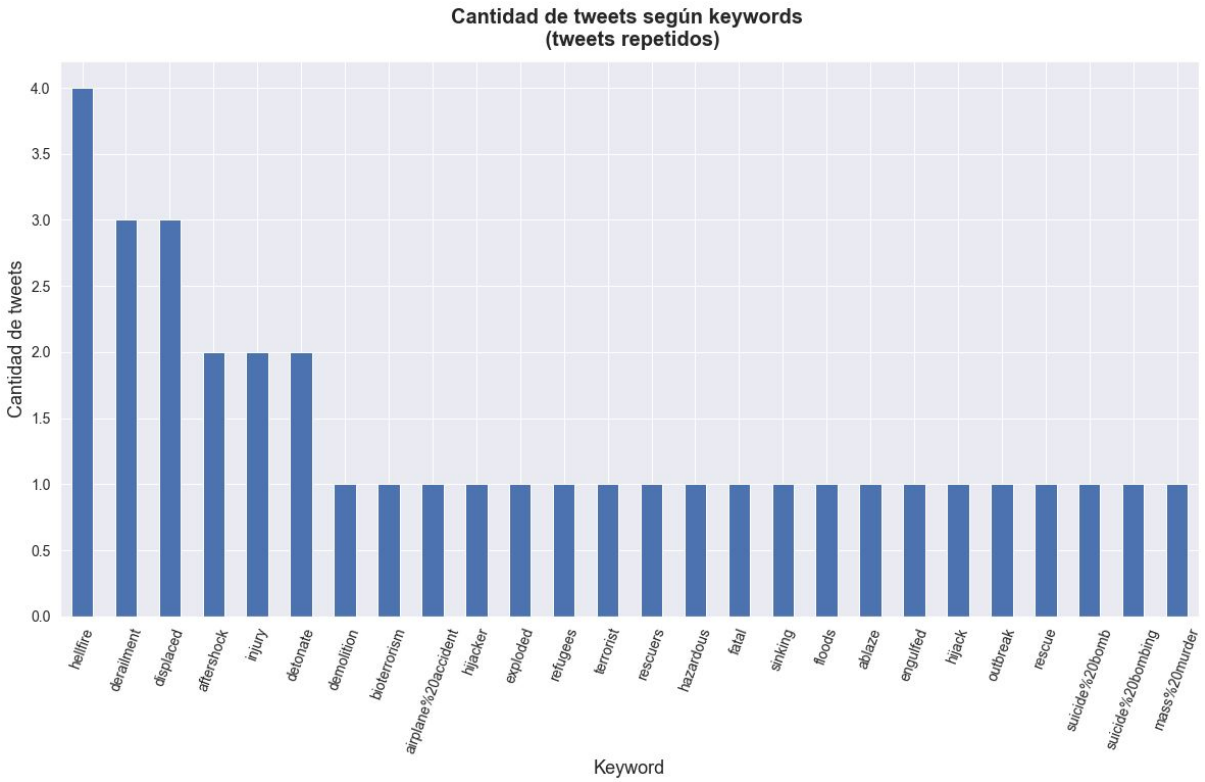

Figura N°4: Tweets repetidos según Keyword

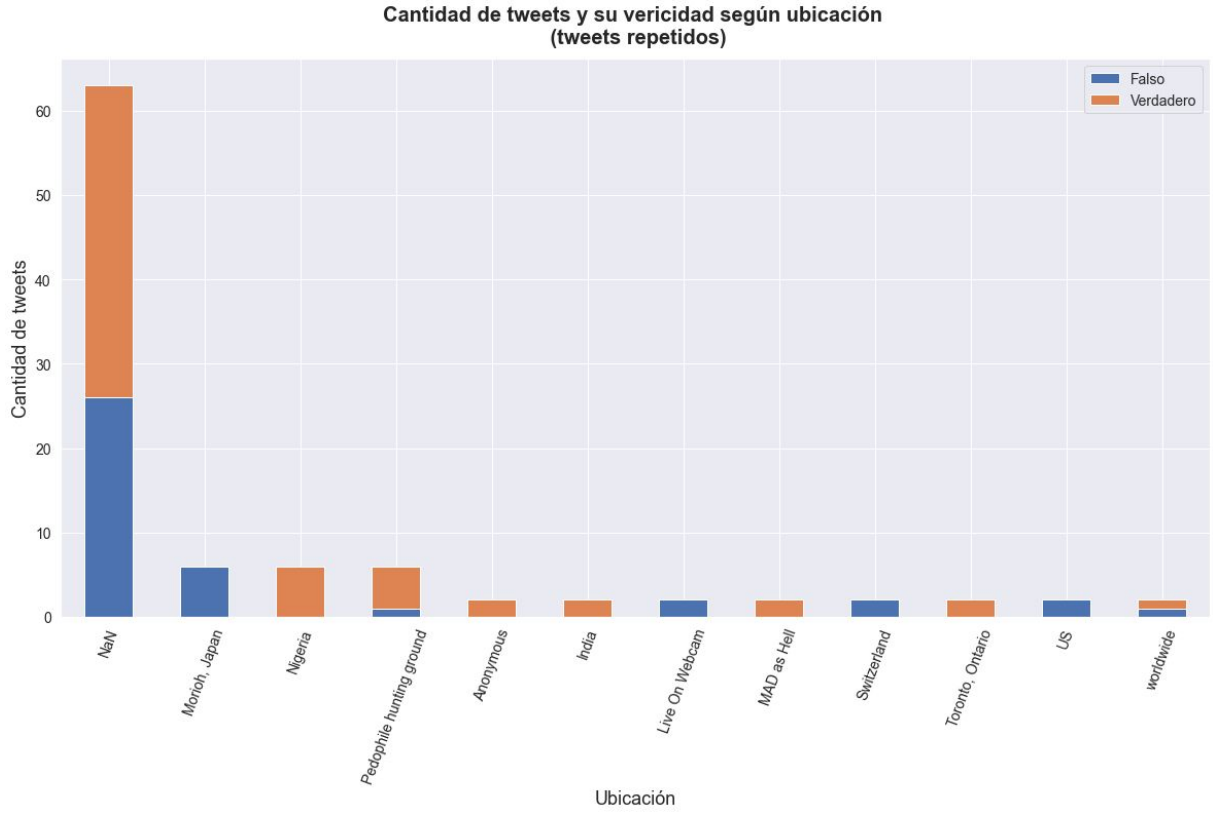

Figura N°5: Tweets repetidos según veracidad y ubicación

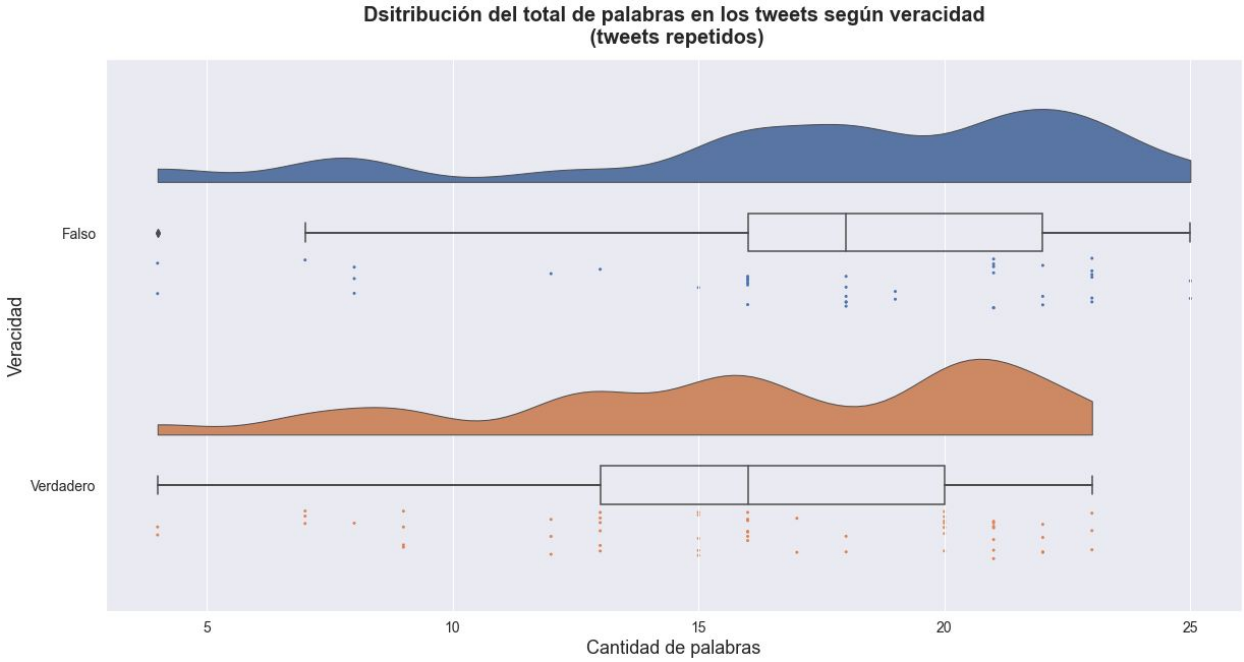

Figura N° 6: Distribución del total de palabras de los Tweets repetidos según veracidad

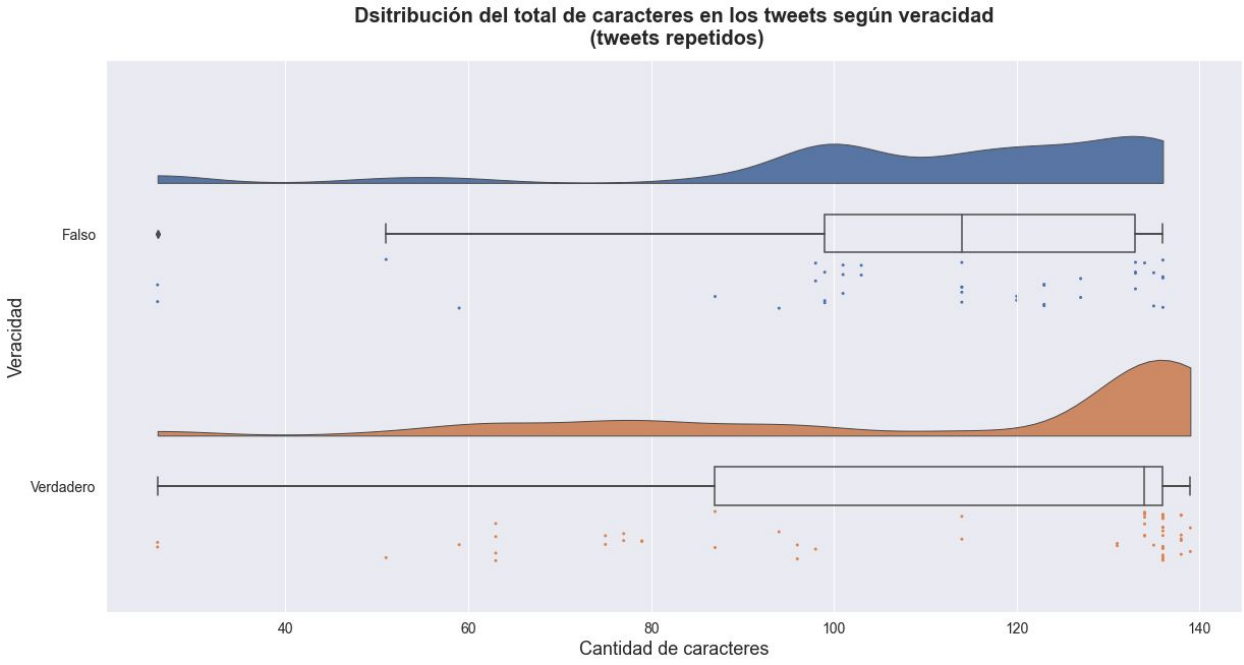

Figura N°7: Distribución del total de caracteres de los Tweets repetidos según veracidad

5. Características generales de Columnas

5.1 Análisis de la Columna **target**

¿Como se encuentra constituida la columna 'target'?

En esta columna se incluye la clasificación del tweet en función de su veracidad o no del mismo:

- Tweet Falso: Corresponde al valor de Target 0
- Tweet Verdadero: Corresponde al valor de Target 1

En la Figura N°8 se observan los porcentajes correspondientes a cada categoría. Según se observa existe una mayor presencia de tweets clasificados como Falsos.

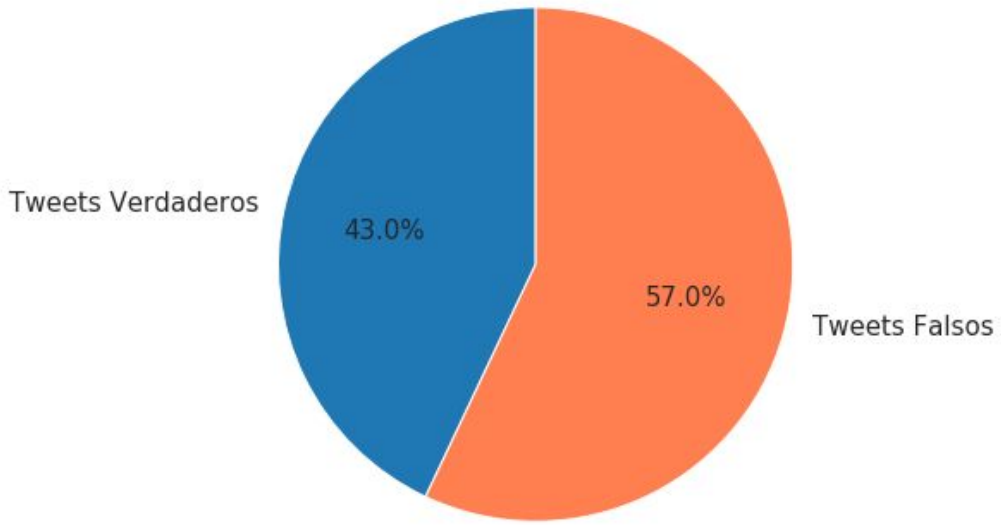

Figura N°8: Tweets Verdaderos/Falsos

5.2 Análisis de la Columna **keyword**

Según la descripción incluida en la página de twitter las keywords (palabras claves) son utilizadas con fines de segmentación.

5.2.1 Análisis de repeticiones

¿Cuantas keywords son unicas? ¿Cuales son las keywords que mayor se repiten?

Con respecto a los datos incluidos en esta columna se identifican 221 keywords diferentes.
En la Figura N°9 se identifican las 20 keywords que más se repiten. Se observa que no existe una prevalencia en la presencia de alguna keyword en particular, sino que la distribución de las mismas presenta cierta uniformidad para estas 20 primeras.

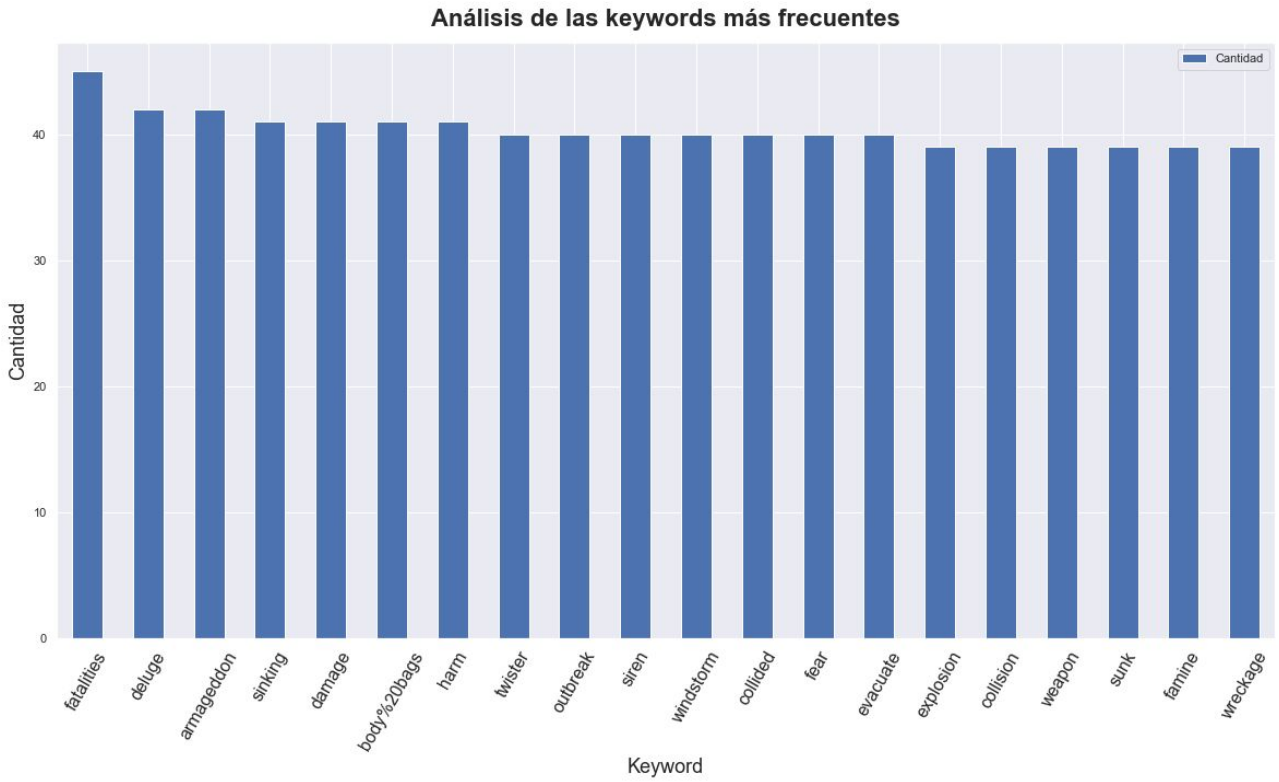

Figura N°9: Keywords más frecuentes

5.2.2 Análisis de contenido - Presencia del símbolo “%20”

¿Las keywords son solo texto o presentan algún signo?

Según se observó en la Figura N°9 algunas keywords incluyen dentro de su estructura el simbolo *%20*, el cual corresponde a la referencia codificada de un espacio en blanco.

Se contabilizaron un total de 1165 de keywords que posee este caracter, lo cual corresponde al 15.4% del total de las mismas.
En la Figura N°10 se identifican las 20 keywords con símbolo %20 que presentan mayor repetición:

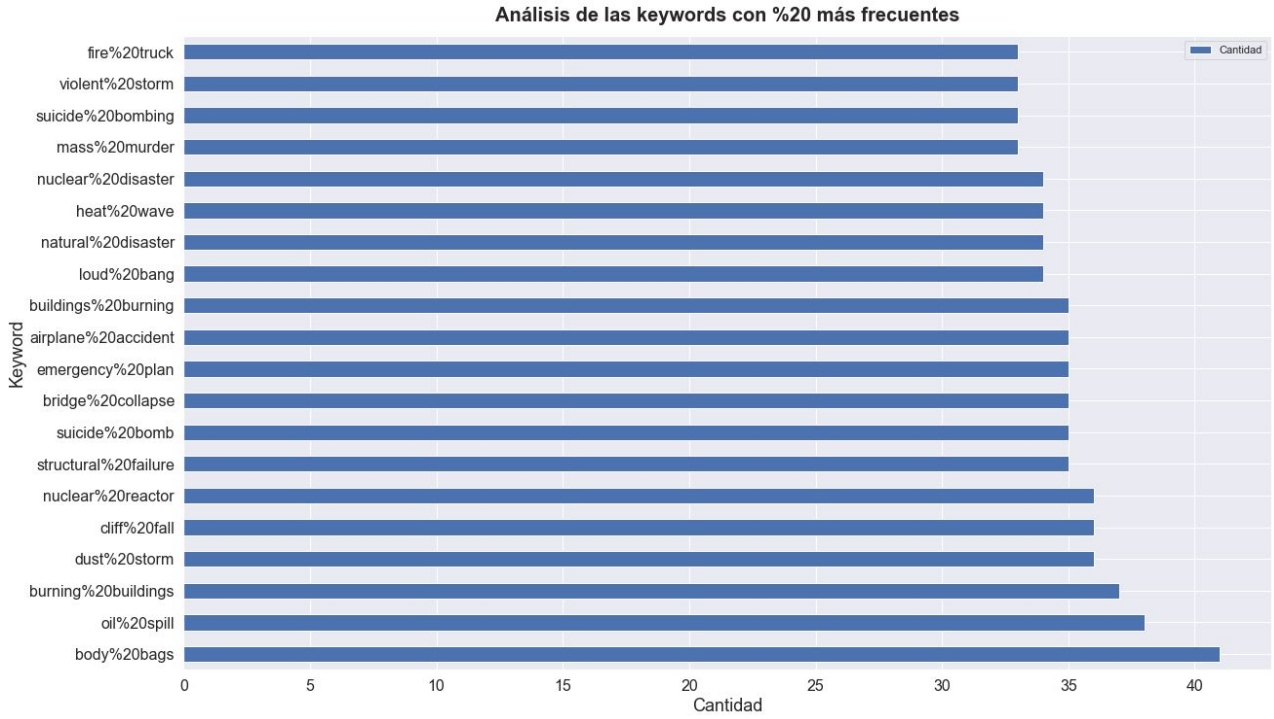

Figura N°10: Keywords más frecuentes que presentan el simbolo %20

5.2.3 Análisis de longitud

¿Como es el comportamiento del largo de las keywords?

En la Figura N°11 se observa la distribución de la cantidad de caracteres que poseen las keywords.
Del análisis del boxplot que se encuentra en esa figura se observa la presencia de outliers en la zona de mayor cantidad de caracteres. Este comportamiento en principio creemos que se debe a la presencia del símbolo %20. En la Figura N°12 se observa el comportamiento de aquellas keywords que tienen el simbolo %20 y aquellas que no lo tienen. Según se observa las keywords que tienen el simbolo %20 presentan una longitud mayor.

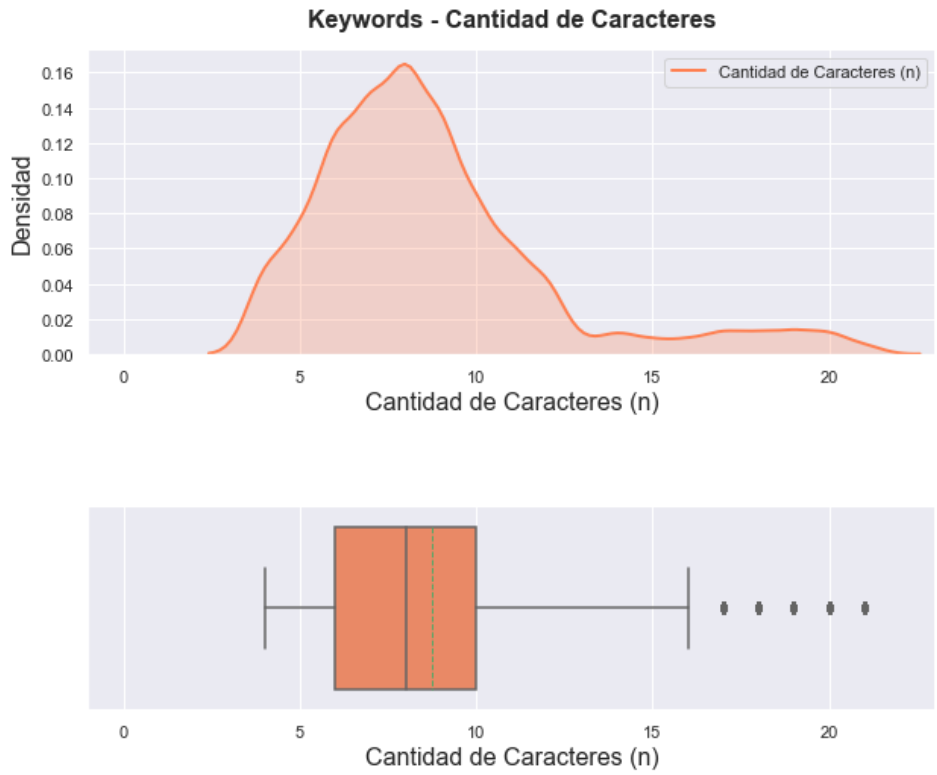

Figura N°11: Cantidad de caracteres en Keywords

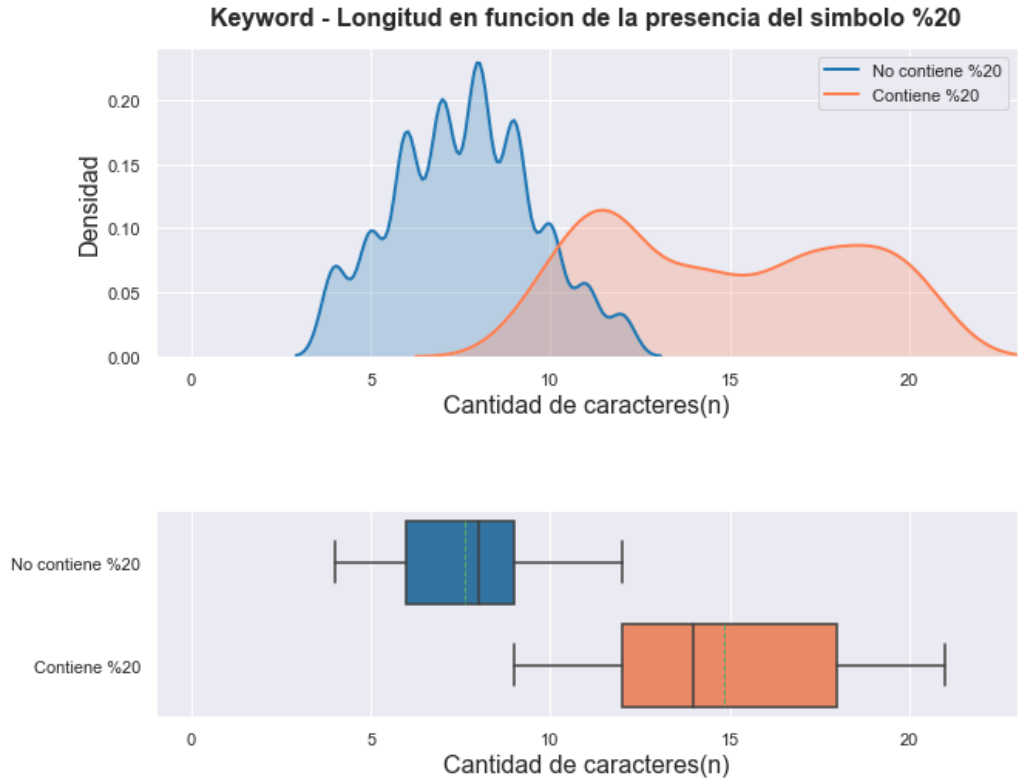

Figura N°12: Cantidad de caracteres en Keywords con y sin presencia del símbolo %20

5.3 Análisis de la Columna **location**

¿Cuántos valores nulos tiene 'location'?

Según se observa en la Figura N°13 la columna *location* cuenta con una gran cantidad de valores nulos, el 33% de los datos analizados son nulos. Esto es relevante desde el punto de vista de la limpieza de los datos, eliminar cada registro que tiene valor nulo significa perder un porcentaje muy grande de datos.

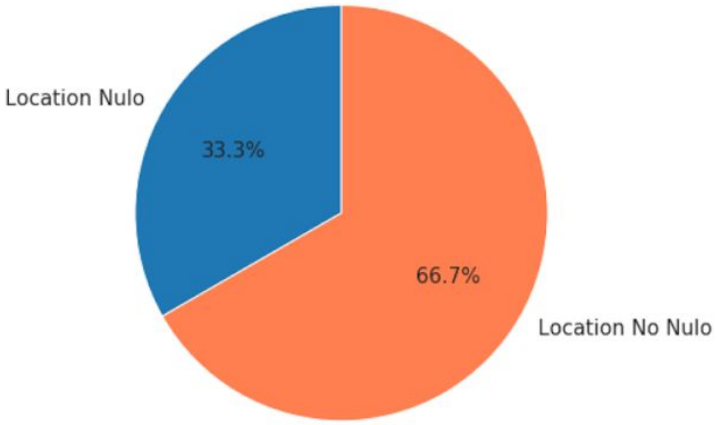

Figura N°13: Nulos en la columna 'location'

¿Cuántos valores distintos tiene la columna 'location'?

Según se observa en la Figura N°14 la columna location tiene 3341 valores distintos (de un total de 5080). Analizando los estadísticos vemos que el valor medio de repeticiones está entre 1 y 2, y la desviación estándar 3, esto quiere decir que dentro de los 3341 valores hay gran cantidad de ellos que se repiten menos de 5 veces, por lo que carece de sentido hacer algún tipo de clasificación utilizando esta columna con los datos como están.

	Cantidad
count	3341.000000
mean	1.520503
std	3.022364
min	1.000000
25%	1.000000
50%	1.000000
75%	1.000000
max	104.000000

Figura N°14: Estadísticos principales de la columna ‘location’

¿Se pueden agrupar los valores con algún criterio?

En la Figura N°15 se observan los distintos valores que toma esta columna y su frecuencia. En primer lugar vemos que hay valores que están incluidos dentro de otros, como por ejemplo podríamos pensar que *New York* está dentro de *USA*, o valores que significan lo mismo pero aparecen escritos de distinta forma, como *USA* y *United States*, o valores que no tienen sentido como *music* o *MayGodHelpUs.*

Ubicación	Cantidad
USA	104
New York	71
United States	50
London	45
Canada	29
...	...
Still. ??S.A.N.D.O.S??	1
music.	1
The Kingdom of Fife, Scotland	1
Arundel	1
#MayGodHelpUS	1

Figura N°15: Datos más frecuentes de la columna ‘location’

En la Figura N°16 se analiza, a modo de ejemplo, el string *USA*, vemos que aparece tanto solo, como dentro de una frase, por ejemplo *California, USA* utilizaremos esto para tratar de agrupar los datos.

Keyword que contiene USA	Cantidad
USA	104
California, USA	15
Pennsylvania, USA	7
New York, USA	5
Florida, USA	5
Texas, USA	5
North Carolina, USA	4
Massachusetts, USA	4
Virginia, USA	3
Hawaii, USA	3

Figura N°16: Estadísticos principales de la columna ‘location’

En la Figura N°17 se observa el resto de los valores que muestran las repeticiones de los valores tanto solos como dentro de una frase.

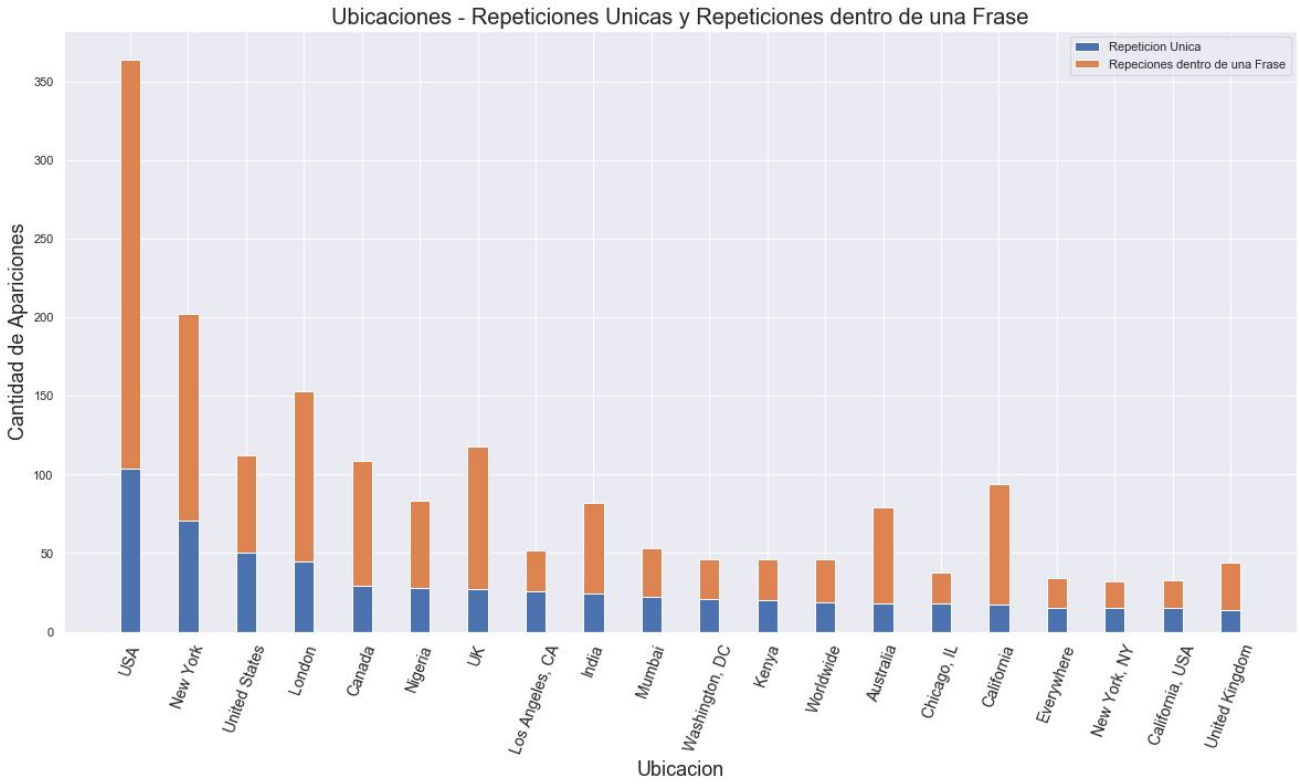

Figura N°17: Repeticiones únicas o dentro de una frase

En la Figura N°18 se indica la cantidad de registros para los 30 valores más frecuentes de esta columna. Vemos allí que aún los valores más frecuentes no superan las 100 repeticiones, que representan aproximadamente un 3% de los datos.

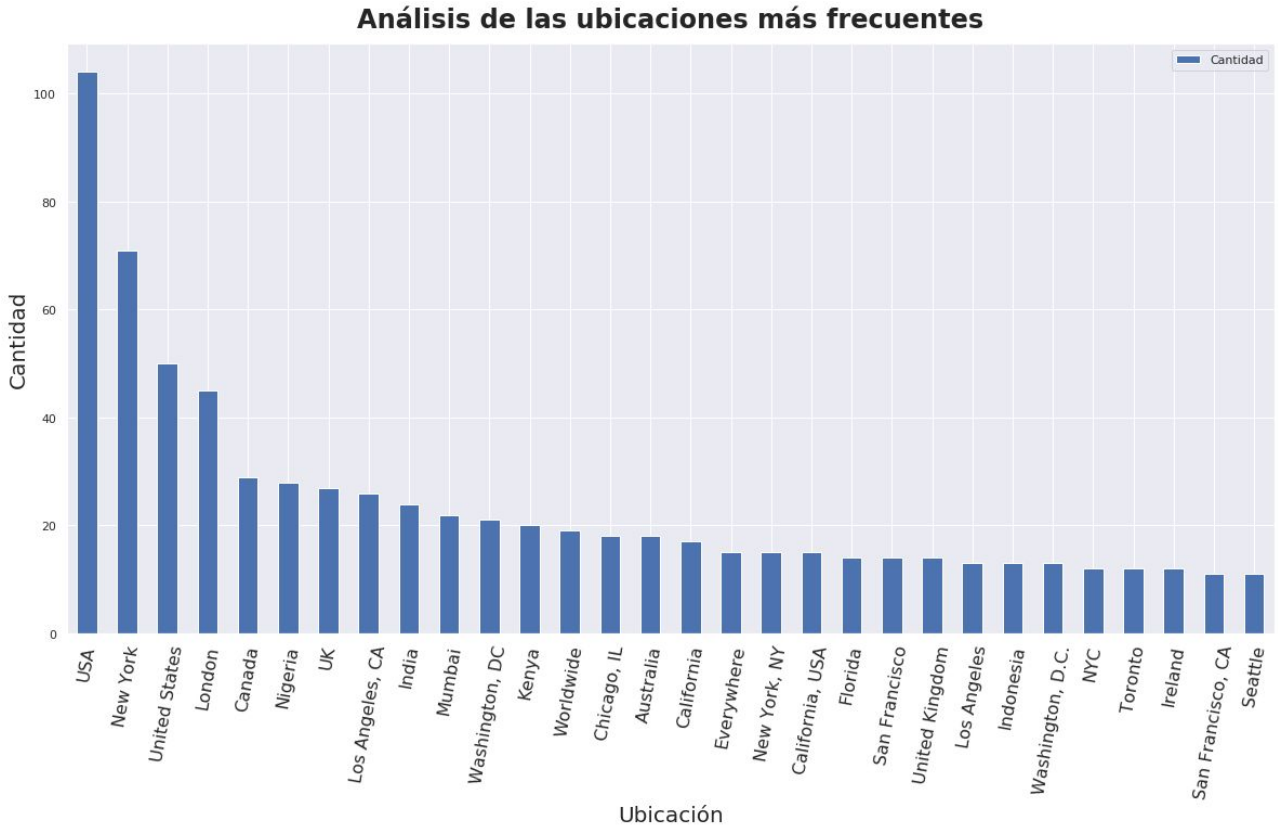

Figura N°18: Repeticiones Únicas

Intentamos agrupar dichos valores por países, en base al análisis realizado anteriormente. Para ello utilizamos la librería 'geotext' y agregamos algunos mapeos manualmente sobre los datos que la librería no reconoció. En la Figura N°19 se observa en la columna *Pais* con valor *other* aquellos que no pudieron ser traducidos.

	Ubicacion	Cantidad	Pais
0	USA	104	US
1	New York	71	US
2	United States	50	US
3	London	45	GB
4	Canada	29	CA
5	Nigeria	28	NG
6	UK	27	GB
7	Los Angeles, CA	26	US
8	India	24	IN
9	Mumbai	22	IN

Figura N°19: Repeticiones Únicas agrupadas por país

Luego de la limpieza, aún nos quedan 1714 registros que no pudieron ser agrupados (recordemos que el total de valores distintos era 3341), esto nos lleva a pensar dos posibilidades o bien descartamos la columna completa o bien podemos pensar a *los valores no agrupados* como otra categoría y ver si esto tiene sentido a la hora de determinar la veracidad de los tweets.

5.4 Análisis de la Columna **text**

5.4.1 Análisis del idioma

¿ Cuál es el idioma de los tweets?

Hemos evaluado el idioma predominante en esta columna. En la Figura N°20 se muestra el idioma detectado para cada registro:

id	text	Idioma
1	Our Deeds are the Reason of this #earthquake M...	English
4	Forest fire near La Ronge Sask. Canada	English
5	All residents asked to 'shelter in place' are ...	English
6	13,000 people receive #wildfires evacuation or...	English
7	Just got sent this photo from Ruby #Alaska as ...	English
...	...	...
10869	Two giant cranes holding a bridge collapse int...	English
10870	@aria_ahrary @TheTawniest The out of control w...	English
10871	M1.94 [01:04 UTC]?5km S of Volcano Hawaii. htt...	English
10872	Police investigating after an e-bike collided ...	English
10873	The Latest: More Homes Razed by Northern Calif...	English

Figura N°20: Idioma por tweet

En la Figura N° 21 analizamos cuántos idiomas distintos aparecían y la frecuencia de cada uno.

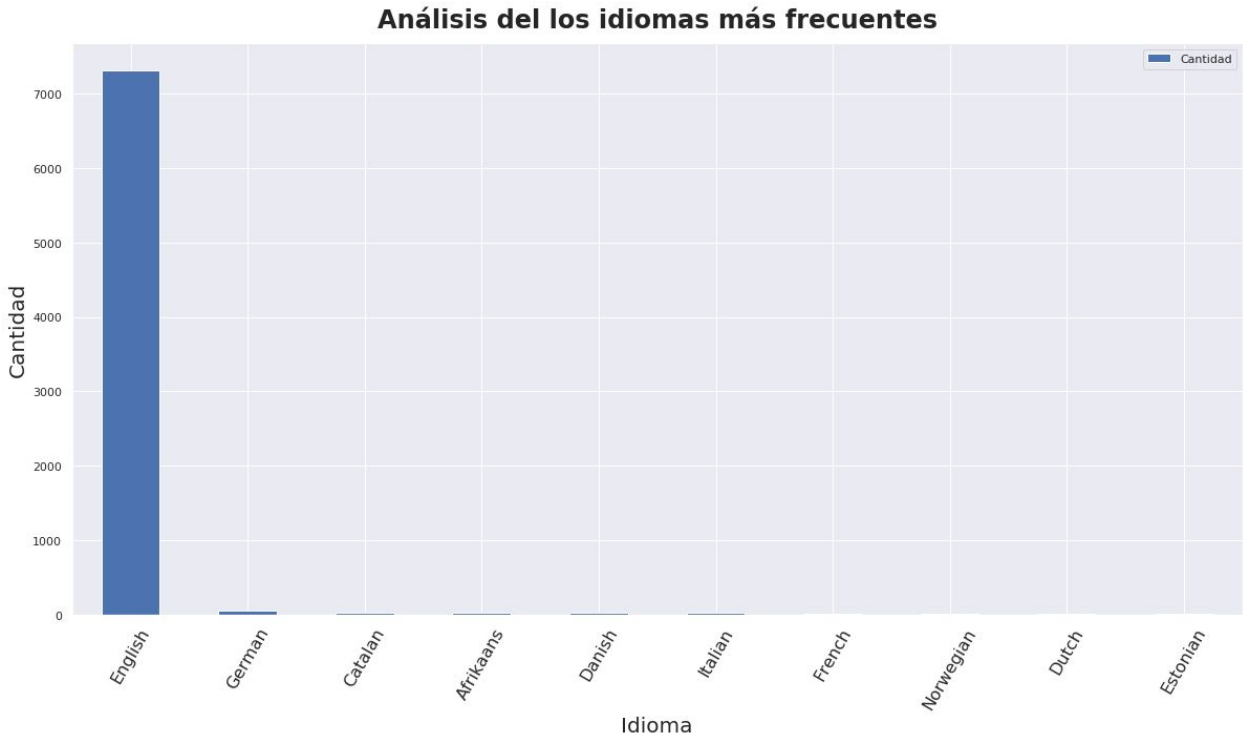

Figura N°21: Cantidades por Idioma

En la Figura N°22 se observa que la mayoría de los textos están en idioma Inglés, con un 96%.

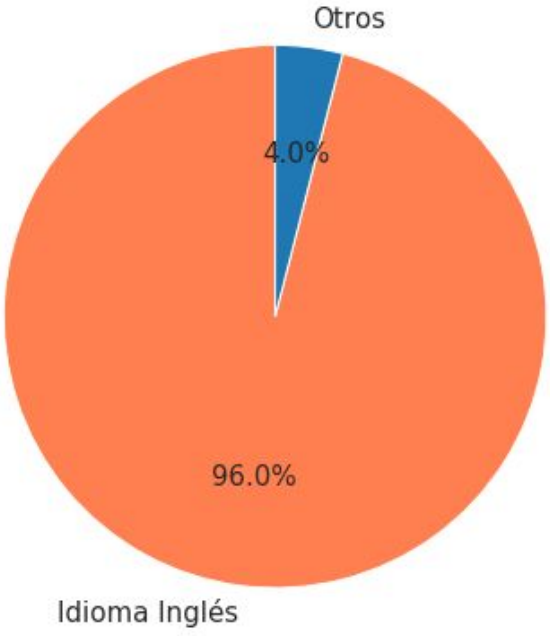

Figura N°22: Porcentaje de Idioma Inglés y otros

5.4.2 Análisis de cantidad de caracteres

¿Cuál es la longitud de los textos de los tweets?

En la Figura N°23 se observa la distribución de la longitud de caracteres del texto de los tweets. Según se observa en largo en caracteres tiene una distribución centrada en valores cercanos a 100 y con un pico de densidad en torno a los 135

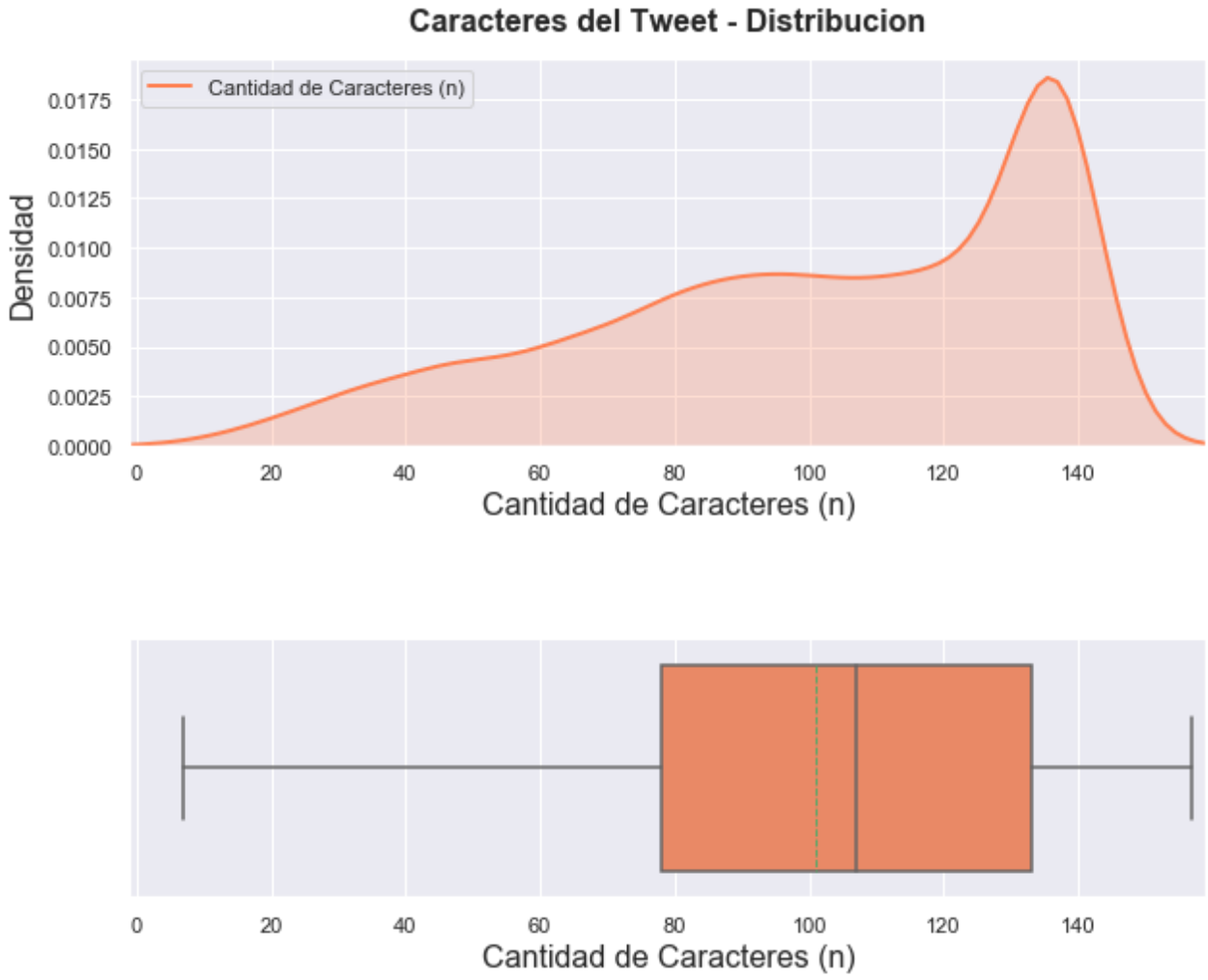

Figura N°23: Cantidad de Caracteres

5.4.3 Análisis de cantidad de palabras

¿Qué cantidad de palabras tienen los tweets?

En la Figura N°24 se observa la distribución de la longitud de palabras del texto de los tweets. Según se observa el largo en caracteres tiene una distribución centrada en valores cercanos a 15 y con un pico de densidad cercano a este valor

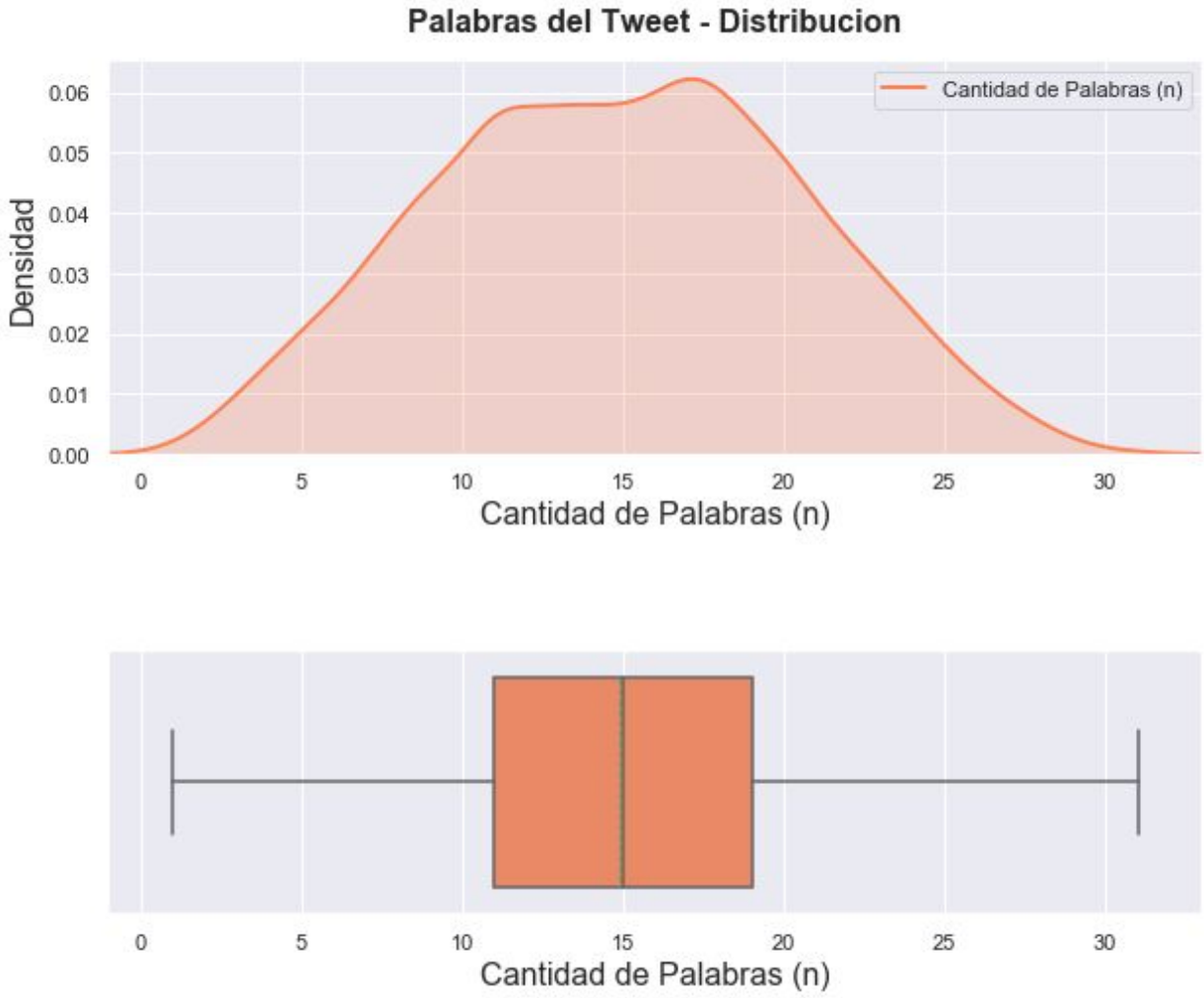

Figura N°24: Cantidad de Caracteres

5.4.5 Análisis de palabras de mayor aparición

¿Cuales son las palabras más repetidas en los tweets analizados?

En la Figura N°25 se pueden observar las repeticiones totales de cada palabra que aparece en los tweets, junto con el indicador de veracidad según color.

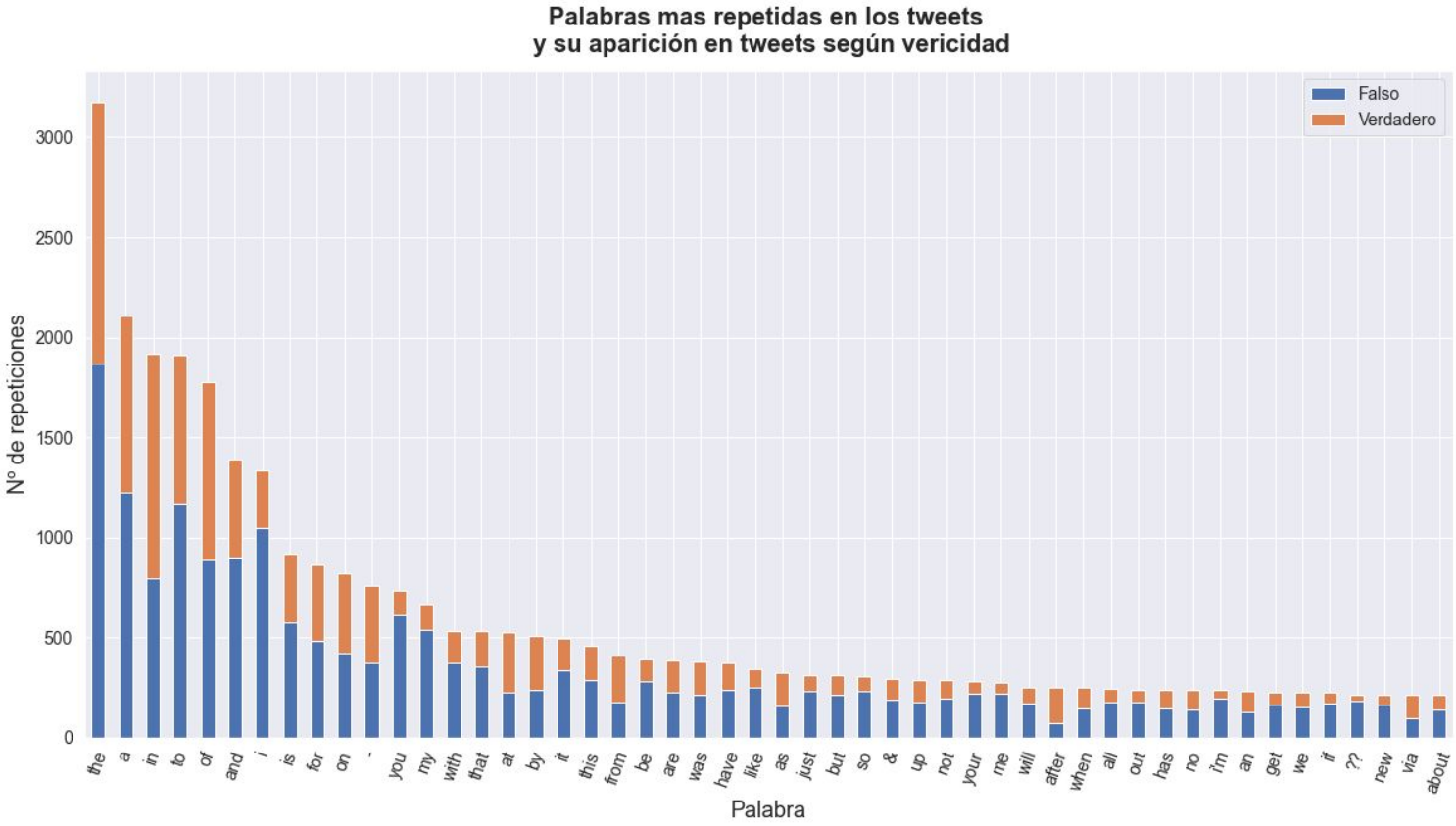

Figura N°25: Palabras más repetidas

5.4.6 Análisis de Hashtags (#)

¿Cuales tweets tienen hashtags?

Según se observa en la figura N°26 al evaluar el texto de cada tweet vemos que el 22.8 % de ellos contienen hashtags.

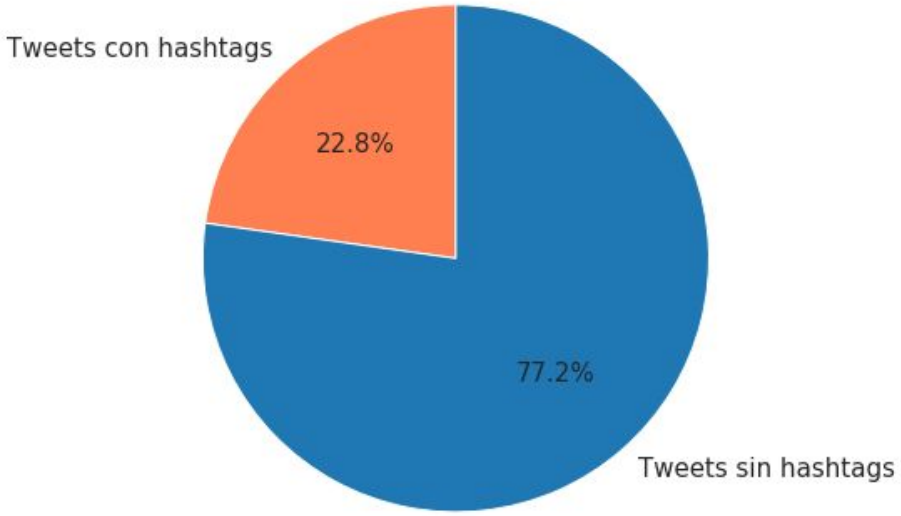

Figura N°26: Presencia de hashtags en los tweets

¿Cuales son los hashtags que más aparecen?

En la Figura N°27 se pueden observar un WordCloud con los hashtags con mayor repetición. Entre ellos tenemos palabras como *news*, *prebreak*, *hot* y *best.*
En la Figura N°28 se ve el detalle la frecuencia de repeticiones de cada hashtag, se puede advertir allí que aún los de mayor repetición, aparecen alrededor de unas 30 veces, siendo este un valor bajo comparado con el total de registros que está en el orden de los 7000.

Figura N°27: Hashtags de mayor aparicion

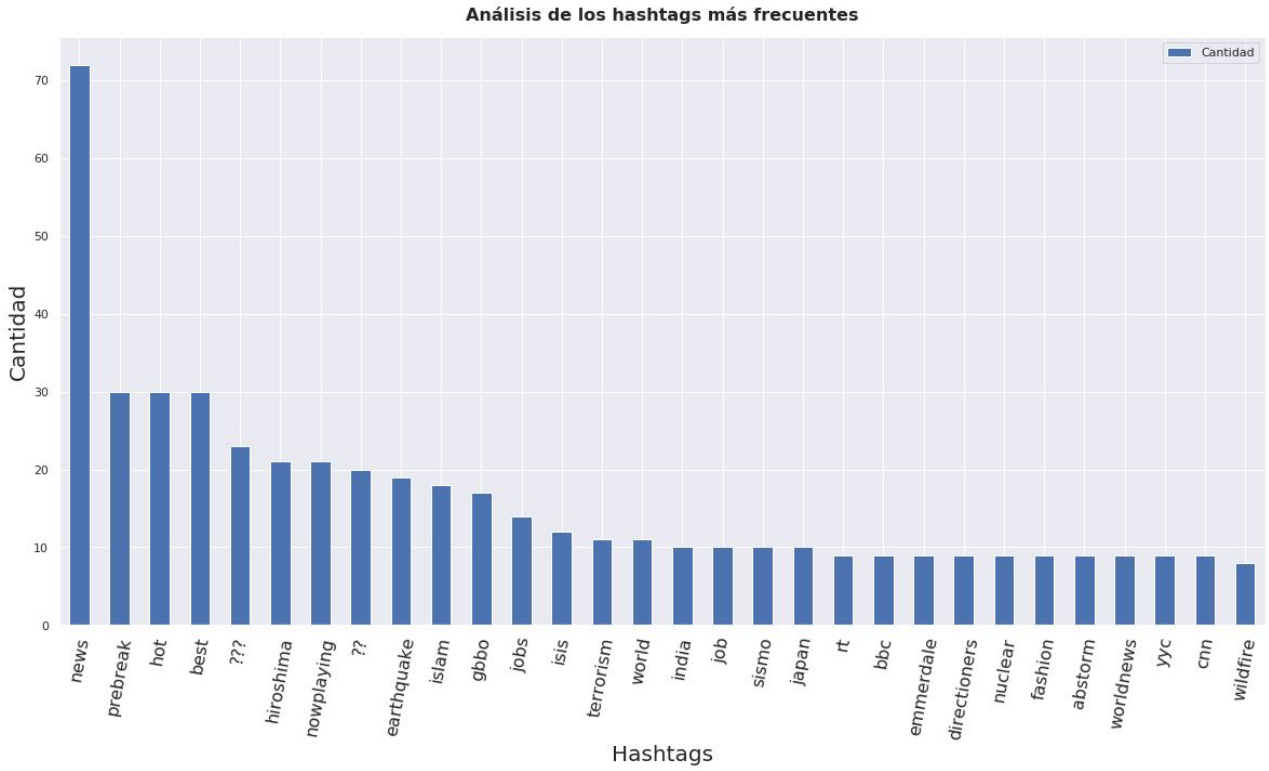

Figura N°28: Hashtags de mayor aparicion

5.4.7 Análisis de menciones

¿Los tweets tienen menciones? ¿Cuales son las menciones que mayor se repiten?

En la Figura N°29 se observan las menciones de mayor ocurrencia, entre las cuales se destaca *YouTube.* Al igual que con los hashtags la cantidad de menciones es bajo comparado con el total de registros.

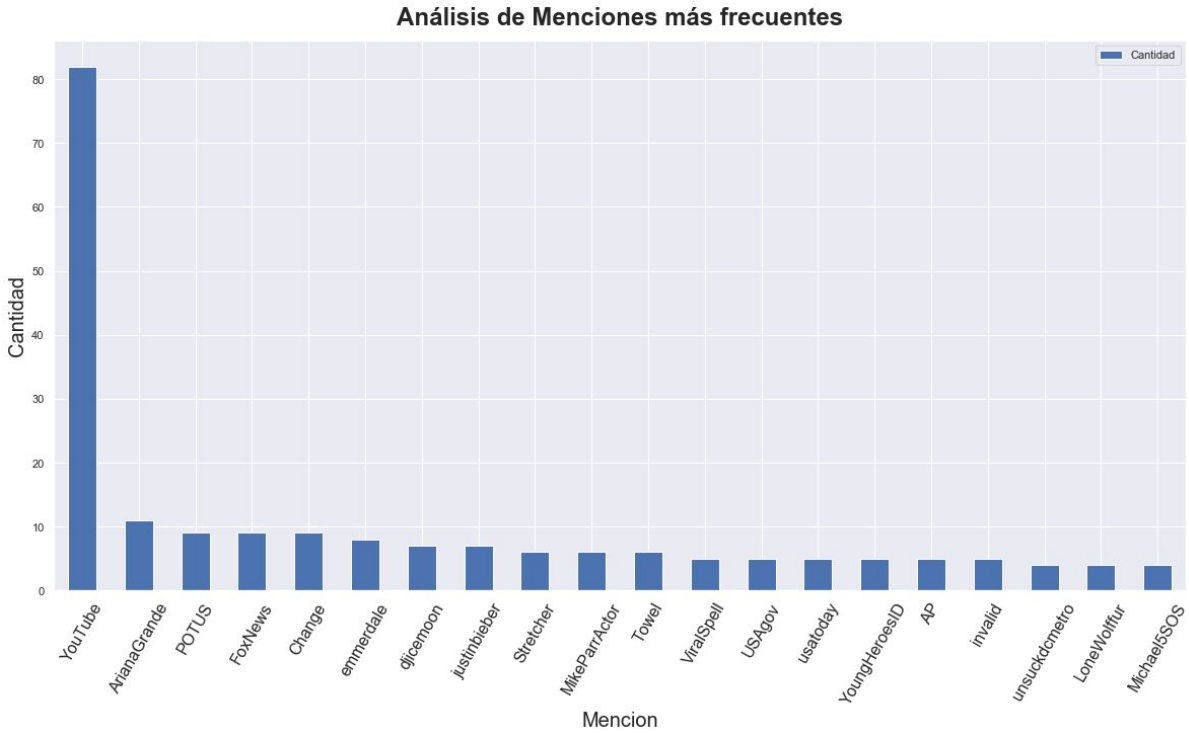

Figura N°29: Menciones de mayor aparición

5.4.8 Análisis de URLs

¿Cuantos tweets contienen URLs?

Hemos detectado que hay ciertos tweets que contienen URLs, esto puede incidir en el análisis, por ejemplo si queremos estudiar la longitud de los tweets, y estos contienen URLs, seguramente serán más largos que el resto.
En la Figura N°30 se observan algunos de ellos.

id	keyword	location	text	target	longitud	Tiene URL
48	ablaze	Birmingham	@bbcmtd Wholesale Markets ablaze http://t.co/l...	True	55	True
49	ablaze	Est. September 2012 - Bristol	We always try to bring the heavy. #metal #RT h...	False	67	True
50	ablaze	AFRICA	#AFRICANBAZE: Breaking news:Nigeria flag set a...	True	82	True
53	ablaze	London, UK	On plus side LOOK AT THE SKY LAST NIGHT IT WAS...	False	76	True
55	ablaze	World Wide!!	INEC Office in Abia Set Ablaze - http://t.co/3...	True	55	True
...	...	...	...	...	...	...
10866	NaN	NaN	Suicide bomber kills 15 in Saudi security site...	True	121	True
10867	NaN	NaN	#stormchase Violent Record Breaking EF-5 El Re...	True	134	True
10869	NaN	NaN	Two giant cranes holding a bridge collapse int...	True	83	True
10871	NaN	NaN	M1.94 [01:04 UTC]?5km S of Volcano Hawaii. htt...	True	65	True
10873	NaN	NaN	The Latest: More Homes Razed by Northern Calif...	True	94	True

Figura N°30: Detección de tweets con URLs

Continuando con el análisis, vemos en la Figura N°31 un total de 3971 tweets que contienen URLs, es decir aproximadamente un 52%

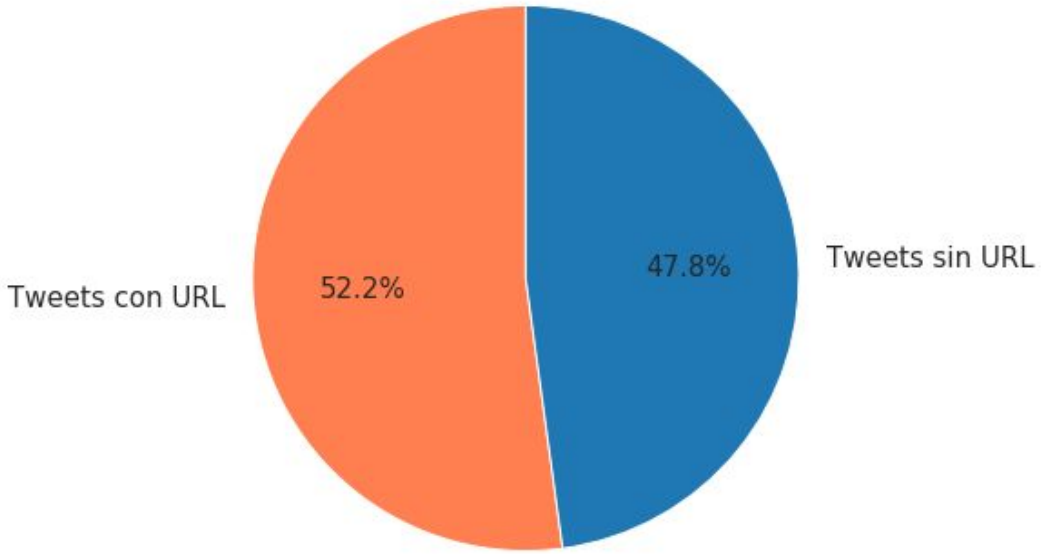

Figura N°31: URLs en los tweets

¿Cómo inciden las URLs en la longitud de los tweets?

A partir del análisis de la Figura N°32 podemos decir que al extraer las URLs de los tweets, su distribución de longitudes se reduce considerablemente, vemos además que la media disminuye, como era de esperarse, al extraer una parte del texto.

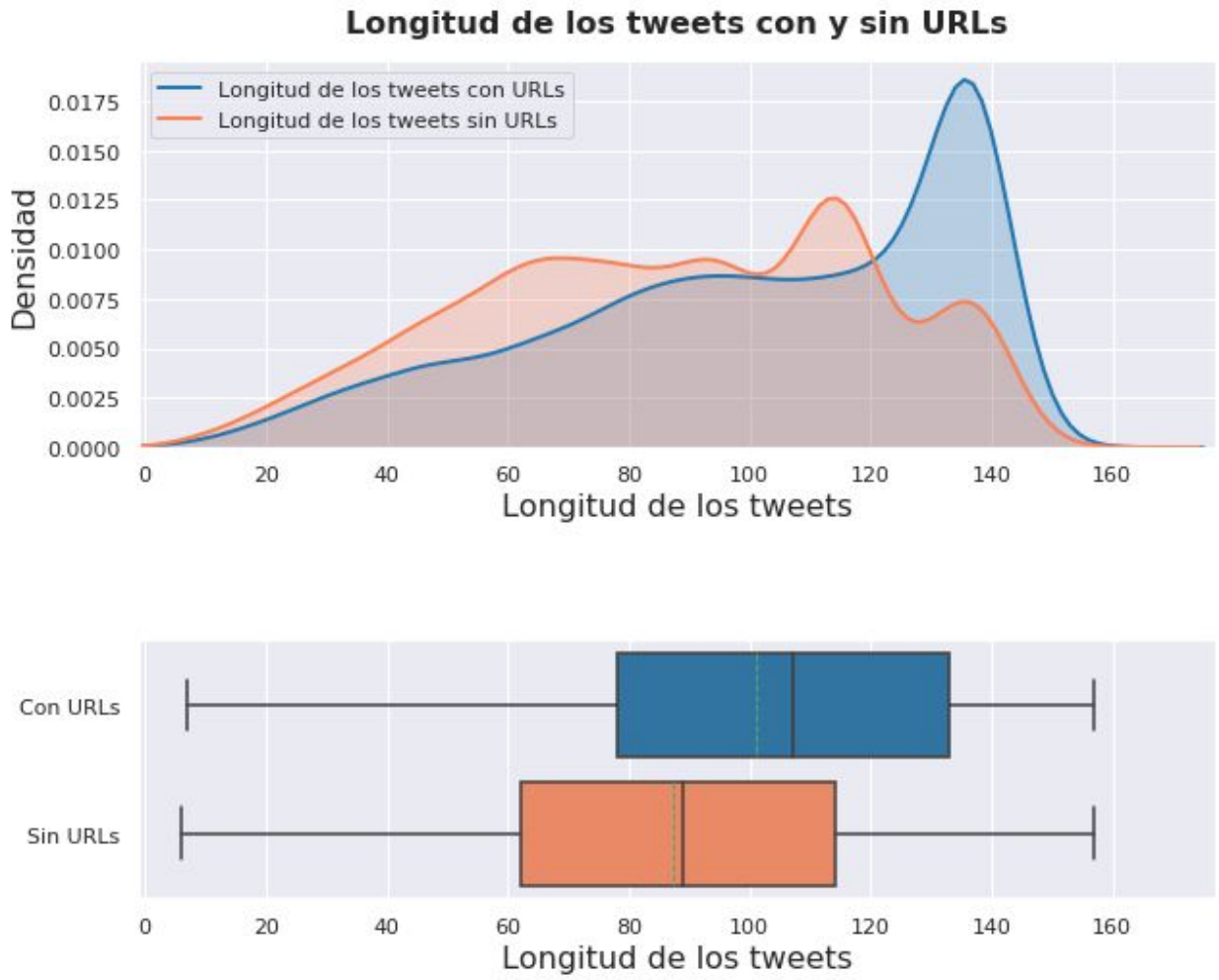

Figura N°32: Presencia de URLs en los tweets

5.4.9 Análisis de stopwords, puntuaciones y números

¿Cómo inciden los signos de puntuación y stopwords en la longitud de los los tweets?

En la Figura N°33 se observa cómo disminuye el valor medio de las longitudes de los tweets, luego de realizar la extracción de : URLs, signos de puntuación, números y stopwords.

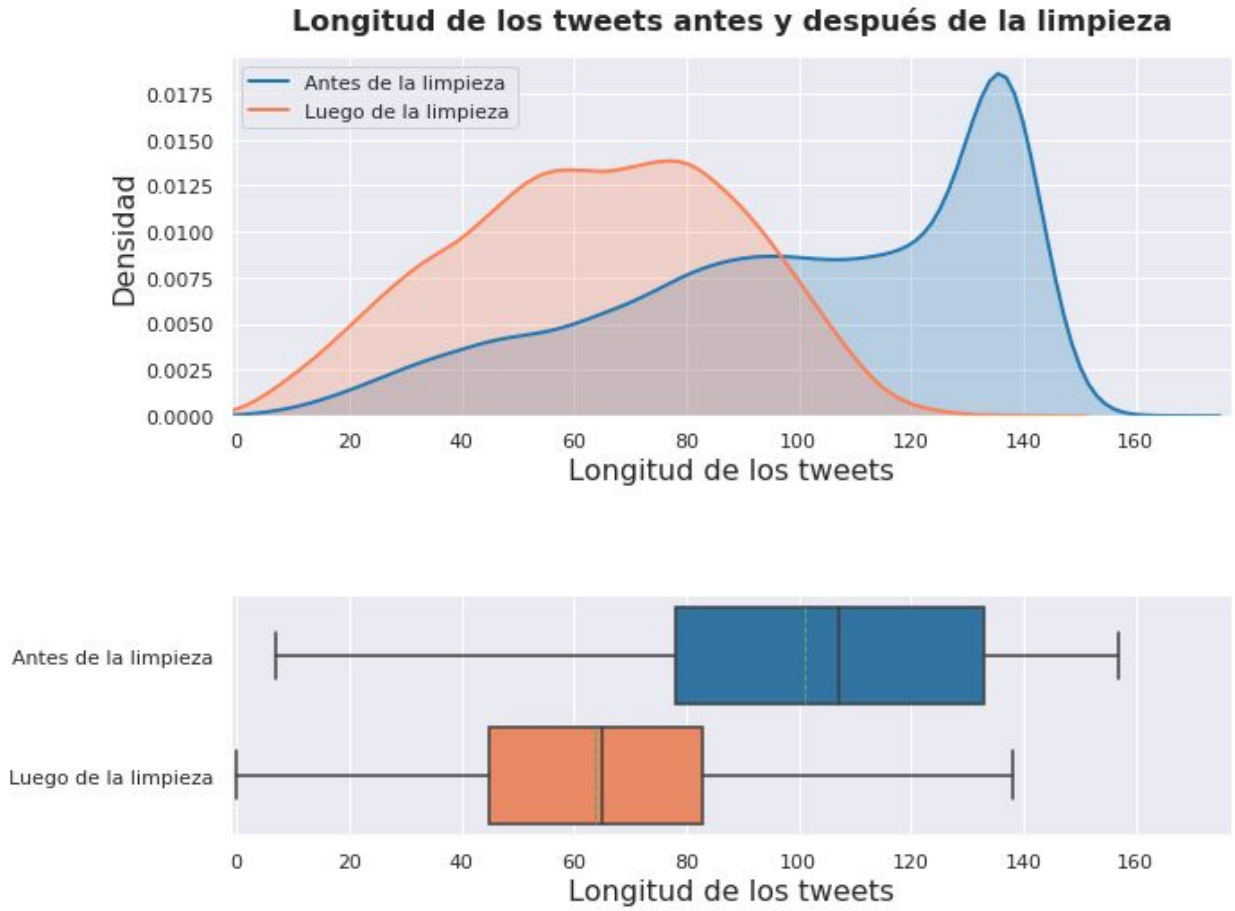

Figura N°33: Analisis de URLs, signos de puntuación, números y stopwords.

6. Características de Columnas analizadas en conjunto

6.1 Análisis de la Columna 'keyword' y 'location'

¿Como es la relación entre keywords y location? ¿Para un lugar específico existen keywords con mayor repetición?

Según se identificó en el ítem **5.3 Análisis de la columna 'location'** los lugares con mayor frecuencia de aparición son aquellos que poseen las palabras *USA*, *New York* y *United States*. Para cada uno de estos lugares se identificaron en las Figuras N°34/35/36 aquellas keywords con mayor ocurrencia.

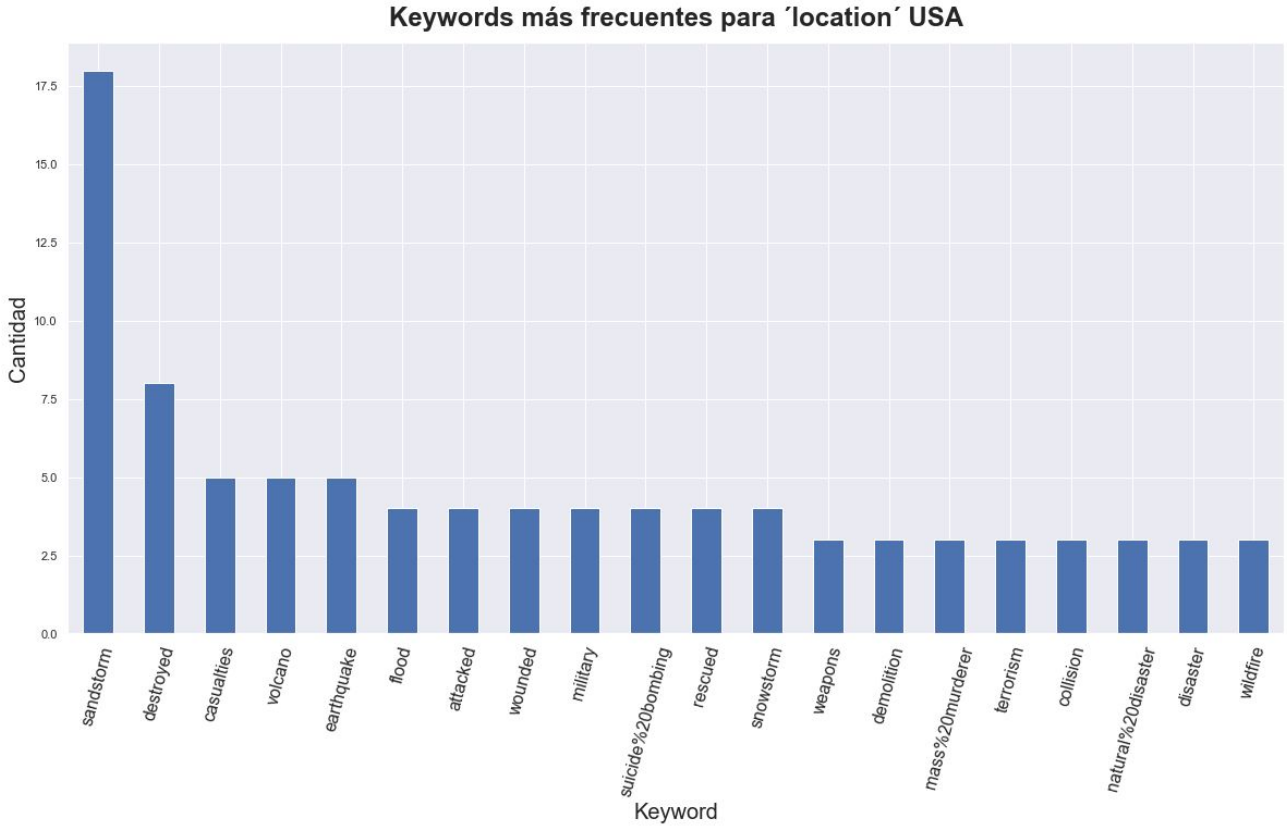

Figura N°34: Keywords más frecuentes para la ubicación 'USA'

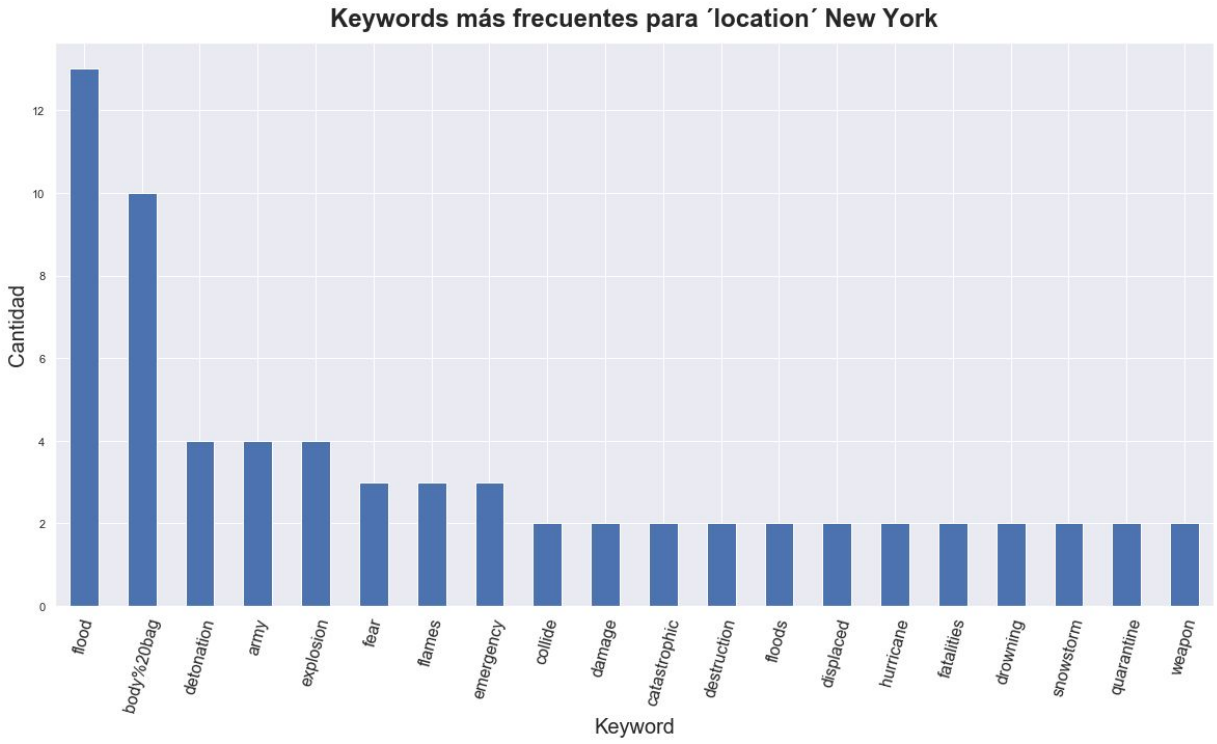

Figura N°35: Keywords más frecuentes para la ubicación ‘New York’

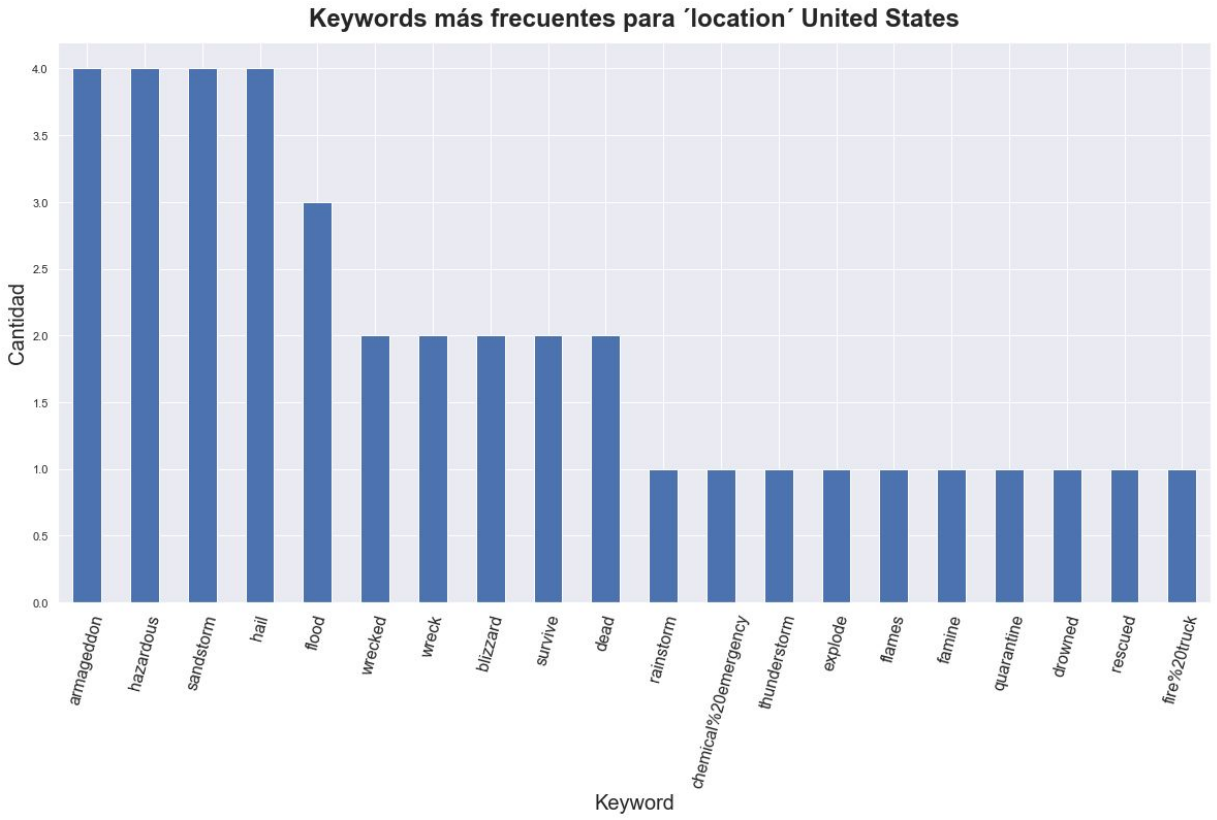

Figura N°36: Keywords más frecuentes para la ubicación ‘United States’

Adicionalmente en la Figura N°37 se ha querido comprobar si existía algún tipo de relación entre estos dos parámetros, para ello tomamos los conjuntos de valores keyword-location que más se repetían llegando a la conclusión de que dicha cantidad de repeticiones es muy pequeña para poder determinar alguna relación entre ellos.

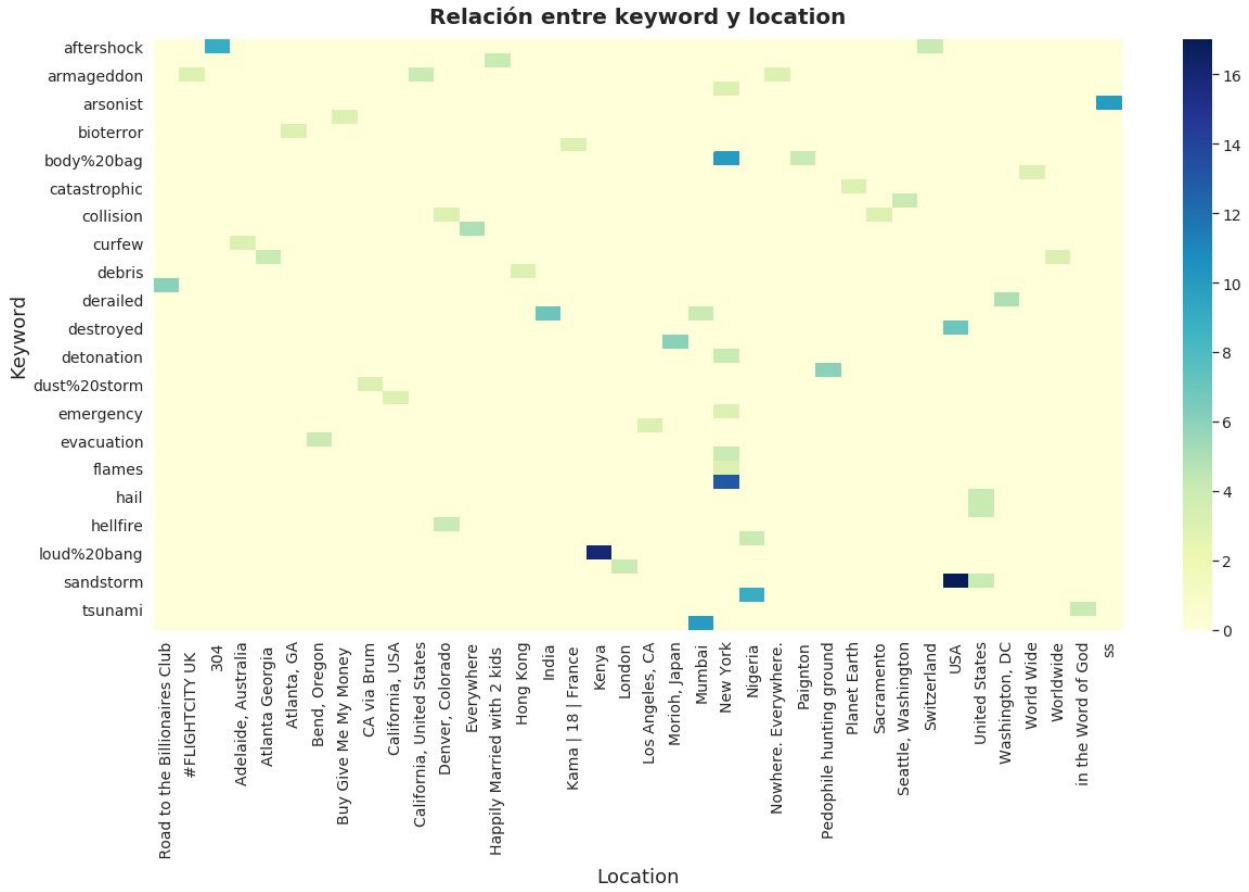

Figura N°37: Relación entre keyword y location

6.2 Análisis de la Columna 'keyword' y 'target'

¿Como es la relación entre keywords y target? ¿Existen keywords con mayor repetición para tweets Falsos y Verdaderos?

En la Figura N°38 se observan las 20 keywords de mayor ocurrencia junto con las proporciones de tweets que son Falsos y Verdaderos. Según se observa algunas keywords tienen una proporción mayor de tweets Falsos (deluge, amageddon, body%20bags) y otras una proporción mayor de tweets Verdaderos (outbreak)

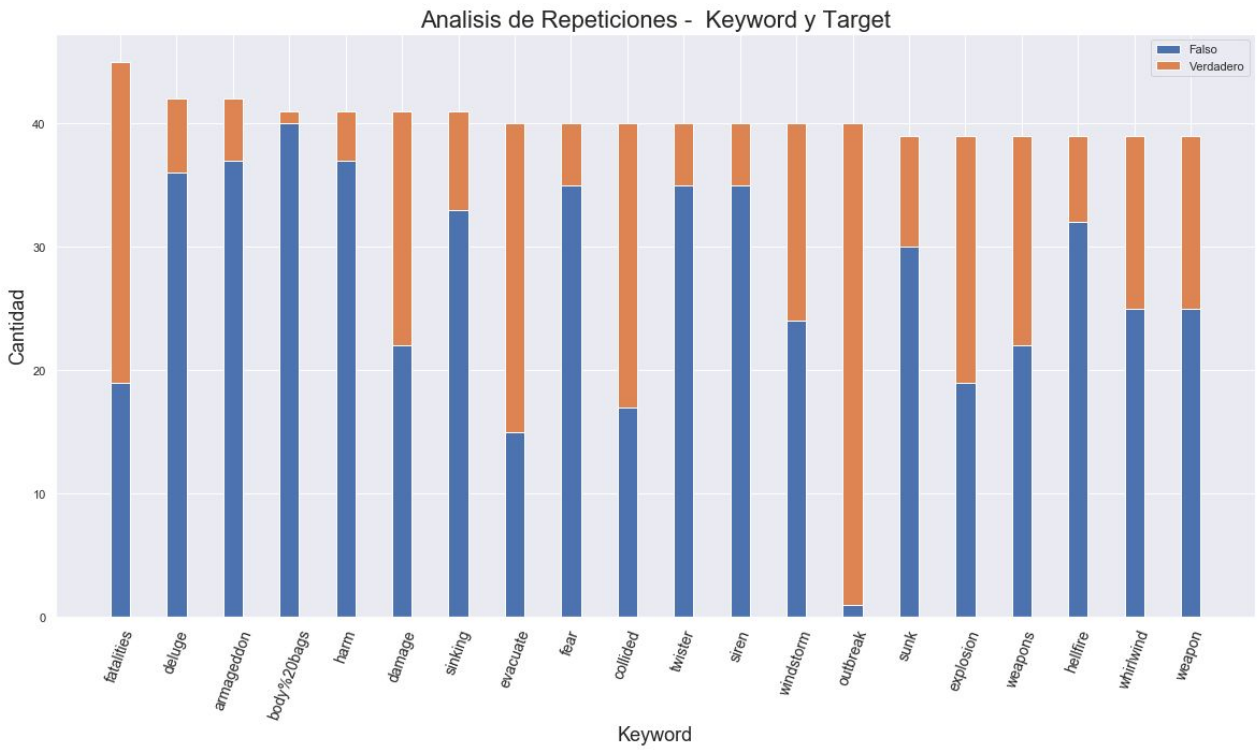

Figura N°38: Keywords más frecuentes y veracidad

6.3 Análisis de la Columna 'location' y 'target'

¿Como es la relación entre location y target? ¿Existen lugares donde la cantidad de tweets Falsos y Verdaderos es mayor?

En la Figura N°39 se observan las 50 ubicaciones de mayor ocurrencia junto con las proporciones de tweets que son Falsos y Verdaderos. Es fácilmente observable que hay una enorme cantidad de ubicaciones nulas, por lo tanto en la Figura Nº40 vamos a hacer un nuevo análisis que no los incluya y se puedan observar de mejor manera los datos.

Según se puede observar hay varias ubicaciones donde predominan tanto los falsos como los verdaderos.

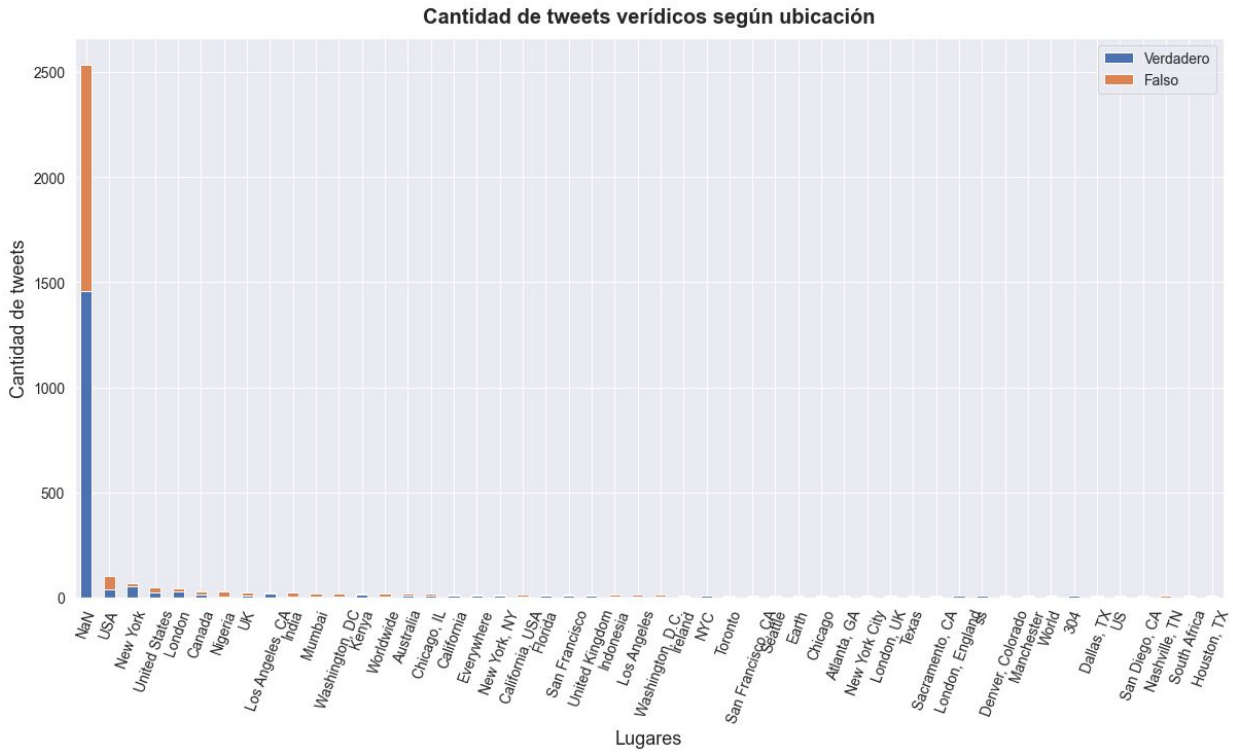

Figura N°39: Tweets y su veracidad según ubicación

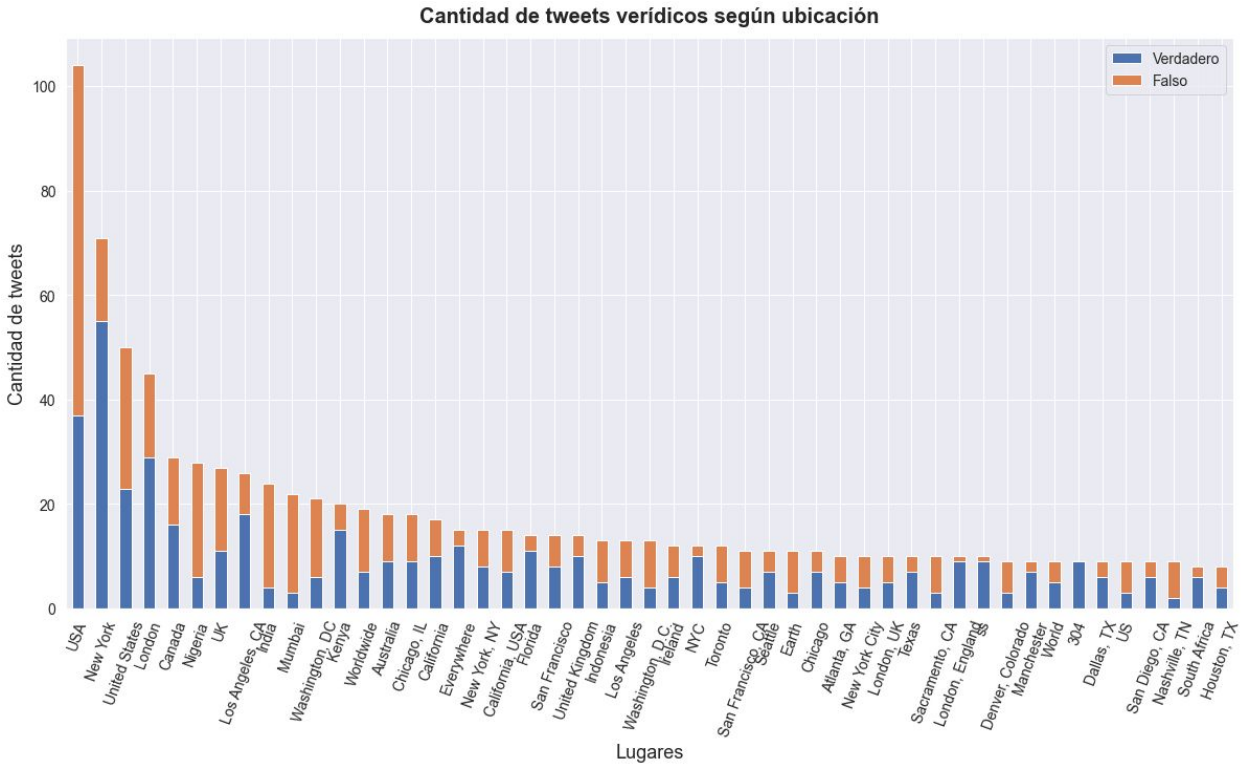

Figura N°40: Tweets y su veracidad según ubicación (excl. nulos)

6.4 Análisis de la Columna 'text' y 'target'

6.4.1 Análisis del largo del tweet en caracteres y su veracidad

¿Como es el comportamiento del largo (caracteres) de los tweets y la veracidad de los mismos?

En la Figura N°41 se observan las distribuciones de caracteres para tweets catalogados como Verdaderos y Falsos. En ambos casos se observa que la mayor densidad de produce en valores cercanos a 135 caracteres. Con respecto las medianas se observa una diferencia en su posición a que para tweets Falsos su posición se encuentra en torno a los 100 caracteres y para tweets Verdaderos se encuentra en tornos a los 115 caracteres

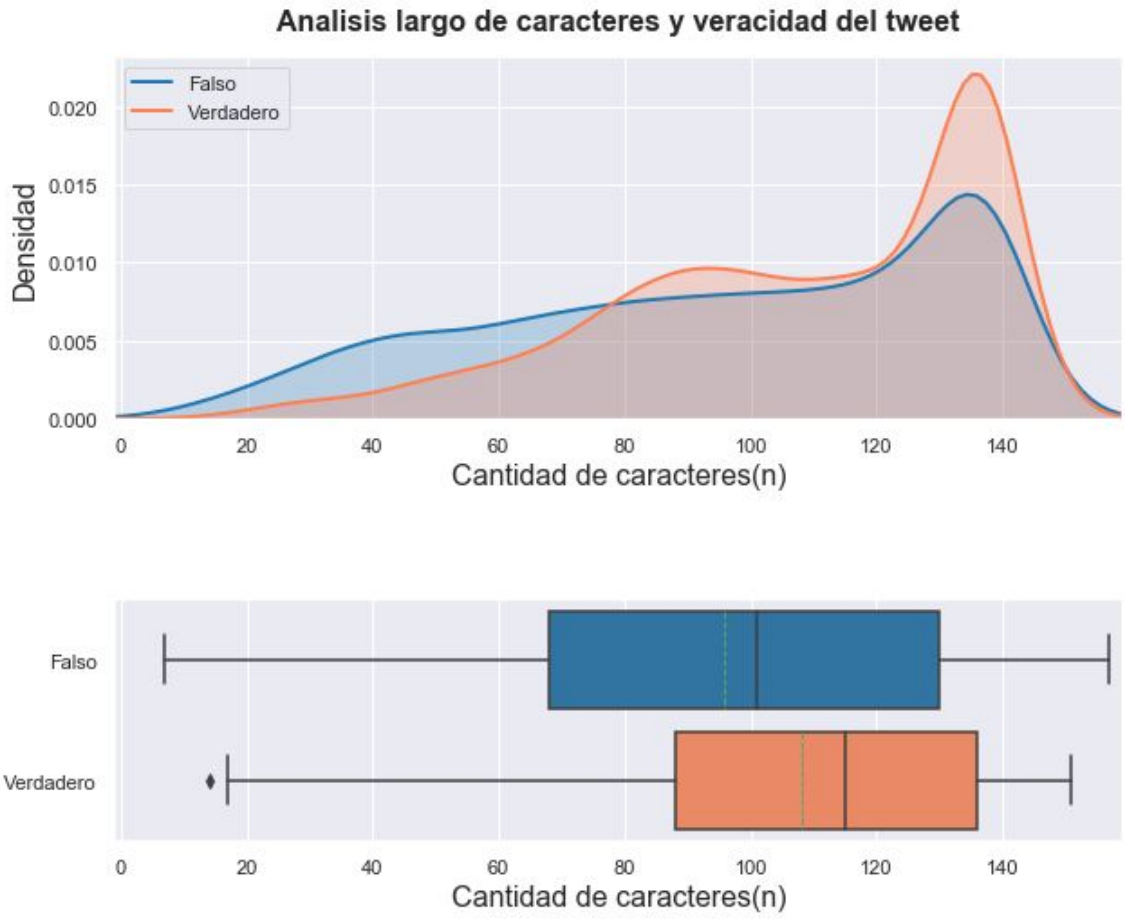

Figura N°41:Longitud en caracteres de los tweets y veracidad

6.4.2 Análisis del largo del tweet en palabras y su veracidad

¿Como es el comportamiento del largo (palabras) de los tweets y la veracidad de los mismos?

En las Figuras N°42/43 se observan las distribuciones de palabras para tweets catalogados como Verdaderos y Falsos. En ambos casos se observa un comportamiento acampanado en la densidad. Con respecto las medianas se observan valores similares en torno a 15 palabras.

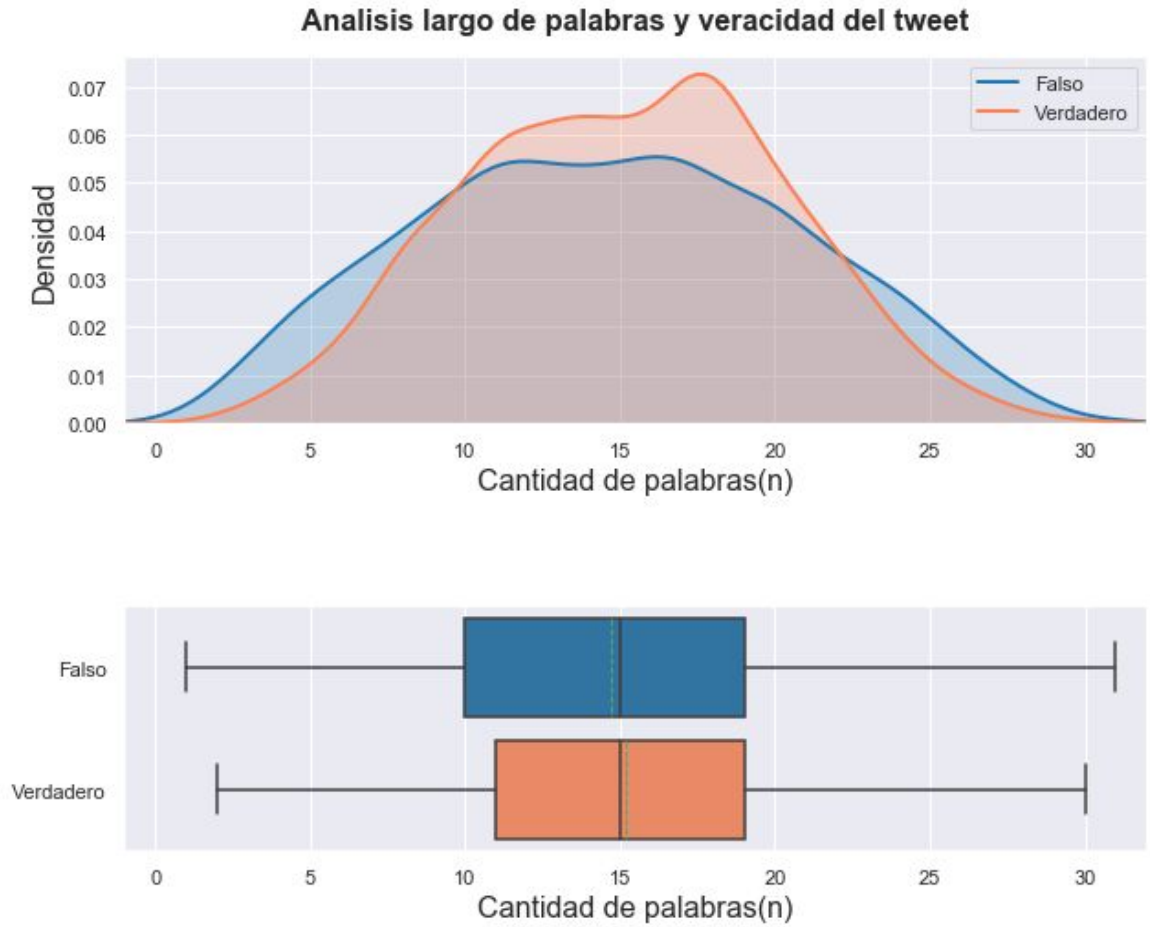

Figura N°42:Longitud en palabras de los tweets y veracidad

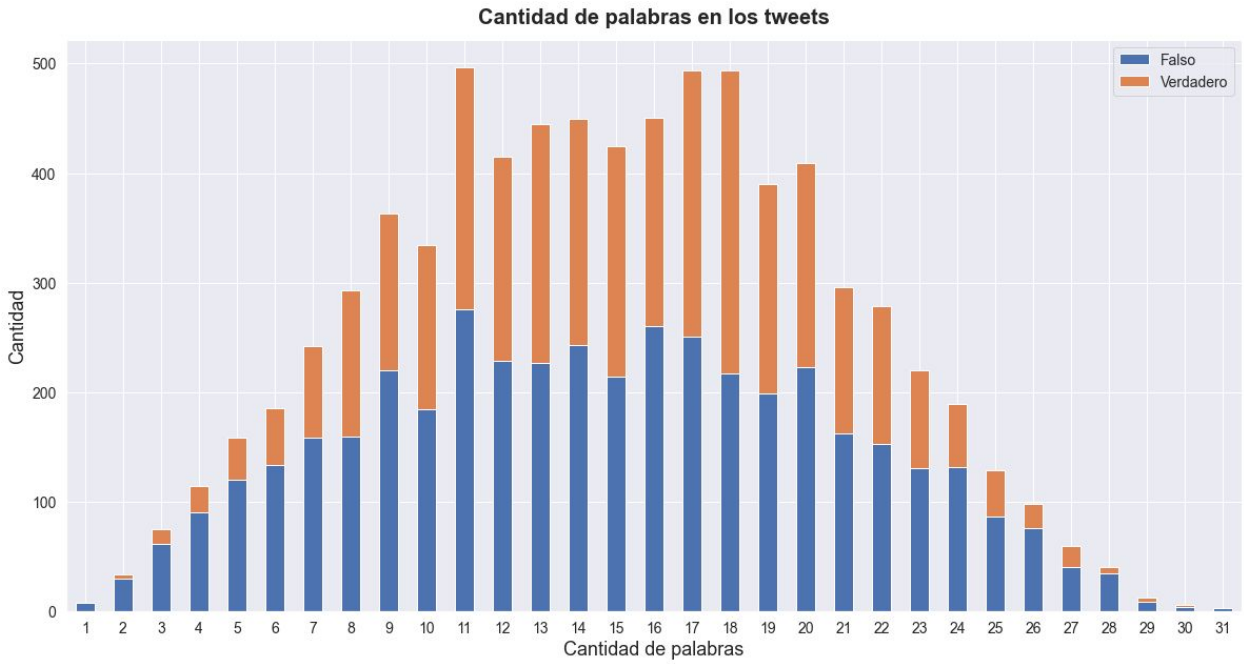

Figura N°43:Longitud en palabras de los tweets y veracidad

6.4.3 Análisis cantidad de párrafos de cada tweet y su veracidad

¿Cuales son las cantidades de párrafos más usuales en un tweet?

En la Figura N°44 se observa la cantidad de párrafos totales en un tweet, junto con el indicador de veracidad según el color. Como se puede observar la inmensa mayoría de los tweets tienen tan solo un párrafo. En la Figura Nº45 se puede apreciar la distribución de esa variable de mejor manera.

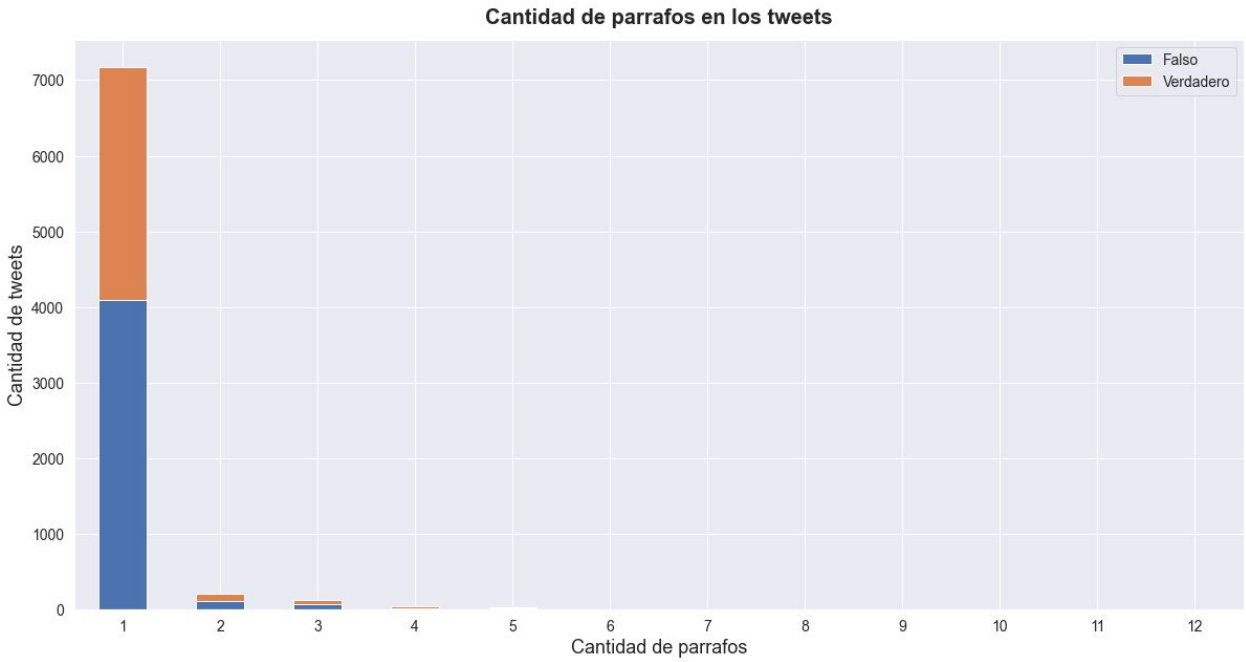

Figura N°44: Cantidad de párrafos

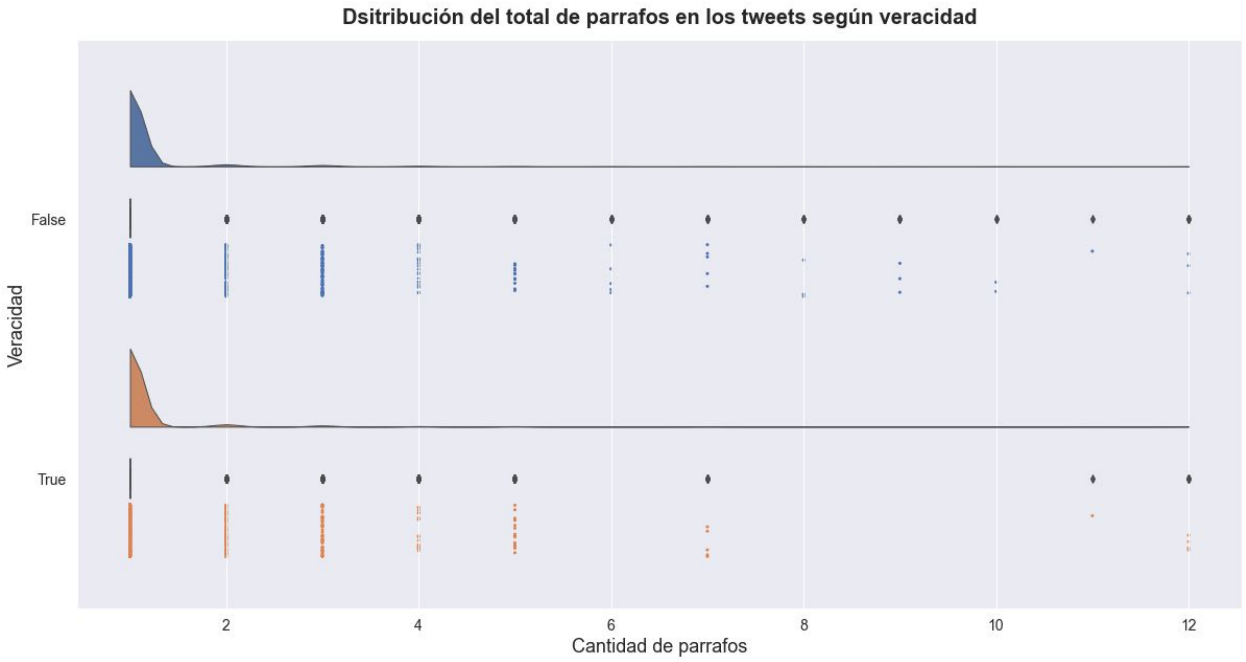

Figura N°45: Distribución de la cantidad de párrafos

6.4.4 Análisis del largo promedio de caracteres en las palabras de cada tweet y su veracidad

¿Como es el comportamiento del largo promedio de caracteres en las palabras que forman cada tweet?

En las Figuras N°46 se observa este comportamiento para tweets catalogados como Verdaderos y Falsos. Se observa un promedio mayor para tweets catalogados como Verdaderos

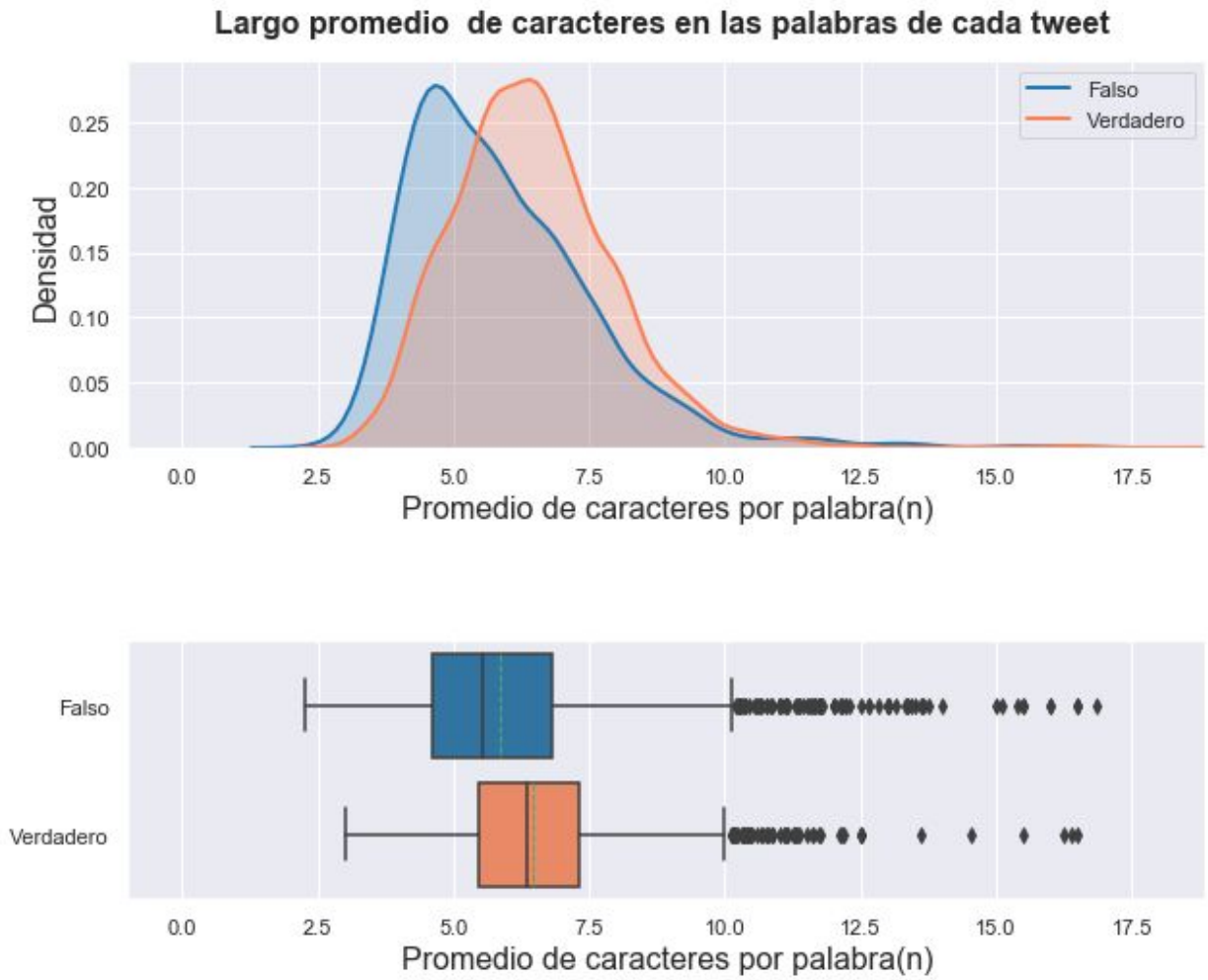

Figura N°46: Longitud promedio de caracteres para las palabras de los tweets y veracidad

7. Conclusiones sobre limpieza de datos

Para la utilización del dataset en otras actividades (ej:entrenamiento de un modelo de aprendizaje automático) consideramos necesario realizar previamente las siguientes tareas que fueron evaluadas en este Trabajo Práctico:

- Limpieza de Hashtags en la columna text
- Limpieza de puntuaciones y stopwords en la columna text
- Limpieza de URLs en la columna text
- Limpieza de HTML tags en la columna text
- Limpieza de datos en la columna location o bien descarte completo de la misma.
- Limpieza de registros con valores nulos en la columna keyword
- Limpieza de registros repetidos.

8. Conclusiones generales

En base al análisis que hemos realizado hemos llegado a las siguientes conclusiones:

- Para análisis posteriores se deberá realizar una limpieza completa de los datos según se ha detallado en el ítem 7.
- No se observan mayores dificultades para reemplazar los valores nulos con el string “NaN” de manera temporal para después volver a ponerlos en nulo
- Los datos tienen una cantidad mayor de tweets que representan noticias falsas que verdaderas.
- Las keywords más frecuentes son ‘fatalities’, ‘deluge’ y ‘armageddon’

- Las keywords tienen una longitud promedio de 8 caracteres.
- Hay 221 keywords distintas, por lo cual tiene sentido considerarla como una variable categórica.
- La columna location tiene una gran cantidad de valores nulos, además por su gran diversidad de valores resulta muy difícil agruparlos para considerarla como una variable categórica.
- El idioma de los tweets es inglés.
- La longitud media de los tweets es de 100 caracteres, presentando un pico de densidad en 135 caracteres.
- Los tweets contienen en promedio 15 palabras.
- Las palabras que más se repiten el los tweets son ‘the’, ‘a’, ‘in’ todas ellas consideradas stopwords y de allí su importancia en la limpieza de los datos.
- Sólo el 22.8% de los tweets contienen hashtags, de los cuales los más frecuentes son ‘news’, ‘prebreak’ y ‘hot’
- El 52.2% de los tweets contienen URLs y por ello es importante suprimirlos para análisis posteriores, ya que afectan la longitud de los tweets.
- No se pudo establecer una correlación entre keyword y location.
- Para la keyword ‘oubreak’ la mayoría de los tweets son verdaderos.
- Para la keyword ‘body%20bags’ la mayoría de los tweets son falsos.
- La media de los tweets falsos es de 100 caracteres, mientras que en los tweets verdaderos es de 115 caracteres.
- Un altísimo porcentaje de los tweets contienen textos con un solo párrafo.