Universidad de Buenos Aires
Facultad de Ingeniería

75.06/95.58 Organización de Datos
Primer Cuatrimestre de 2020
Trabajo Práctico 1

Grupo NameError

Julián Comandé
Omar Cardenas
Daniel Collico
Lucas Escudero

Link al repositorio: https://github.com/MonoColorado/7506-NameError

Contenido

Introducción

Este informe busca analizar los tweets del set de datos de la competencia de Kaggle https://www.kaggle.com/c/nlp-getting-started.

Los campos que contiene son:

- `id` - identificador único 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 de la competencia de Kaggle es predecir si un tweet es sobre un desastre real o no. En este análisis buscamos entender los datos, no necesariamente para acercarnos al objetivo.

Análisis general de los datos

El set de datos de train.csv contiene los siguientes datos:

```
RangeIndex: 7613 entries, 0 to 7612
Data columns (total 5 columns):
id          7613 non-null int64
keyword     7552 non-null object
location    5080 non-null object
text        7613 non-null object
target      7613 non-null bool
```

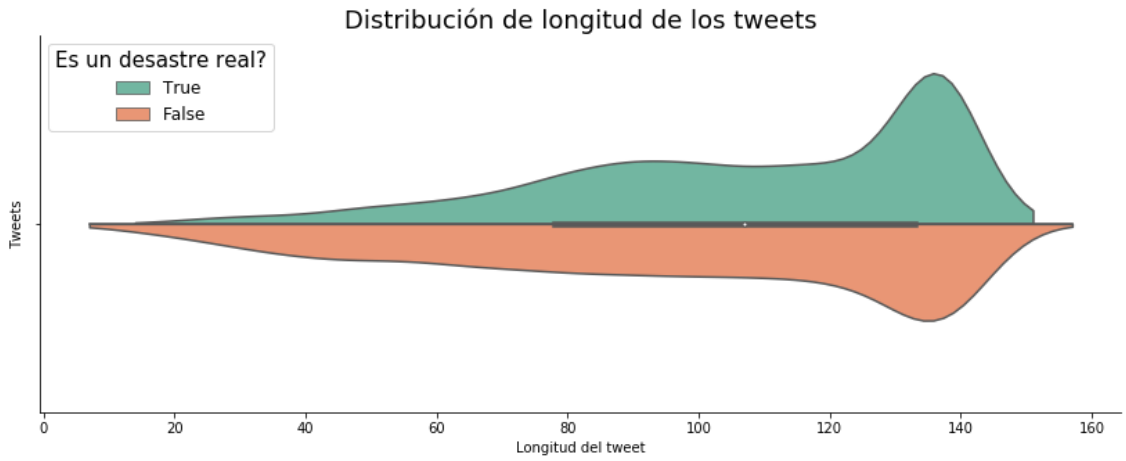
Validamos los datos del enunciado: `id`, `text` y `target` no contienen valores nulos. La gran mayoría de los tweets tienen una `keyword` asignada. Alrededor de 2/3 de los datos contienen algún valor en el campo `location`.

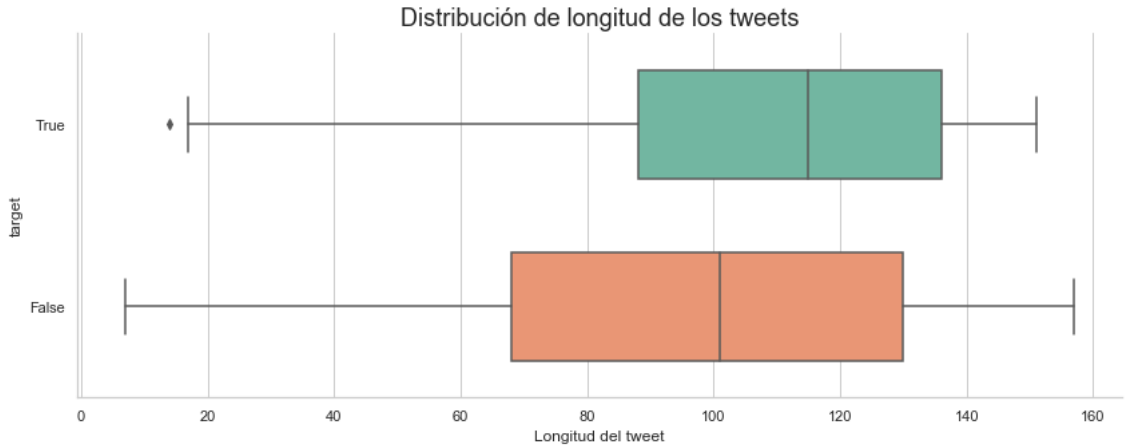
Analizando el contenido de `target` podemos obtener los siguientes conteos:

```
False    4342
True     3271
Name: target
```

Tenemos una gran cantidad de datos, relativamente balanceada entre tweets sobre desastres y tweets sobre falsos desastres.

Longitud de los tweets





Como podemos observar en los gráficos, los tweets sobre desastres reales tienden a ser más largos:

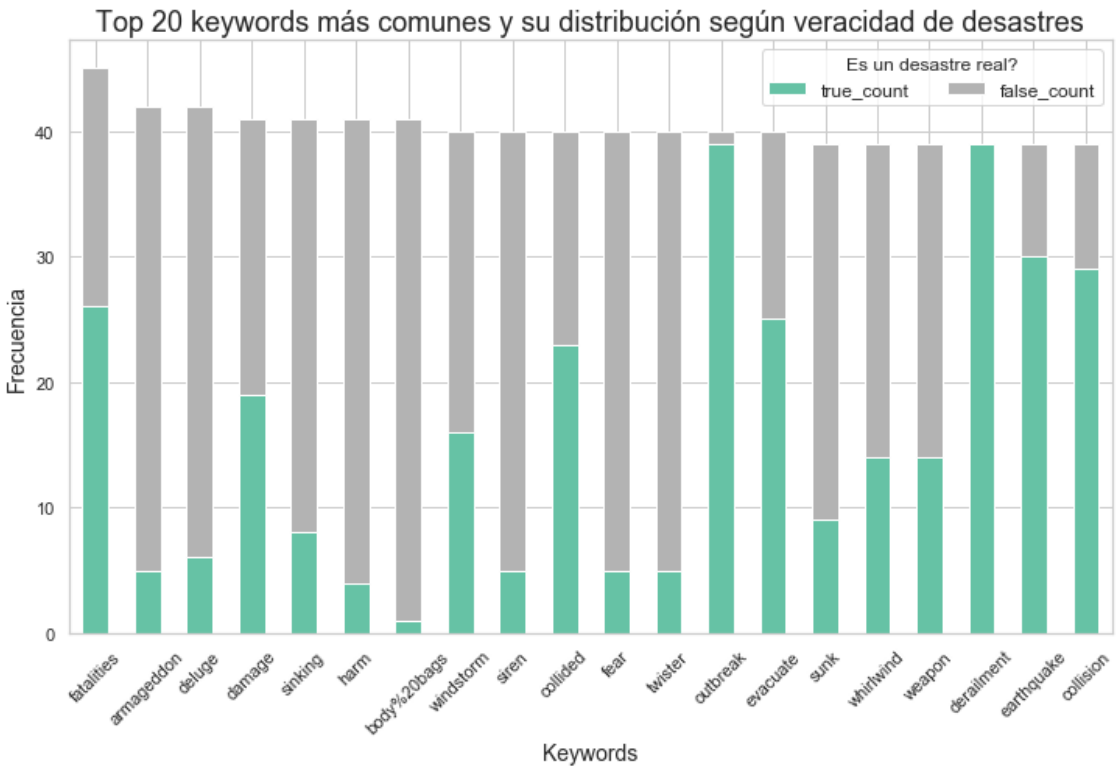
El 50% de todos los tweets tienen entre 80 y 140 caracteres y un pico importante entre los 130 y 140 caracteres.

En el caso de los desastres verdaderos, esta distribución es mucho más compacta, se sitúa entre los 90 y 135 caracteres. Y la proporción en longitudes menores a 60 caracteres es más pequeña comparada a los tweets falsos.

En el caso de los desastres falsos, el 50% de los tweets están entre los 70 y 130 caracteres, con una distribución más plana y uniforme, incluso en el pico.

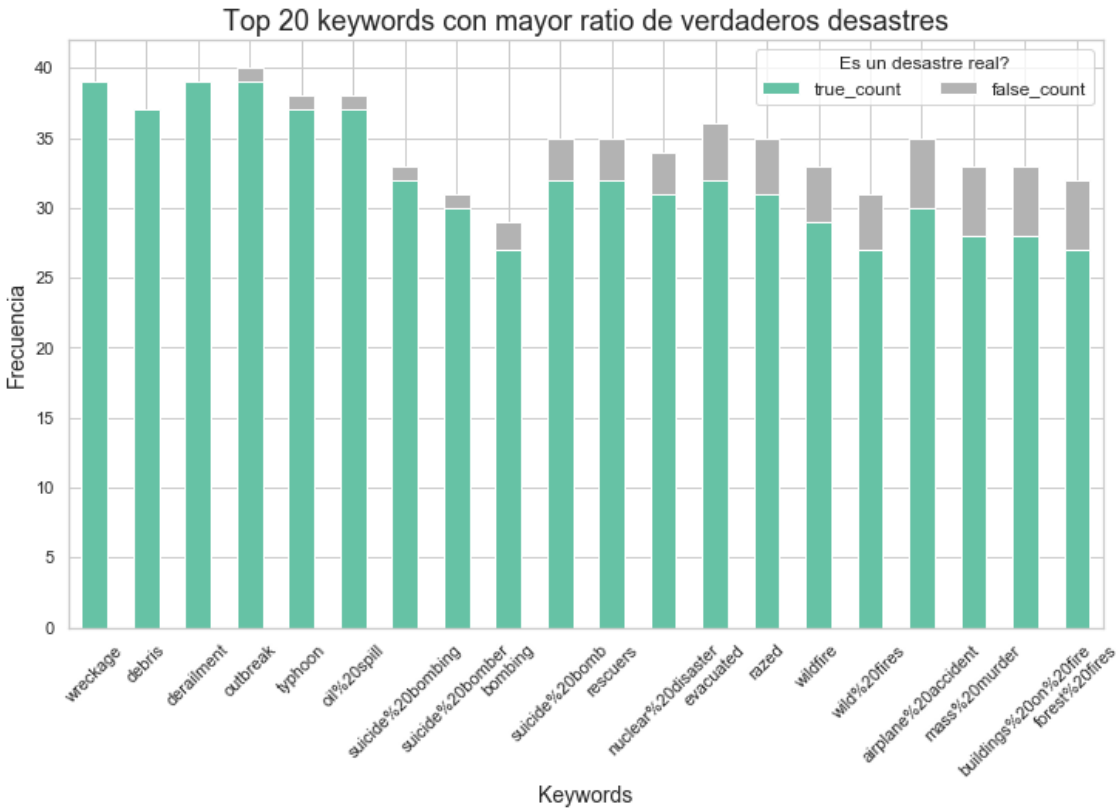
Keyword de los tweets

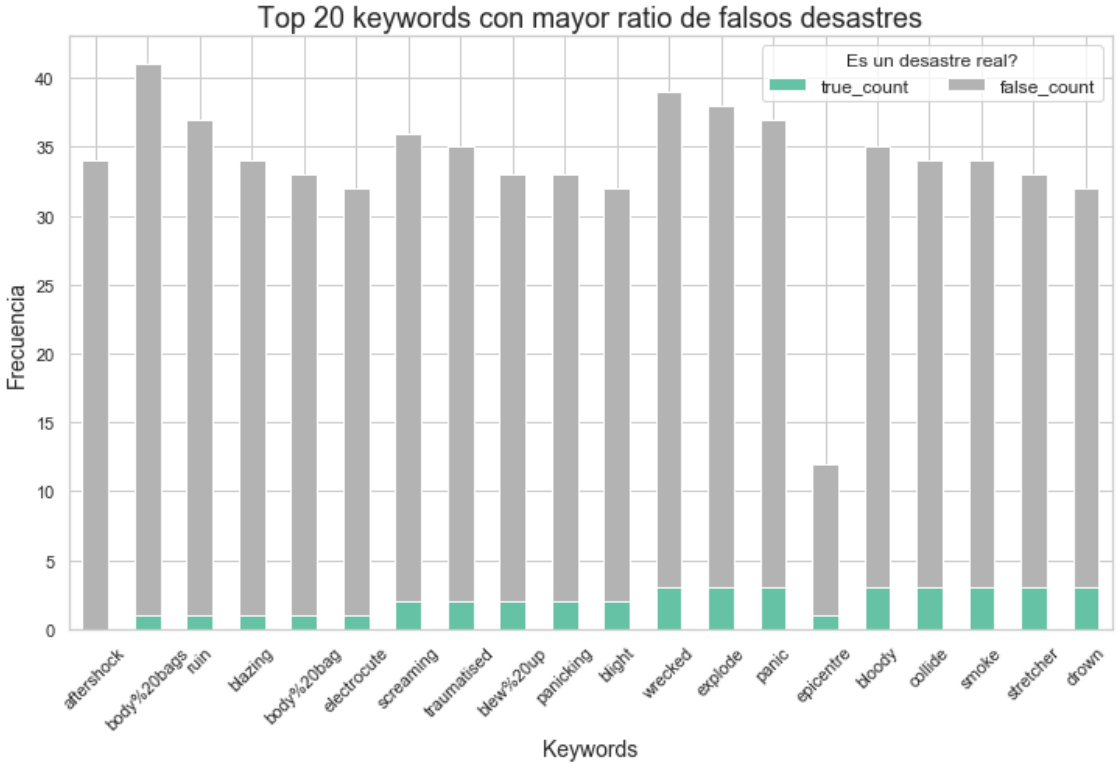
El 99% de los tweets tiene asignado un keyword. Desconocemos el método por el cuál fue asignado el mismo.



Como podemos ver en este gráfico, algunos casos tienen una tendencia más marcada que otros, pero no es uniforme, y sin realizar un análisis más profundo sobre las características de estas keywords, no podemos obtener un indicio claro.

Reordenemos el gráfico anterior para poder observar los keyword con mayor ratio de casos verdaderos y falsos:





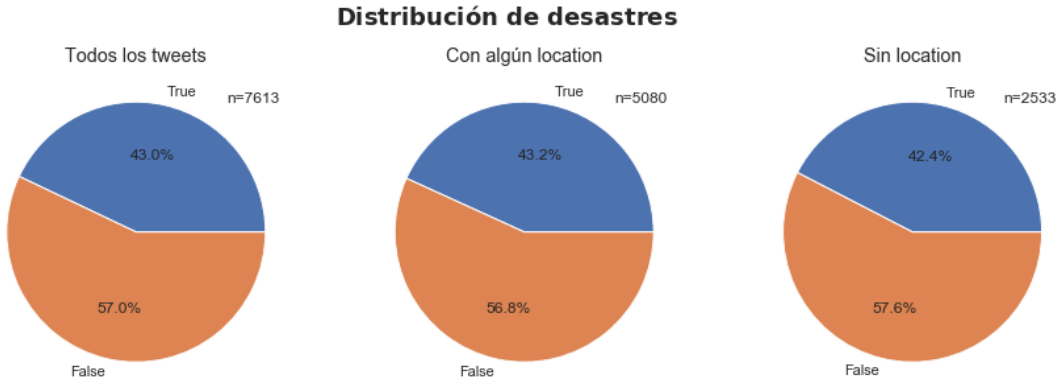
De esta manera podemos observar los casos más extremos, en donde vemos por ejemplo que keywords que implican desastres masivos y con un riesgo inmediato en el presente son los que mantienen un ratio de desastres verdaderos más alto, en contraposición con los que nos dan

un ratio de desastres falsos, que son menos específicos o hablan sobre las consecuencias de un desastre.

Podríamos usar este simple cálculo o realizar un análisis más profundo sobre estas palabras para ayudar con la predicción.

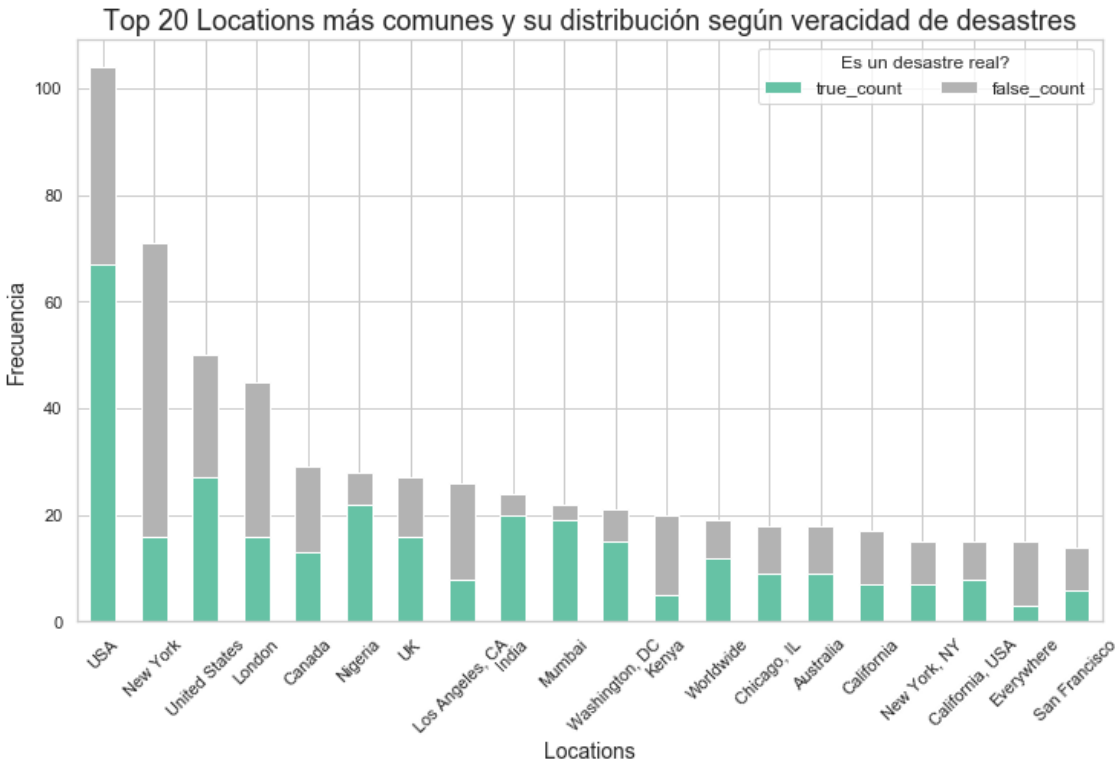
Location de los tweets

La proporción de tweets sin locación es de aproximadamente un tercio del universo total. El resto contienen algún valor en este campo.



En este gráfico podemos ver la relación entre que un tweet tenga o no location y la proporción de casos con desastres verdaderos/falsos. Observamos que no hay, al menos con esta medida simple, una diferencia significativa en la distribución.

En el gráfico siguiente, segmentaremos por los valores que puede tomar location y graficaremos el top 20.



Podemos ver que el set de datos tiene tweets que parecen provenir de distintas ubicaciones en el mundo, que algunos locations particulares como "India" o "Everywhere" tienen una tendencia más marcada que otros y que algunos casos como "New York" y "New York, NY" hacen referencia a un mismo lugar.

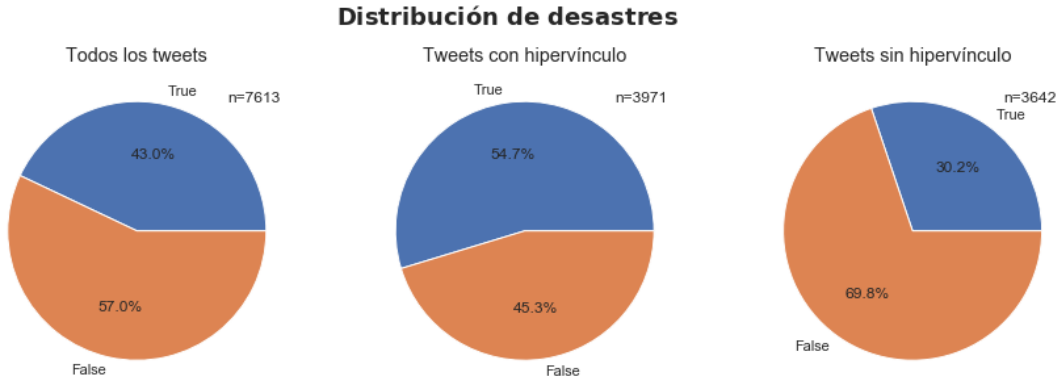
Encontraremos casos como "Everywhere" u otros que no corresponden a un lugar real. Sería interesante analizar si estos se correlacionan con más casos falsos.

No vamos a proceder con un análisis similar al que hicimos con los keywords, sobre aquellos con mayor/menor ratio, debido a que solo el 14% de los tweets pertenecen a keywords con 6 o más casos, reduciéndonos muchísimo la muestra.

Tweets que contienen links

Podemos ver que varios de los tweets – aproximadamente la mitad – contienen un hipervínculo (o al menos la cadena 'http').

La intuición nos diría que aquellos con un vínculo pueden hacer referencia a un artículo de noticias sobre el desastre o algún tema similar. Y esto debería impactar en el ratio de desastres verdaderos.



Se ve claramente en los datos que el hecho de contener un hipervínculo nos podría ayudar a entender si un tweet es sobre un desastre verdadero o falso.

Conclusiones

- Los tweets más largos tienden a ser verdaderos.
- Seguramente se puedan obtener observaciones interesantes de los campos de keyword, location y text si aplicamos algoritmos de procesamiento de lenguaje natural.
- El 99% de los tweets tiene una keyword.
- Hay keywords que claramente se correlacionan con desastres reales o falsos.
- La ausencia de locación no afecta el ratio de desastres en comparación a tener una location cualquiera.
- Sin embargo, distintas locations nos dan distintos ratios, que sí nos podrían ayudar a mejorar la predicción.
- Sería interesante poder distinguir locations reales de ficticias y observar cómo afectan a este ratio.
- Los tweets con hipervínculos tienden a ser sobre desastres verdaderos.

Análisis pendientes que podrían aportar valor

- Faltas de ortografía en tweets
- Mayor análisis de hipervínculos
- Location real o ficticia
- Sentiment análisis sobre los tweets
- Tipificación de keywords