UNIVERSIDAD DE BUENOS AIRES FACULTAD DE INGENIERÍA

Trabajo Práctico N°1

Análisis exploratorio de datos sobre Tweets

Primer Cuatrimestre de 2020

Alumno	Padrón	Mail
Cenizo, Facundo	96657	facundocenizo@hotmail.com
Cruz, Pablo	97553	cruzpa95@gmail.com
Vicente, Braian	96542	braianvicente@gmail.com

Repositorio: https://github.com/BraianVicente/7506-real_or_not

Repositorio: https://github.com/BraianVicente/7506-real_or_not

INTRODUCCIÓN

Acerca de Twitter

Twitter es un servicio de microblogging, La red permite enviar mensajes de texto plano de corta longitud llamados tweets, que se muestran en la página principal del usuario. Los usuarios pueden suscribirse a los tweets de otros usuarios – a esto se le llama "seguir" y a los usuarios abonados se les llama seguidores,11 followers. Por defecto, los mensajes son públicos, pudiendo difundirse privadamente mostrándose únicamente a unos seguidores determinados.

Twitter se ha convertido en un importante canal de comunicación en tiempos de emergencia.
La ubicuidad de los teléfonos inteligentes permite a las personas anunciar una emergencia que están observando en tiempo real. Debido a esto, más agencias están interesadas en monitorear programáticamente Twitter (es decir, organizaciones de ayuda ante desastres y agencias de noticias).

OBJETIVO

El objetivo del primer TP es realizar un análisis exploratorio del set de datos. Queremos ver qué cosas podemos descubrir sobre los datos que puedan resultar interesantes. Estas cosas pueden estar relacionadas al objetivo del TP2 (predecir si un cierto tweet es real o no) o no, ambas son de interés.

Repositorio: https://github.com/BraianVicente/7506-real_or_not

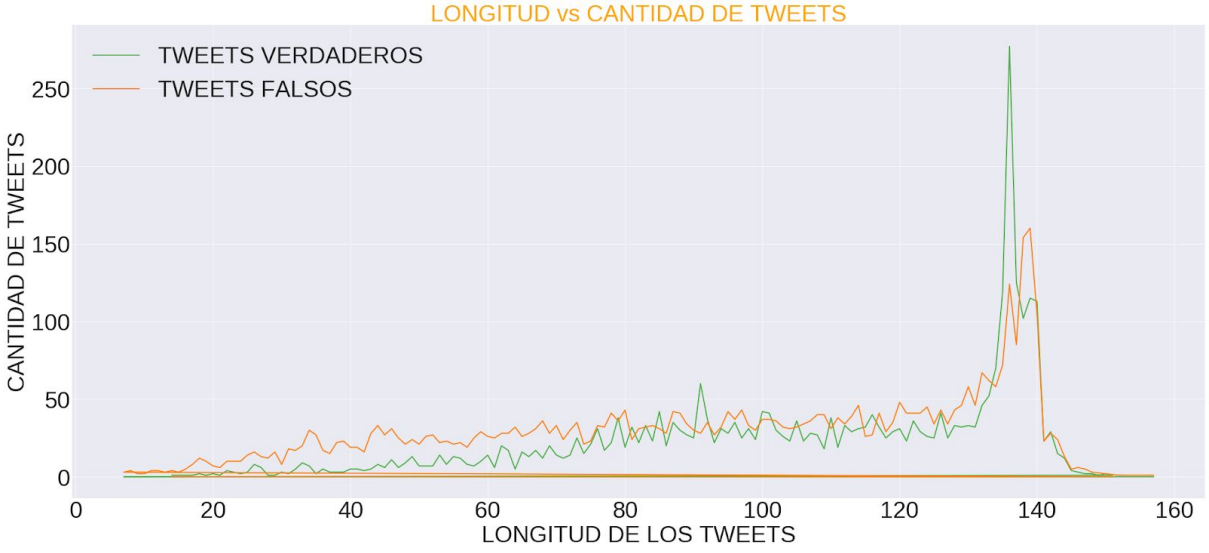
ANÁLISIS

Longitud y veracidad

Como un análisis preliminar comenzamos buscando la cantidad de tweets según su veracidad:

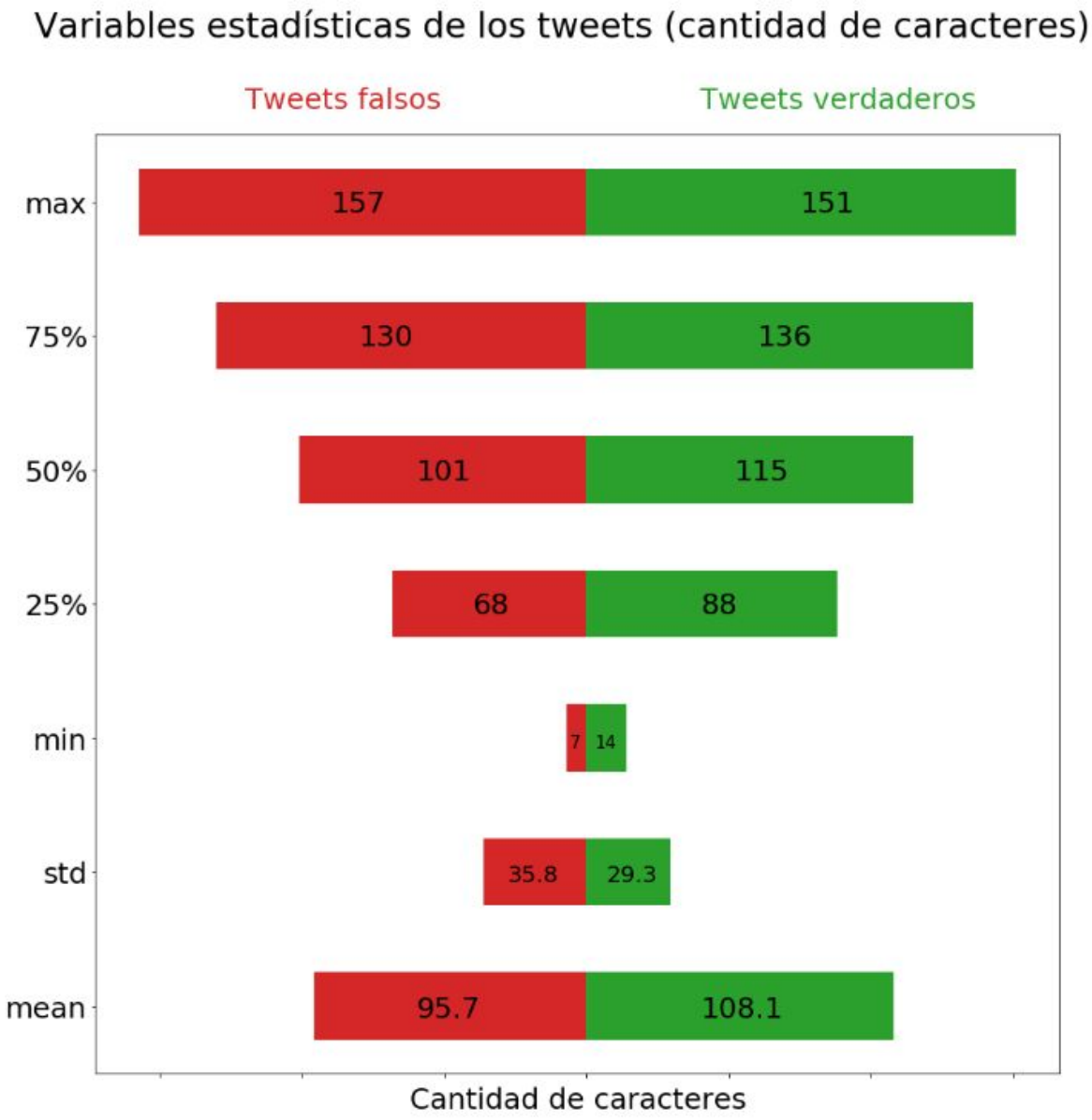
Podemos observar que hay más cantidad de tweets falsos que verdaderos.

A continuación, analizamos la relación entre longitud, cantidad y veracidad de los tweets.



Pudimos encontrar una clara tendencia de que, para todas las longitudes, hay más tweets falsos que verdaderos, exceptuando el rango alrededor de los 90 y 137 caracteres donde para los tweets verdaderos hay dos picos en donde se invierte la relación.

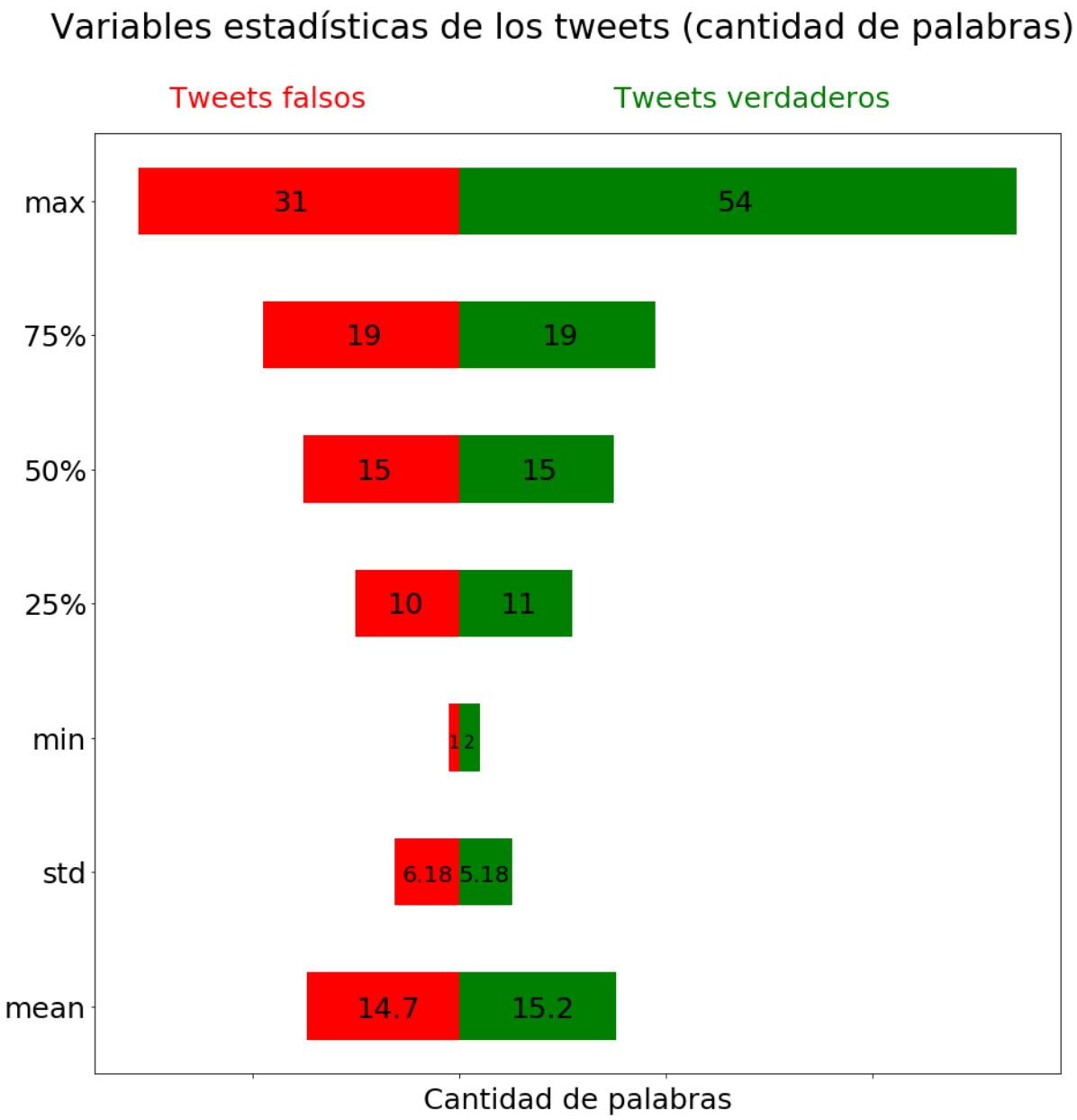
Repositorio: https://github.com/BraianVicente/7506-real_or_not



Analizando las variables estadísticas de los tweets en base a la cantidad de caracteres, se puede observar que los tweets verdaderos son en promedio más largos que los tweets falsos, que la desviación estándar de los falsos es mayor que la de los verdaderos y que los cuantiles de los verdaderos son mayores que los de los falsos.

Repositorio: https://github.com/BraianVicente/7506-real_or_not

Por otro lado podemos analizar cantidad de palabras y veracidad del tweet:



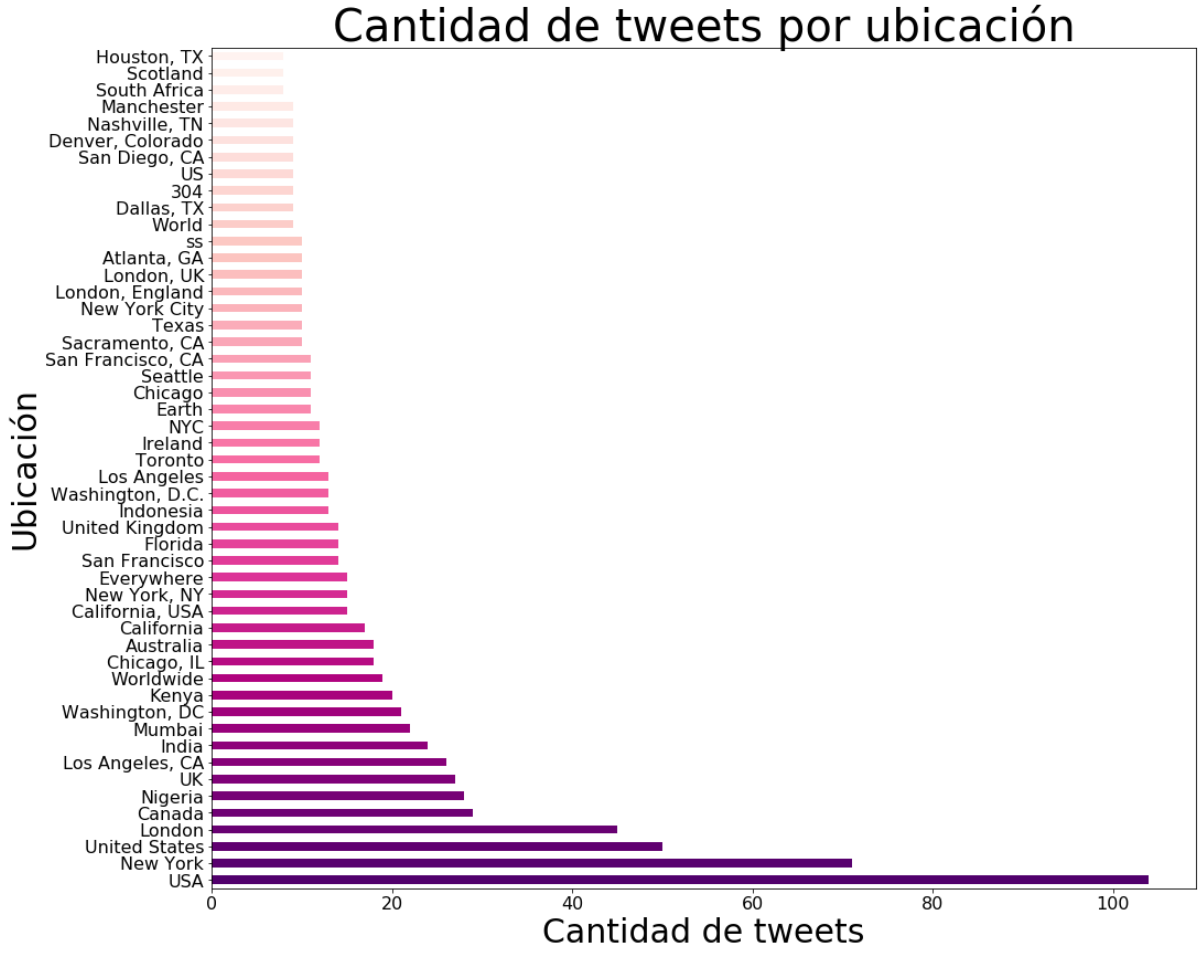
Al igual que en el gráfico de cantidad de caracteres, podemos observar que se mantiene la tendencia de las variables estadísticas.

Repositorio: https://github.com/BraianVicente/7506-real_or_not

Ubicación geográfica

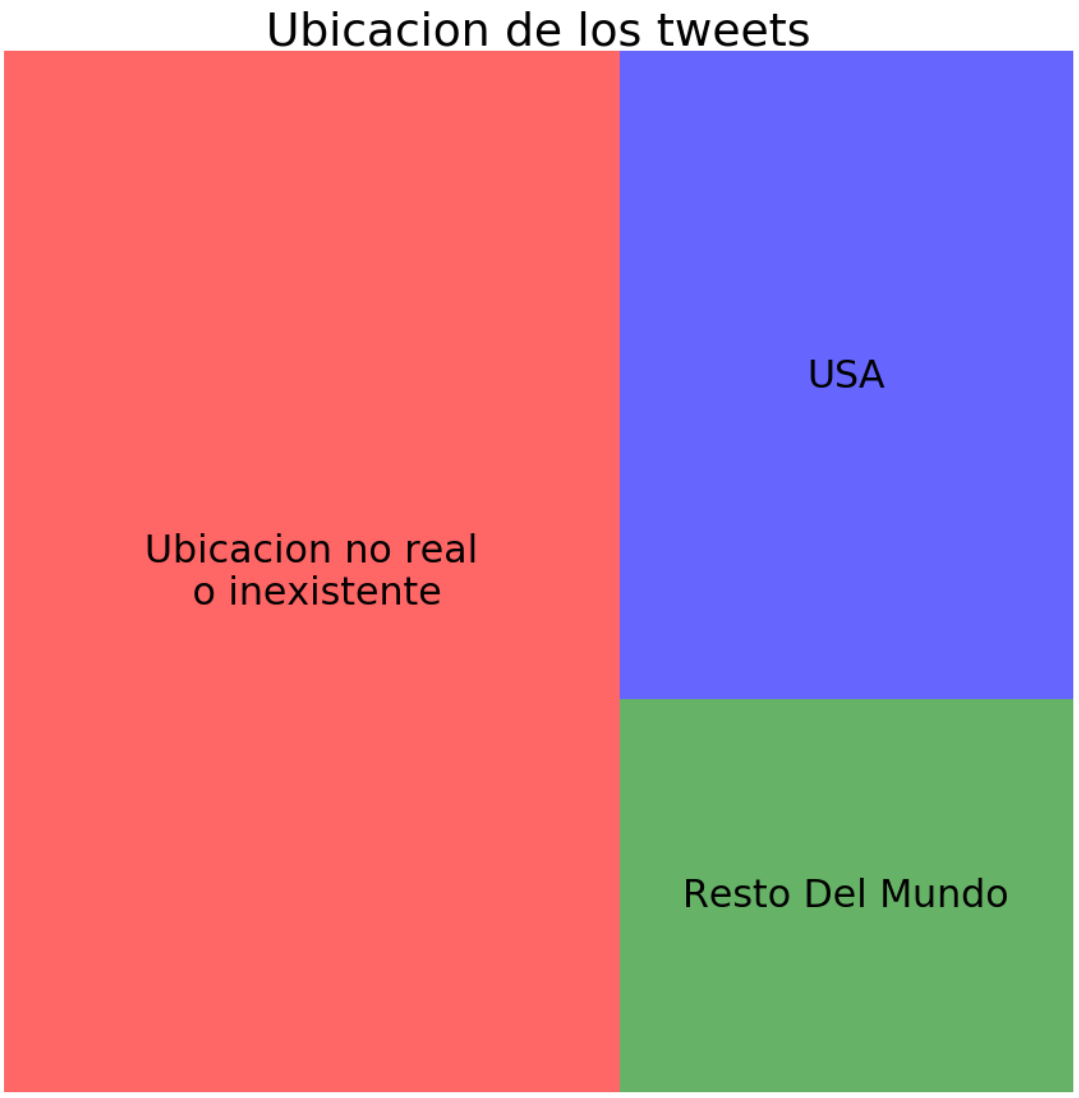
Aprovechando que el set de datos nos provee la ubicación(location) de la publicación de los tweet, decidimos investigar las ubicaciones.

En el siguiente gráfico, podemos observar que las ubicaciones pueden tomar cualquier valor como por ejemplo: Earth, 304, Everywhere, Worldwide, entre otros. También podemos encontrarnos con países o ciudades, ya sea con el nombre completo o abreviaturas.



Como la mayoría de las ubicaciones pertenecen a estados o ciudades de Estados Unidos, decidimos hacer foco en estos datos y filtrarlos de manera tal que nos quede por un lado los situados en USA, los el resto del mundo y los “Ubicación no real o inexistente”.

Repositorio: https://github.com/BraianVicente/7506-real_or_not

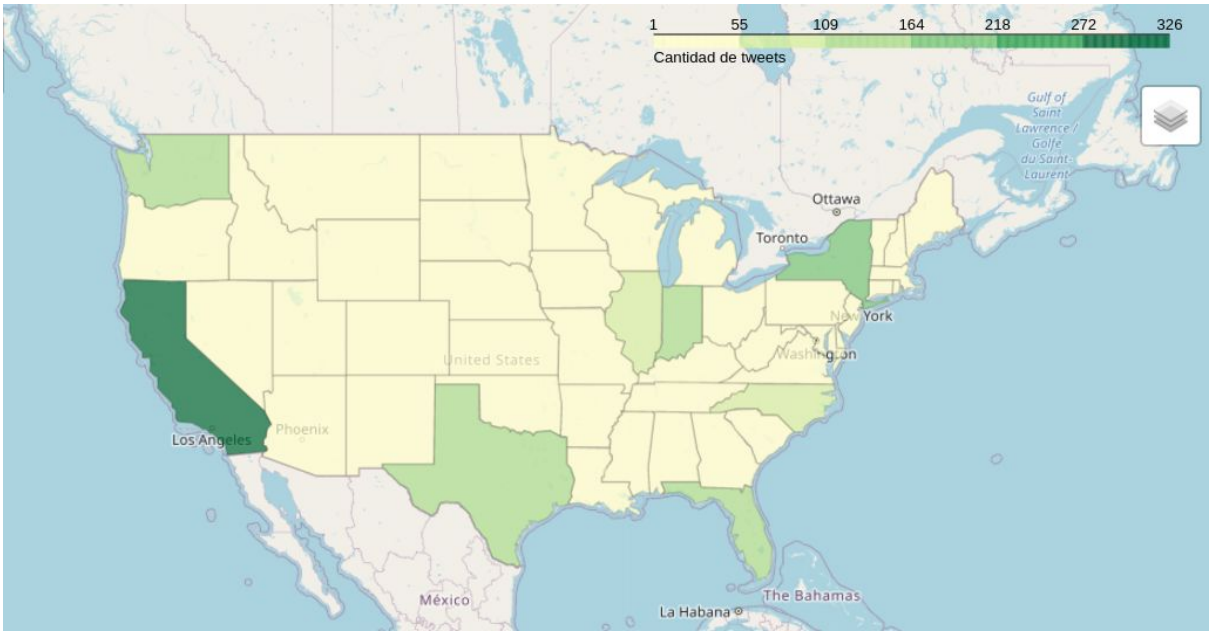


Podemos apreciar que la gran mayoría de los tweets poseen una ubicación no real o inexistente.

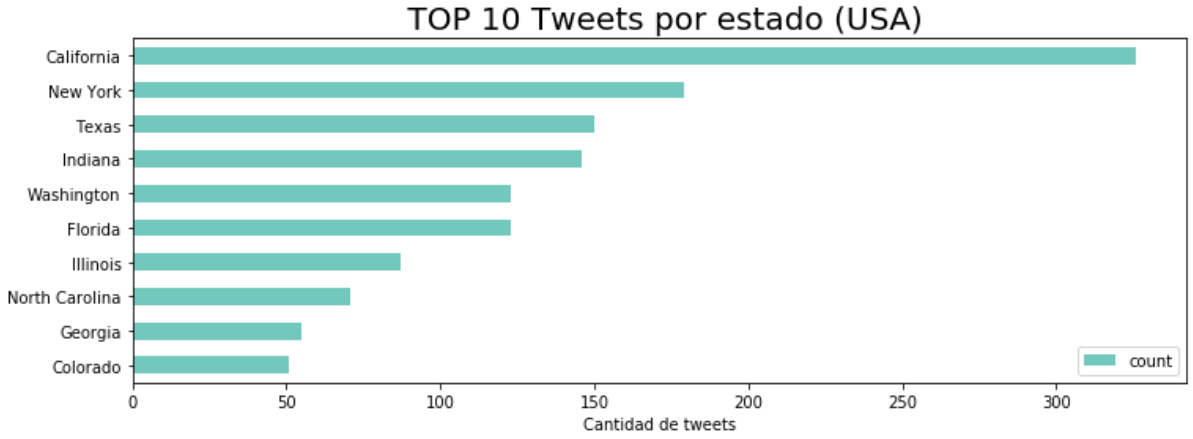
A pesar de que la cantidad de tweets con ubicación en USA no representa la cifra más significativa, nos pareció interesante generar el siguiente mapa y entender la distribución geográfica de los tweets pertenecientes a este pais:

Repositorio: https://github.com/BraianVicente/7506-real_or_not

Mapa de cantidad de tweets generados por cada estado.



En el siguiente gráfico mostramos el top 10 de estados con mayor cantidad de tweets:



Podemos observar que California, New York y Texas posee la mayor cantidad de tweets, esto puede deberse a que son los estados con mayor población.

Los estados más poblados de Estados Unidos son:

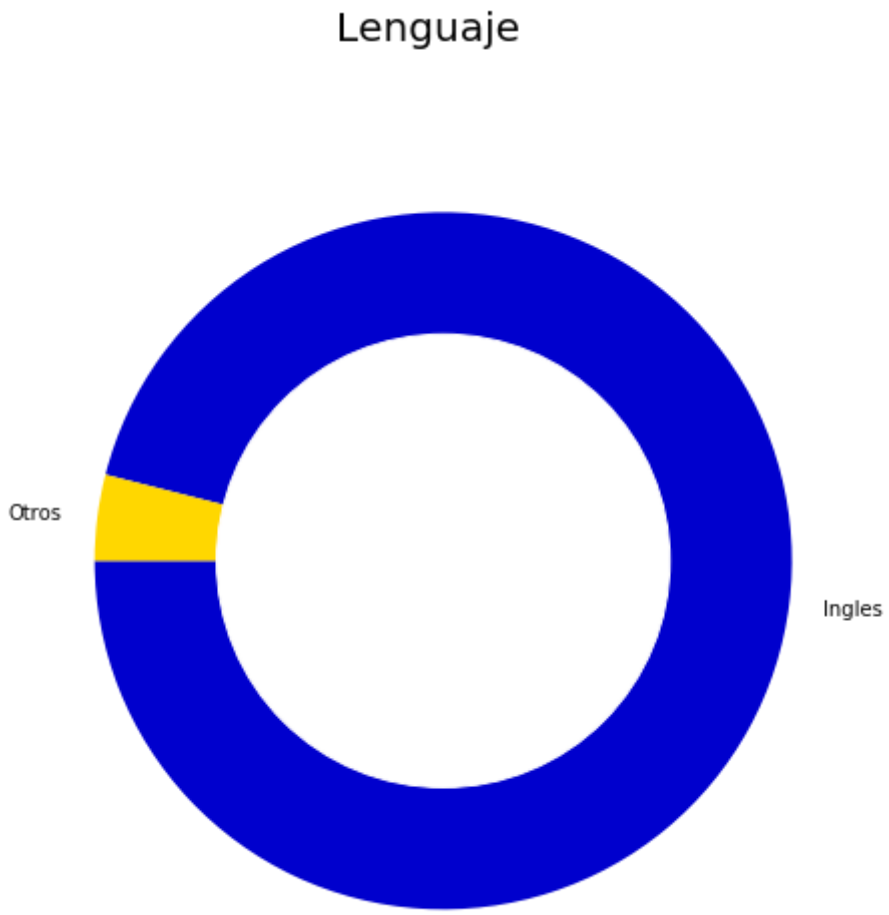
- California: 38.041.430 habitantes.
- Texas: 28.701.845 habitantes.
- Florida: 21.299.325 habitantes.
- Nueva York: 19.542.209 habitantes.

https://es.wikipedia.org/wiki/Anexo:Estados_de_los_Estados_Unidos_por_poblaci%C3%B3n

Repositorio: https://github.com/BraianVicente/7506-real_or_not

Lenguaje y gramática

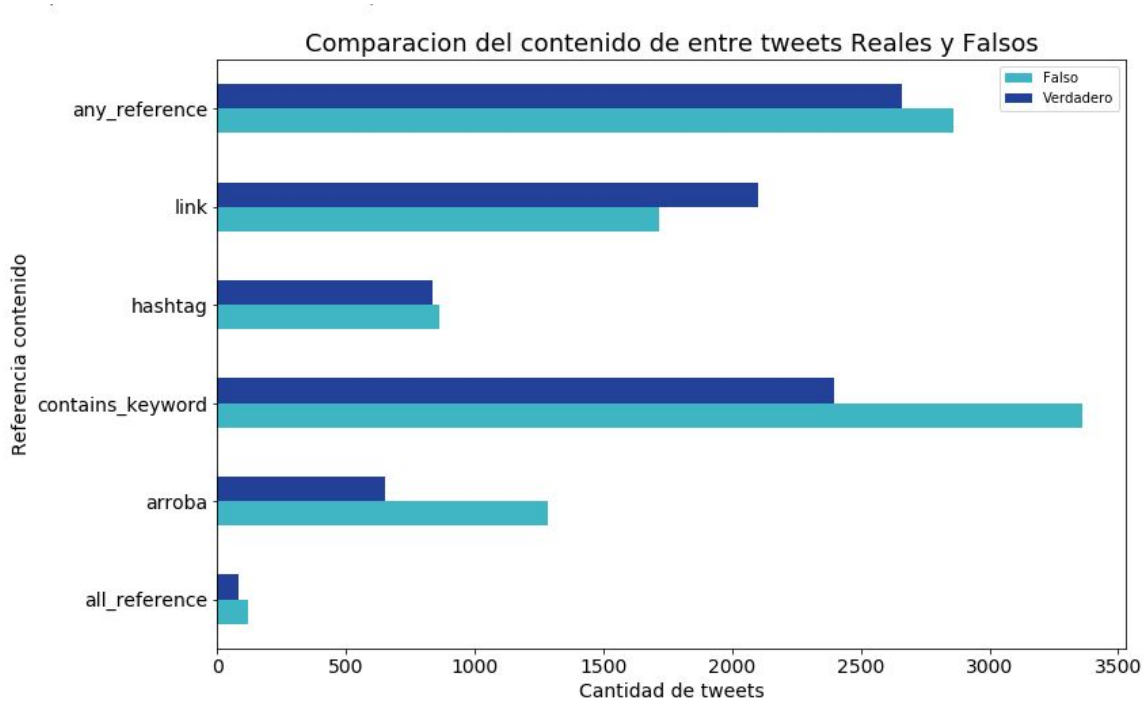
Como notamos que gran cantidad de tweets provenían de Estados Unidos nos llevó a investigar los lenguajes que se encuentran presentes en el texto de cada tweet.
Esto lo llevamos a cabo con el framework langdetect.



Con esto pudimos concluir que casi la totalidad de los tweets están escritos en Inglés.

Siguiendo con la dinámica de analizar el contenido de cada tweet, nos dispusimos analizar las referencias que contienen, es decir, link(http), hashtags(#), arrobas(@) y además si los tweets contienen las palabras claves(keywords).

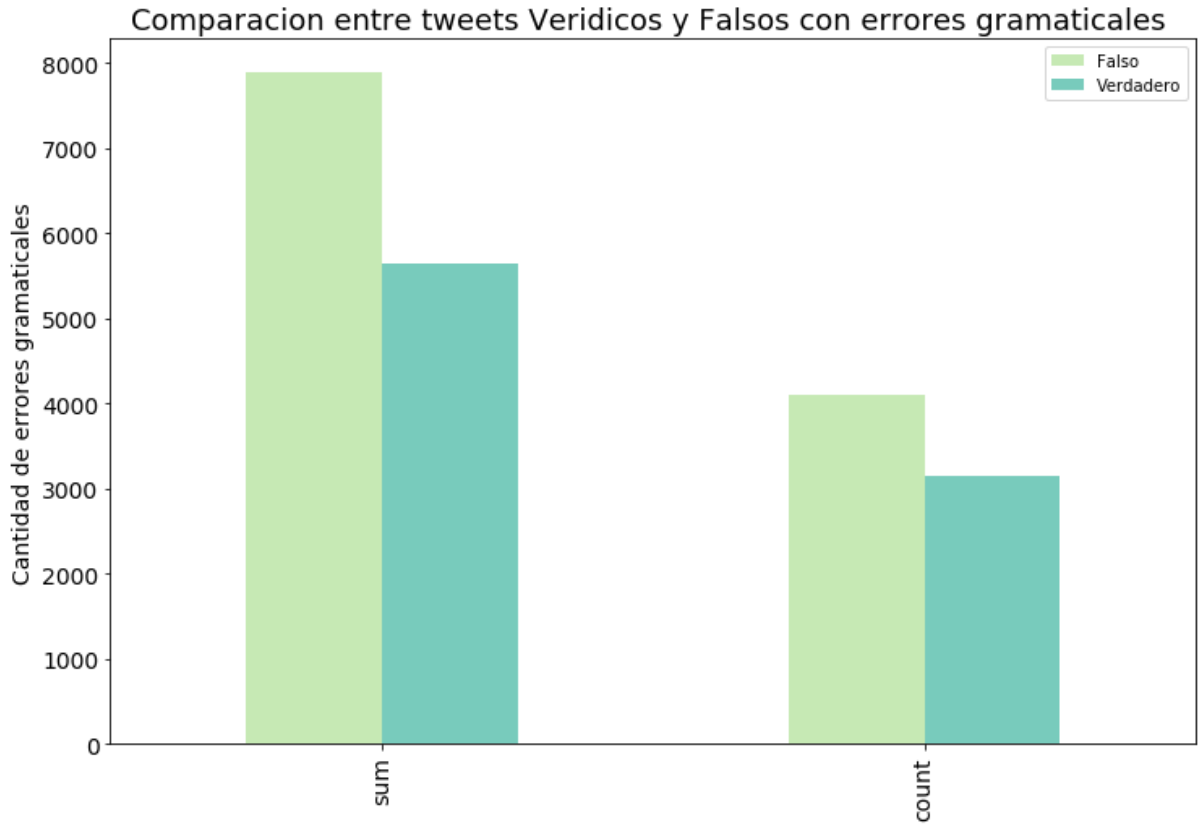
Repositorio: https://github.com/BraianVicente/7506-real_or_not



No se logra distinguir una tendencia respecto a esta comparación. Aun que, en particular, en el caso del arroba (@) hay una diferencia significativa que puede relacionarse con el hecho de que en la plataforma este se utiliza para mencionar a alguien y/o responder un tweet, que interpretamos, se debe a que significa una utilización más informal de este tipo de recursos de twitter.

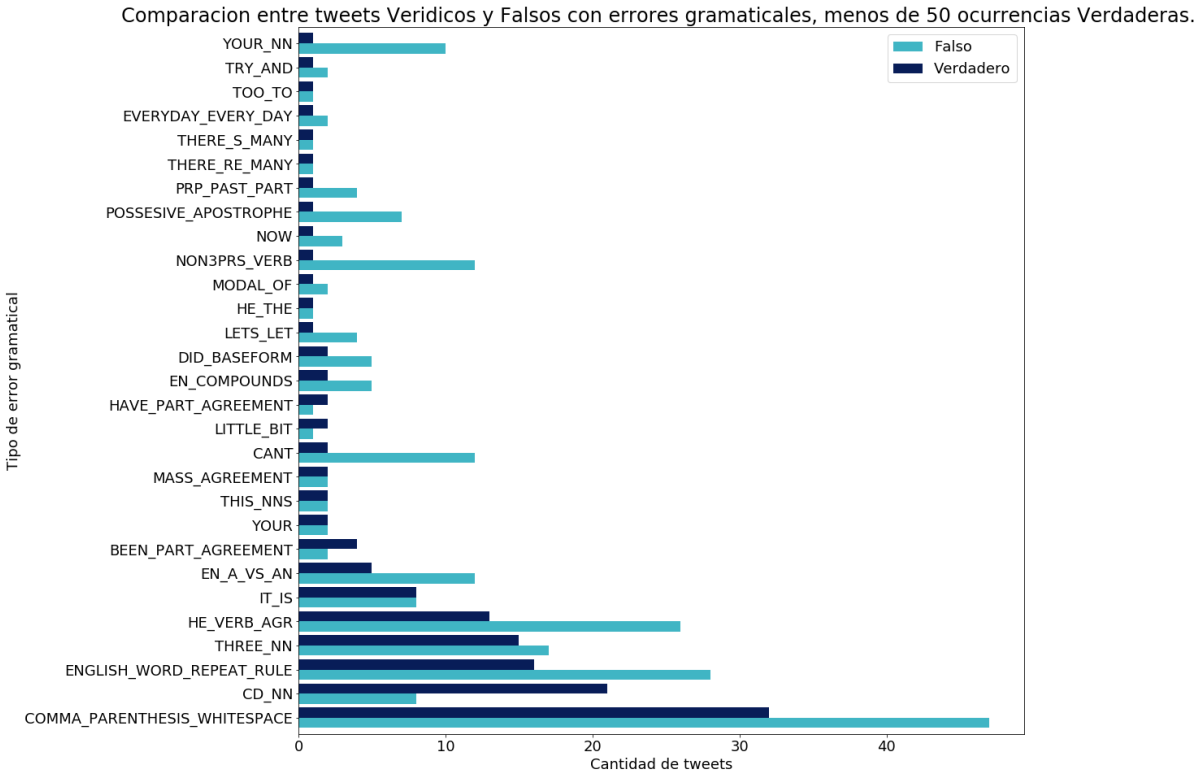
Es por esto, que decidimos hacer un análisis gramatical para entender que tantos errores gramaticales hay en los tweets, utilizando la librería language_check.

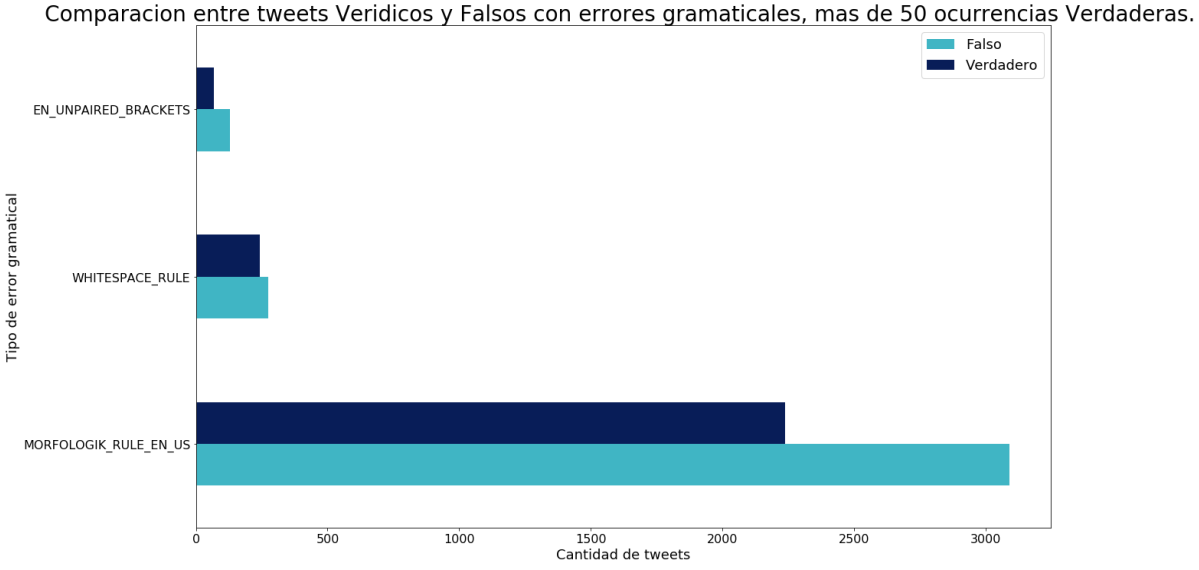
Repositorio: https://github.com/BraianVicente/7506-real_or_not



Aquí se graficó por un lado la suma total de errores, ya que identificamos que hay textos que pueden contener más de un error gramatical de distintos tipos.
A partir de esto decidimos investigar más a fondo de qué tipo errores gramaticales se trataban.
Para una mejor apreciación de los datos, decidimos dividir entre los que cuentan con menos de 50 ocurrencias y posteriormente con más de 50 ocurrencias, ya que hay una gran dispersión de datos respecto a los errores gramaticales presentes.

Repositorio: https://github.com/BraianVicente/7506-real_or_not



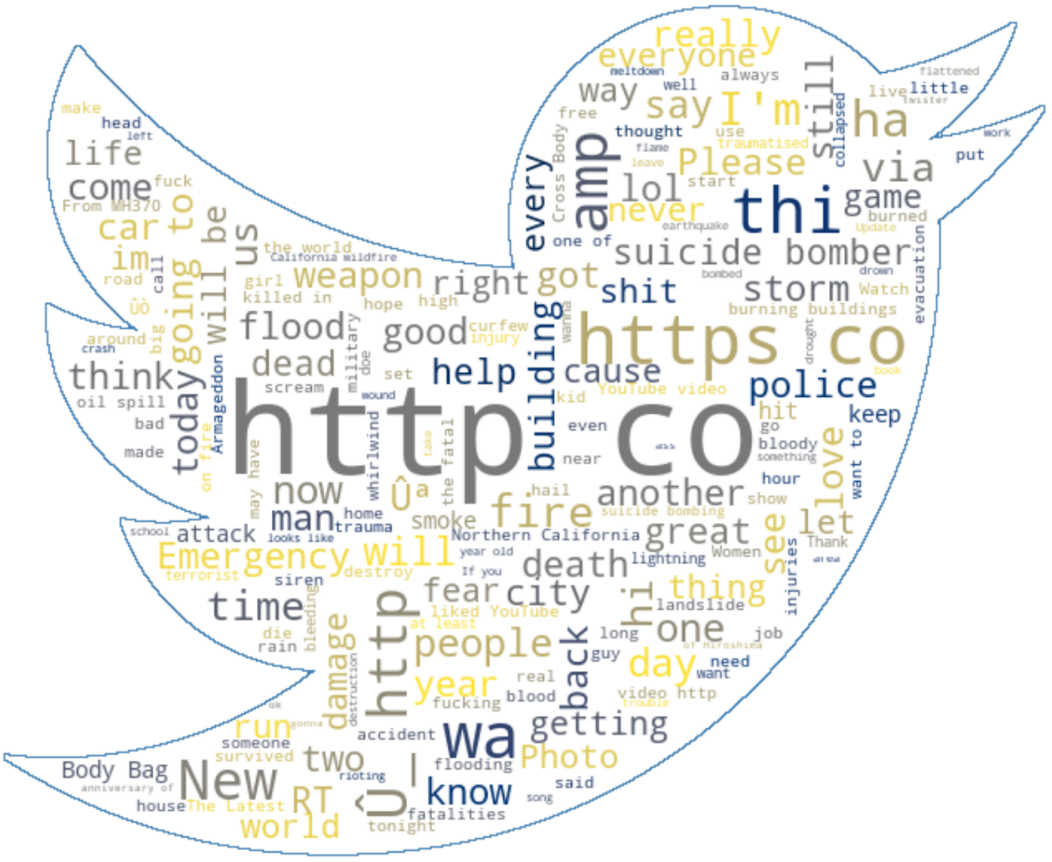


En los dos gráficos anteriores se distingue que los tweets falsos tiene mayor ocurrencia sobre los verdaderos.

Repositorio: https://github.com/BraianVicente/7506-real_or_not

Frecuencia de palabras

En el siguiente wordcloud se puede observar que las palabras más frecuentes tienen relación con lo que es compartir links, como por ejemplo 'http', 'via', entre otras. También se ven palabras como 'New', 'Emergency', 'fire', 'people', Please.

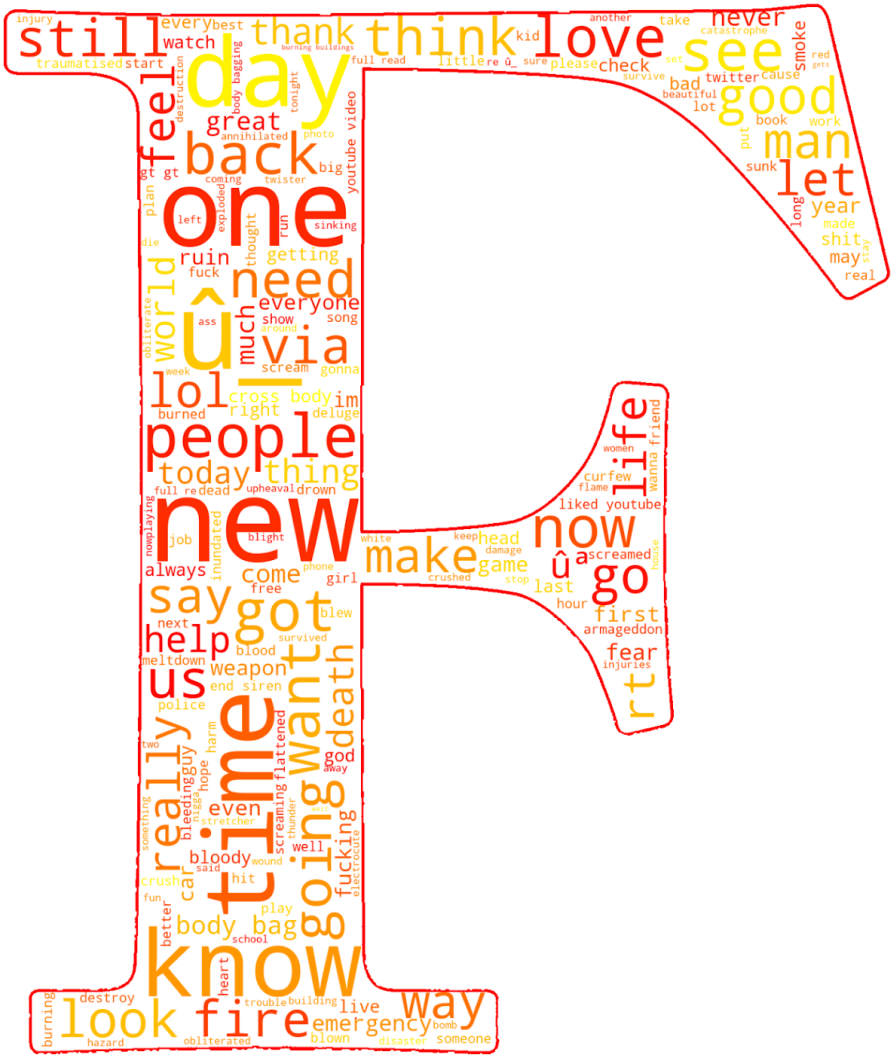


Repositorio: https://github.com/BraianVicente/7506-real_or_not

A continuación, analizaremos las palabras más frecuentes tanto para los tweets verdaderos como para los falsos.

Entre las palabras más utilizadas en los tweets verdaderos podemos encontrar: “Death”, “flood”, “new”, “fire”, “storm”, “suicide”, “bomber”. Esto nos puede hacer pensar que cuando se nombran palabras haciendo referencia a alguna catástrofe o algún tema sensible, puede ser un buen indicio para predecir la veracidad.

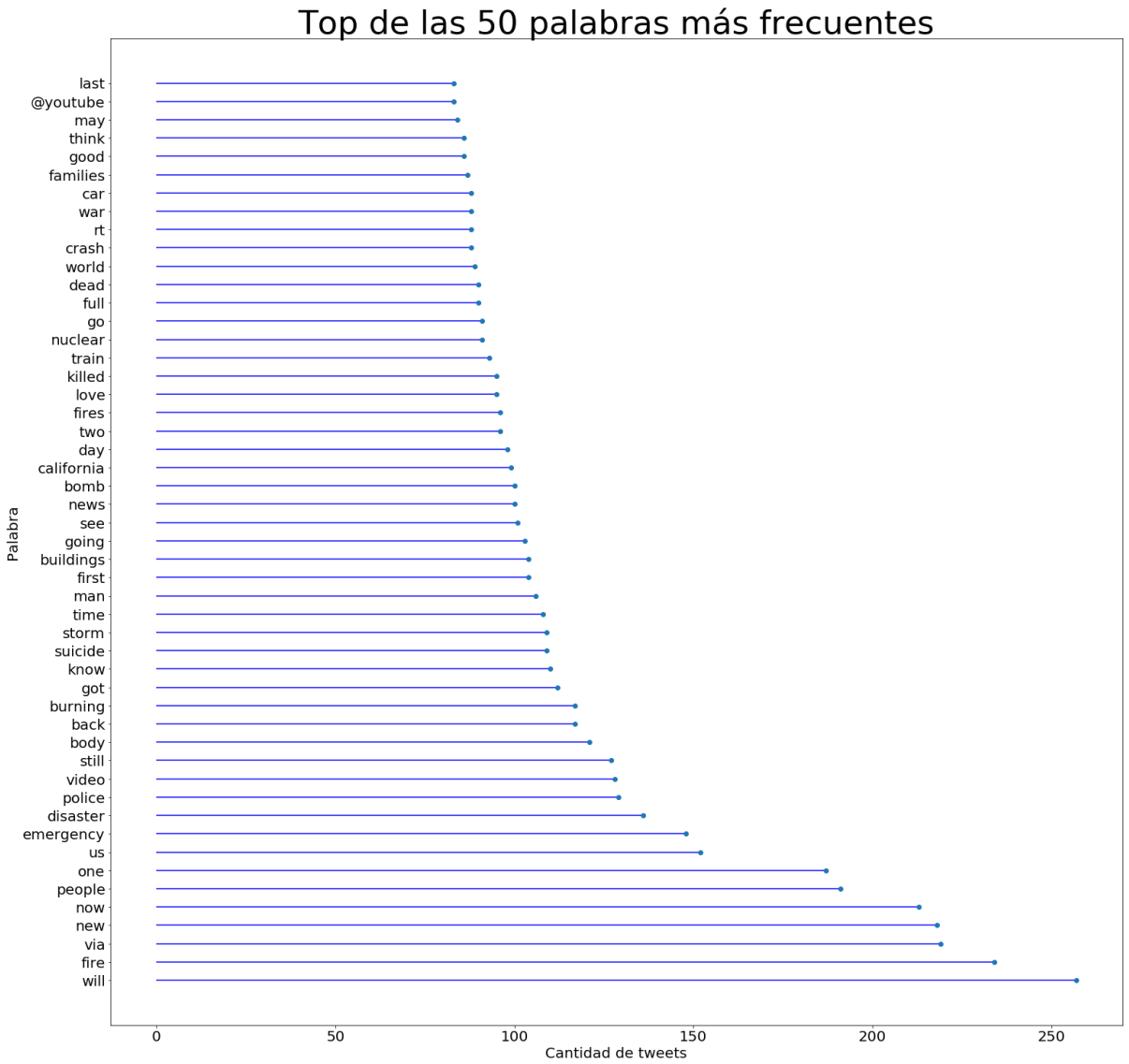
Repositorio: https://github.com/BraianVicente/7506-real_or_not



Entre las palabras más utilizadas en los tweets falsos podemos encontrar: “Love”, “New”, “Time”, “One”, “Think”, “Now”, “lol”, entre otros.

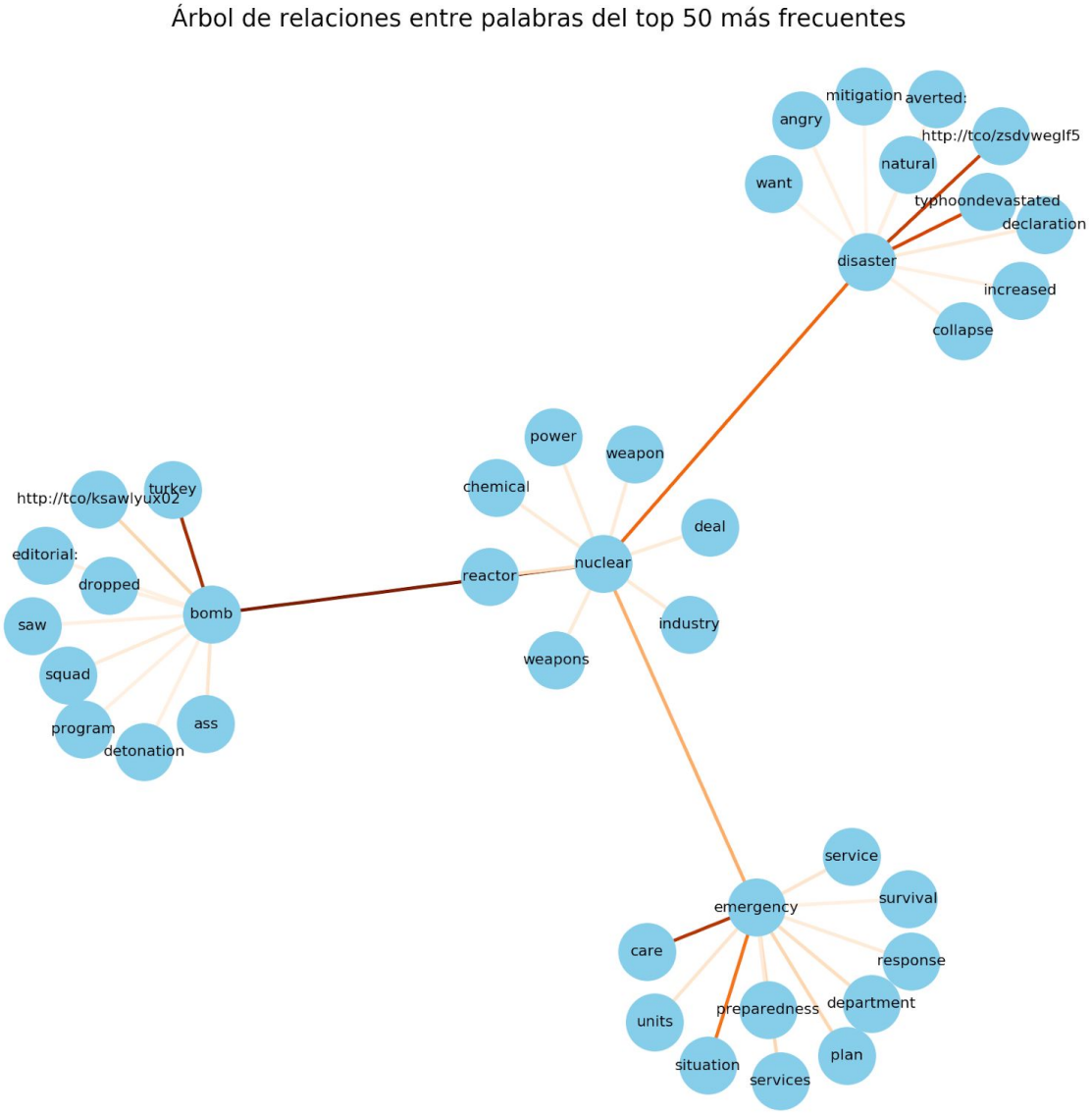
Repositorio: https://github.com/BraianVicente/7506-real_or_not

A continuación, analizamos frecuencias de las palabras en todos los tweets



Observando el Top 50 de las palabras más utilizadas, nos pareció interesante vincular ciertas palabras en una network, donde la tonalidad de las aristas nos dan información sobre la cantidad de veces que se conectan entre sí.

Repositorio: https://github.com/BraianVicente/7506-real_or_not



Se puede ver que “Bomb” se relaciona fuertemente con “Turkey”, esto lo podemos relacionar con los atentados ocurridos en dicho país.
“Nuclear” se relaciona tanto con “weapon” (es decir “armas nucleares”) como con “industry” y “reactor” (es decir plantas nucleares).
“Emergency” se relaciona principalmente con “care” y “situation”.
Y “Disaster” está ligado con “typhoon devastated” (lo cual podemos suponer que es una noticia sobre un accidente relacionado a las áreas devastadas por un ciclón).

Repositorio: https://github.com/BraianVicente/7506-real_or_not

CONCLUSIONES

- Hay más cantidad de tweets falsos que verdaderos.

- Los tweets verdaderos son en promedio más largos que los tweets falsos y la desviación estándar de los falsos es mayor que la de los verdaderos.

- La gran mayoría de los tweets poseen una ubicación no real o inexistente, como por ejemplo “Everywhere”, “behind you”, etc.

- La mayoría de las ubicaciones reales pertenecen a estados o ciudades de Estados Unidos.

- Casi la totalidad de los tweets se encuentra en el idioma inglés.

- Los tweets que contienen menciones (@) son más propensos a ser falsos que verdaderos.

- Los tweets falsos contienen mayores errores gramaticales que los verdaderos.

- Creemos que puede ser un buen indicio para identificar la condición de verdadero de un tweet cuando aparecen palabras haciendo referencia a alguna catástrofe o algún tema sensible, como por ejemplo: “Death”, “flood”, “fire”, “storm”, “suicide”, “bomber”.

Repositorio: https://github.com/BraianVicente/7506-real_or_not