

UNIVERSIDAD DE BUENOS AIRES

FACULTAD DE INGENIERÍA

**75.06 Organización de Datos**

**Trabajo Práctico 1**

**Primer Cuatrimestre de 2020**

Grupo 27

|  |  |
| --- | --- |
| Joaquin Lopez Saubidet | 99252 |
| Santiago Tadini | 104439 |
| Tomas Sabao | 99437 |
| Zugna, Federico | 95758 |

**Link de GitHub:** <https://github.com/tomasSabao/Organizacion_de_datos_tp1>

Índice

[1. Introduccion 3](#_Toc40380977)

[2. Informe 4](#_Toc40380978)

[3. Conclusión Final 4](#_Toc40380979)

# Introduccion

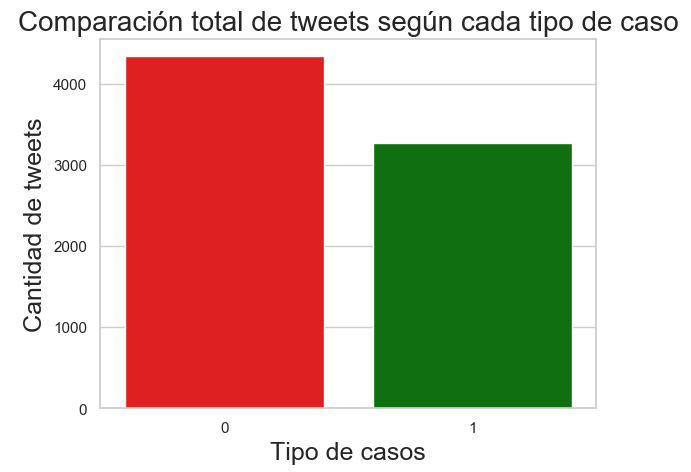
Este informe se encarga de analizar los datos sobre un conjunto de los tweets del set de datos de la competencia: <https://www.kaggle.com/c/nlp-getting-started>.

* id - identificador unico para cada  tweet
* text - el texto del tweet
* location - ubicación desde donde fue enviado (podría no estar)
* keyword - un keyword para el tweet  (podría faltar)
* target - en train.csv, indica si se trata de un desastre real  (1) o no (0)

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.

## Informe

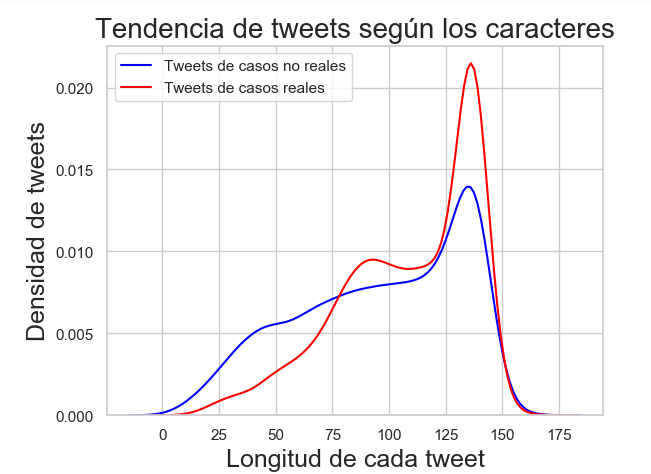
Dentro del archivo podemos ver como cada uno de los tweets presentes tienen un campo target que nos indica si es un caso real o no lo es. De todo el set de datos, comparamos la cantidad de tweets que hay para cada caso.



Se puede observar que hay una mayor cantidad de tweets acerca de casos no reales. En un 57% del total de los tweets son de casos no reales, y por consecuencia, un 43% son casos reales.

Se sabe que la cantidad máxima de caracteres por cada tweet es de 280 caracteres. El promedio de longitud de todos los tweets es de 101 caracteres aproximadamente.

Se puede ver que los tweets que tienen entre aproximadamente 80 caracteres y 150 son en los que predominan los casos reales, en cambio cuando es menor a 75 caracteres y mayor a 150 caracteres la mayoría de tweets son de casos no reales



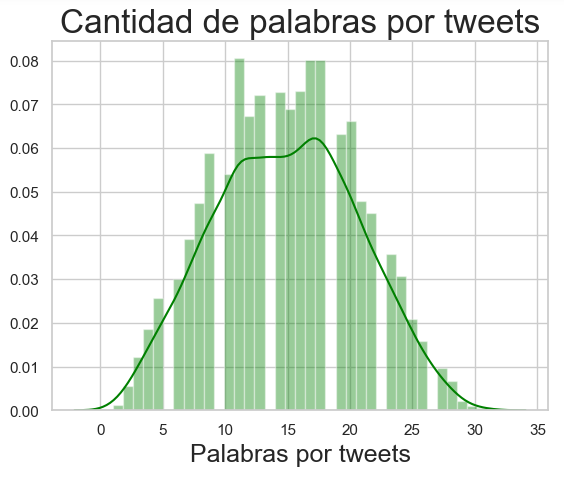
Pues tanto los tweets con menor (7 caract.) longitud como el mayor (163 caract.) se corresponden a tweets de casos no reales.

Acá podemos ver como las palabras más usadas en cada tweet

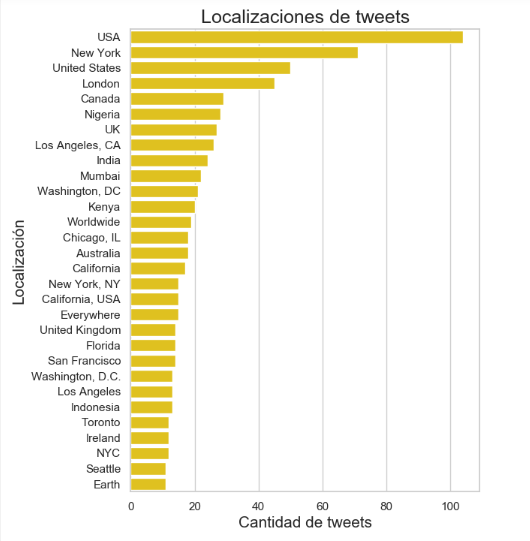


También se puede ver la cantidad de palabras por tweets, predominan los tweets que están entre 10 y 20 palabras.

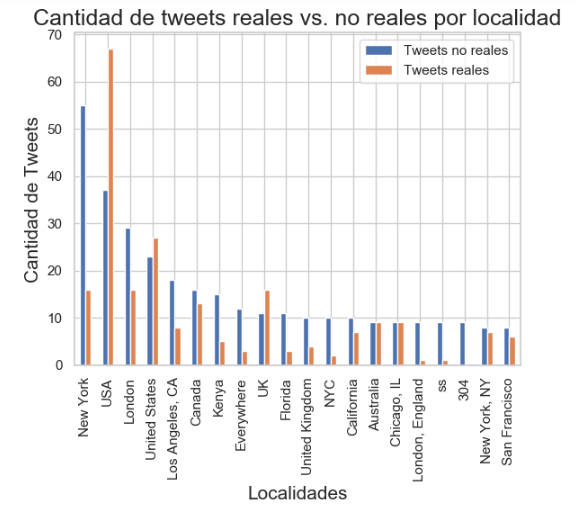
**Obs:** Se considera palabra a todo lo que esté precedido de un espacio.



Del archivo de tweets tenemos la localización de algunos de estos, por lo que nos fijamos de donde vienen cada uno de estos y notamos que la mayoría fueron realizados en Estados Unidos, pues los tres primeros valores son Usa, United States y New York. También se observa que de los primeros 30 lugares de donde hay mayor cantidad de tweets realizados, el 50% son ciudades de EEUU.

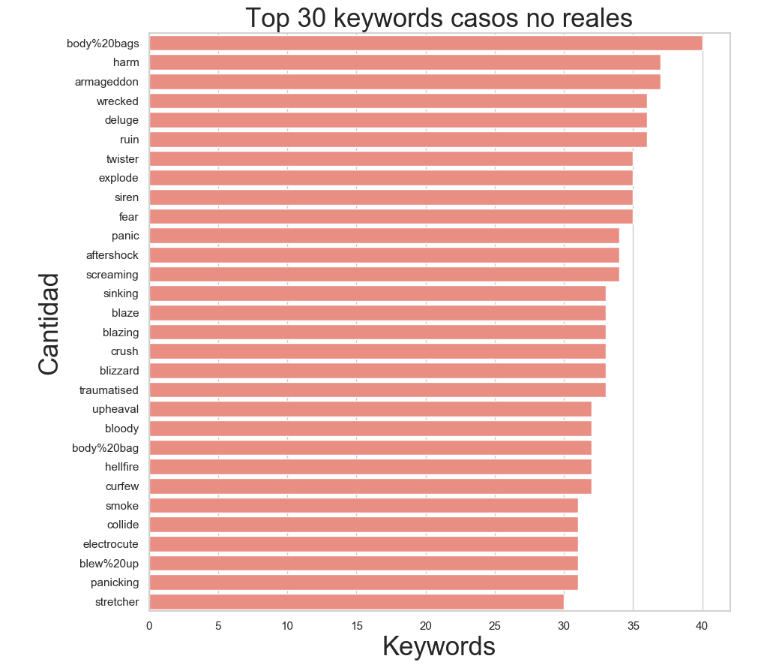


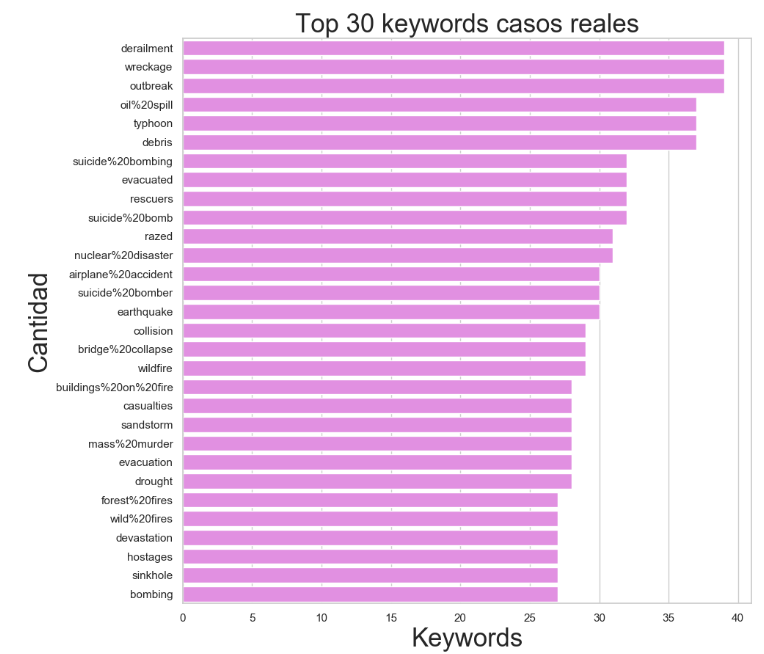
Pero a la vez, separando los casos reales de los no reales, solo en algunos países se logra tener una mayor cantidad de tweets acerca de casos reales.



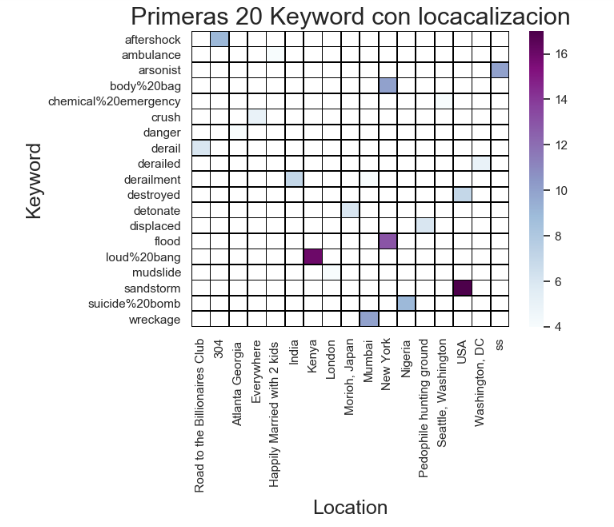
Separando las keywords de los casos reales y no reales, ninguna keyword está dentro del top 30 en ambos casos, siempre pertenece a algún grupo más fuerte que al otro.

También se puede notar que siempre las primeras 6 keywords de cada grupo están por encima de las 35 repeticiones:





Para las keyword en cada localidad, solo en la localización USA hay dos palabras con dentro del top 20 de keywords. También se puede ver que hay localizaciones que no son ciudades o países, y tienen un numero bajo de repeticiones de la misma keyword, por lo que se puede pensar que es la misma persona o gente cercana a esta.



## Conclusión Final