FACULTAD
DE INGENIERIA
Universidad de Buenos Aires

Finger 1

[75.06] Organización de Datos
Primer cuatrimestre de 2020

Alumno:	MARTINEZ SASTRE, Gonzalo Gabriel
Número de padrón:	102321
Email:	gonzalomartinezsastre@gmail.com

[1]:

```python
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

[2]:

```python
tweets = pd.read_csv('../Finger/train.csv')
tweets.head()
```

[2]:

	id	keyword	location	text	target
0	1	NaN	NaN	Our Deeds are the Reason of this #earthquake M...	1
1	4	NaN	NaN	Forest fire near La Ronge Sask. Canada	1
2	5	NaN	NaN	All residents asked to 'shelter in place' are ...	1
3	6	NaN	NaN	13,000 people receive #wildfires evacuation or...	1
4	7	NaN	NaN	Just got sent this photo from Ruby #Alaska as ...	1

[3]:

```python
tweets.info()
```

```
<class 'pandas.core.frame.DataFrame'>
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 int64
dtypes: int64(2), object(3)
memory usage: 297.5+ KB
```

[4]:

```python
tweets_final = tweets.rename(columns={'target':'about_disaster'})
tweets_final['about_disaster'] = ((tweets_final['about_disaster'])==1)
tweets_final['length'] = (tweets_final['text']).str.len()
tweets_final.head(20)
```

[4]:

	id	keyword	location	text	about_disaster	length
0	1	NaN	NaN	Our Deeds are the Reason of this #earthquake M...	True	69
1	4	NaN	NaN	Forest fire near La Ronge Sask. Canada	True	38
2	5	NaN	NaN	All residents asked to 'shelter in place' are ...	True	133
3	6	NaN	NaN	13,000 people receive #wildfires evacuation or...	True	65
4	7	NaN	NaN	Just got sent this photo from Ruby #Alaska as ...	True	88
5	8	NaN	NaN	#RockyFire Update =>California Hwy. 20 closed...	True	110
6	10	NaN	NaN	#flood #disaster Heavy rain causes flash flood...	True	95
7	13	NaN	NaN	I'm on top of the hill and I can see a fire in...	True	59
8	14	NaN	NaN	There's an emergency evacuation happening now ...	True	79
9	15	NaN	NaN	I'm afraid that the tornado is coming to our a...	True	52
10	16	NaN	NaN	Three people died from the heat wave so far	True	43
11	17	NaN	NaN	Haha South Tampa is getting flooded hah- WAIT ...	True	129
12	18	NaN	NaN	#raining #flooding #Florida #TampaBay #Tampa 1...	True	76
13	19	NaN	NaN	#Flood in Bago Myanmar #We arrived Bago	True	39
14	20	NaN	NaN	Damage to school bus on 80 in multi car crash ...	True	56
15	23	NaN	NaN	What's up man?	False	14
16	24	NaN	NaN	I love fruits	False	13
17	25	NaN	NaN	Summer is lovely	False	16
18	26	NaN	NaN	My car is so fast	False	17
19	28	NaN	NaN	What a goooooooaaaaaal!!!!!!	False	28

[5]:

```
tweets_final.groupby('about_disaster').agg({'text':'count', 'length':['mean',␣
 ↪'max', 'min', 'sum']})
```

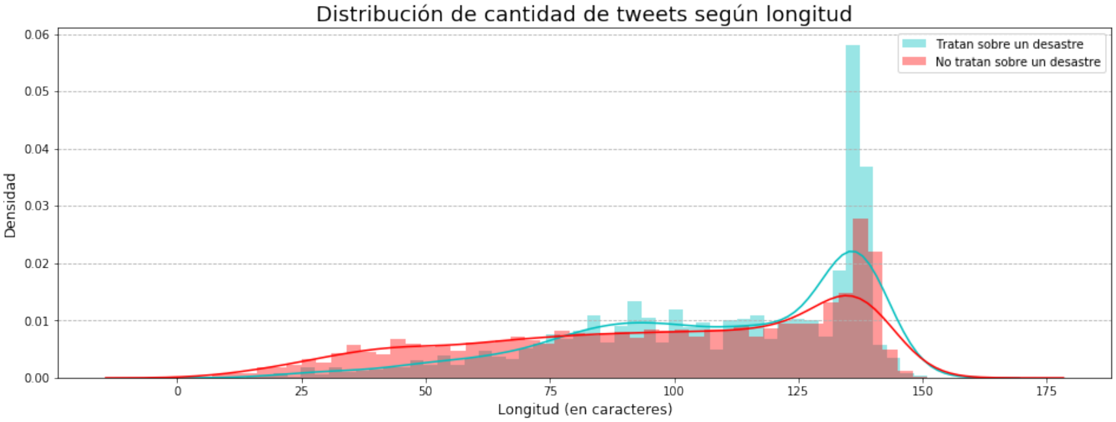
[5]:

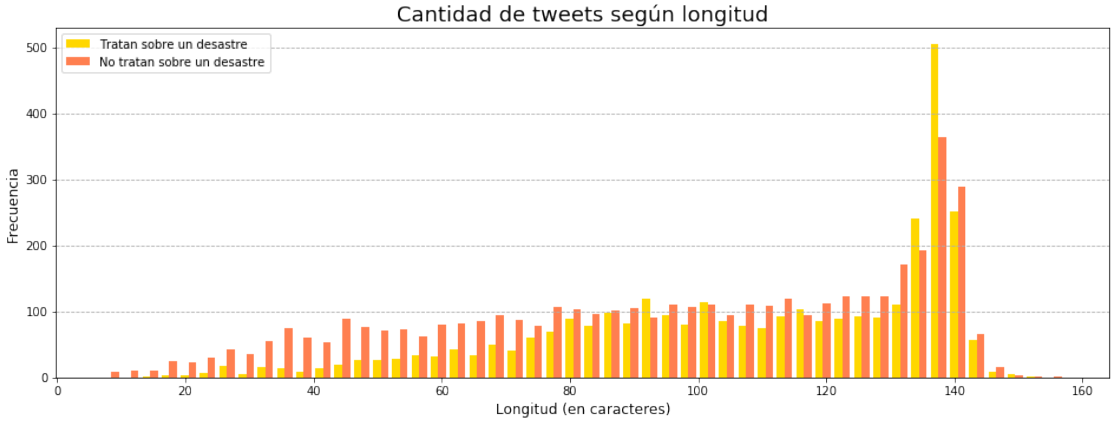
	text	length			
about_disaster	count	mean	max	min	sum
False	4342	95.706817	157	7	415559
True	3271	108.113421	151	14	353639

[6]:

```
tweets_clima = tweets_final[tweets_final['about_disaster']==True]['length']
tweets_no_clima = tweets_final[tweets_final['about_disaster']==False]['length']

# coloco 2 gráficos en una misma visualización
fig, (ax1, ax2) = plt.subplots(nrows=2)
fig.set_figheight(12)
fig.set_figwidth(16)

# density plot
densidad_tweets = sns.distplot(tweets_clima, color='c', \
                    label="Tratan sobre un desastre", bins=50,ax=ax1)

densidad_tweets = sns.distplot(tweets_no_clima, color='r', \
                    label="No tratan sobre un desastre", bins=50,ax=ax1)

densidad_tweets.set_title("Distribución de cantidad de tweets según longitud", \
                          fontsize=18)
densidad_tweets.set_ylabel("Densidad", fontsize=12)
densidad_tweets.set_xlabel("Longitud (en caracteres)", fontsize=12)
densidad_tweets.legend(prop={'size': 10})
densidad_tweets.grid(b=True, axis='y', linestyle='--')
```

```
# histogram
plt.hist([tweets_clima, tweets_no_clima], bins=50, color=['gold','coral'], \
         label=['Tratan sobre un desastre', "No tratan sobre un desastre"])

plt.title("Cantidad de tweets según longitud", fontsize=18)
plt.ylabel("Frecuencia", fontsize=12)
plt.xlabel("Longitud (en caracteres)", fontsize=12)
plt.legend(prop={'size': 10})
plt.grid(b=True, axis='y', linestyle='--')
```


Distribución de cantidad de tweets según longitud
Tratan sobre un desastre
No tratan sobre un desastre
Densidad
Longitud (en caracteres)


Cantidad de tweets según longitud
Tratan sobre un desastre
No tratan sobre un desastre
Frecuencia
Longitud (en caracteres)