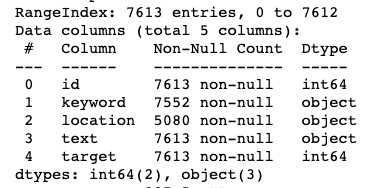
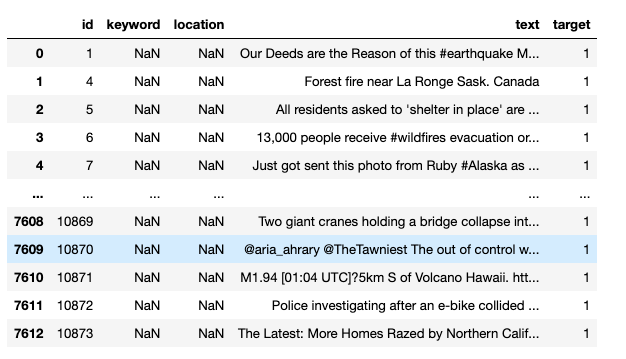
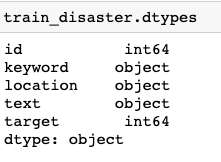
Análisis de train\_disaster

train\_disaster.info()



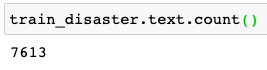


dtypes:



Size: 38065

Shape: (7613,5)



Text es el único feature importante y no hay valores en 0.

Ideas:

1) Proxies de 0 (falso)

cuanto mayor es el número de signos de exclamación, más tiende a ser falso.

Idem con el nro de mayúsculas (trata de atraer la atención)

modismo "ablaze" parece aparecer siempre en hoaxes.

mucho hasthag o arrobar a 3ero es intento de llamar la atención => hoax

2) Proxies de 1 (verdadero)

Uso de http para ofrecer fuente suena razonable.

hashtag o strings que correlacionan bien con desastre: crash, accident.

Hasthag o mención de lugares: buen indicador de true, correlacionar con base de sitios geográficos tal vez sirva.

3) Preguntas:

- Para generar buzz, son más largos los tweets de un falso desastre?

- Qué cadenas son las más comunes, más allá de las evidentes (disaster, crash, etc).

- Prescindir de las palabras más comunes de textos en Inglés (I, the, etc)

- Uso de http lo supongo alineado con True (intención de aportar verosimilitud)