finger1

May 2, 2020

```
[1]: import pandas as pd
     import seaborn as sns
     import matplotlib.pylab as plt
     import matplotlib.patches as mpatches
     import numpy as np
```

```
[2]: df = pd.read_csv("train.csv")
     df
```

```
[2]:          id keyword location  \
     0         1     NaN      NaN
     1         4     NaN      NaN
     2         5     NaN      NaN
     3         6     NaN      NaN
     4         7     NaN      NaN
     …       …       …       …
     7608  10869     NaN      NaN
     7609  10870     NaN      NaN
     7610  10871     NaN      NaN
     7611  10872     NaN      NaN
     7612  10873     NaN      NaN

                                                        text  target
     0     Our Deeds are the Reason of this #earthquake M…       1
     1                Forest fire near La Ronge Sask. Canada       1
     2     All residents asked to 'shelter in place' are …       1
     3     13,000 people receive #wildfires evacuation or…       1
     4     Just got sent this photo from Ruby #Alaska as …       1
     …                                                  …     …
     7608  Two giant cranes holding a bridge collapse int…       1
     7609  @aria_ahrary @TheTawniest The out of control w…       1
     7610  M1.94 [01:04 UTC]?5km S of Volcano Hawaii. htt…       1
     7611  Police investigating after an e-bike collided …       1
     7612  The Latest: More Homes Razed by Northern Calif…       1

     [7613 rows x 5 columns]
```

[3]:
```
#me quedo con las columnas que me interesan para el análisis
df = df.loc[:,['id', 'text', 'target']]
df
```

[3]:
```
         id                                               text  target
0         1  Our Deeds are the Reason of this #earthquake M…       1
1         4             Forest fire near La Ronge Sask. Canada       1
2         5  All residents asked to 'shelter in place' are …       1
3         6  13,000 people receive #wildfires evacuation or…       1
4         7  Just got sent this photo from Ruby #Alaska as …       1
…       …                                                …     …
7608  10869  Two giant cranes holding a bridge collapse int…       1
7609  10870  @aria_ahrary @TheTawniest The out of control w…       1
7610  10871  M1.94 [01:04 UTC]?5km S of Volcano Hawaii. htt…       1
7611  10872  Police investigating after an e-bike collided …       1
7612  10873  The Latest: More Homes Razed by Northern Calif…       1

[7613 rows x 3 columns]
```

[4]:
```
df['length'] = df.loc[:,'text'].str.len()
df
```

[4]:
```
         id                                               text  target  length
0         1  Our Deeds are the Reason of this #earthquake M…       1      69
1         4             Forest fire near La Ronge Sask. Canada       1      38
2         5  All residents asked to 'shelter in place' are …       1     133
3         6  13,000 people receive #wildfires evacuation or…       1      65
4         7  Just got sent this photo from Ruby #Alaska as …       1      88
…       …                                                …     …     …
7608  10869  Two giant cranes holding a bridge collapse int…       1      83
7609  10870  @aria_ahrary @TheTawniest The out of control w…       1     125
7610  10871  M1.94 [01:04 UTC]?5km S of Volcano Hawaii. htt…       1      65
7611  10872  Police investigating after an e-bike collided …       1     137
7612  10873  The Latest: More Homes Razed by Northern Calif…       1      94

[7613 rows x 4 columns]
```

[5]:
```
#chequeo que no falten datos
df.isnull().any()
```

[5]:
```
id        False
text      False
target    False
length    False
dtype: bool
```

[6]:
```
df['length'].describe()
```

```
[6]: count    7613.000000
     mean      101.037436
     std        33.781325
     min         7.000000
     25%        78.000000
     50%       107.000000
     75%       133.000000
     max       157.000000
     Name: length, dtype: float64
```

```
[7]: df['target'] = df['target'].astype('bool')
     df
```

```
[7]:          id                                               text  target  length
     0         1  Our Deeds are the Reason of this #earthquake M…    True      69
     1         4             Forest fire near La Ronge Sask. Canada    True      38
     2         5  All residents asked to 'shelter in place' are …    True     133
     3         6  13,000 people receive #wildfires evacuation or…    True      65
     4         7  Just got sent this photo from Ruby #Alaska as …    True      88
     …       …                                                 …       …       …
     7608  10869  Two giant cranes holding a bridge collapse int…    True      83
     7609  10870  @aria_ahrary @TheTawniest The out of control w…    True     125
     7610  10871  M1.94 [01:04 UTC]?5km S of Volcano Hawaii. htt…    True      65
     7611  10872  Police investigating after an e-bike collided …    True     137
     7612  10873  The Latest: More Homes Razed by Northern Calif…    True      94

     [7613 rows x 4 columns]
```

```
[8]: df[df.loc[:, 'target'] == True]['length'].describe()
```

```
[8]: count    3271.000000
     mean      108.113421
     std        29.309854
     min        14.000000
     25%        88.000000
     50%       115.000000
     75%       136.000000
     max       151.000000
     Name: length, dtype: float64
```

```
[9]: df[df.loc[:, 'target'] == False]['length'].describe()
```

```
[9]: count    4342.000000
     mean       95.706817
     std        35.885924
     min         7.000000
     25%        68.000000
```

```
50%        101.000000
75%        130.000000
max        157.000000
Name: length, dtype: float64
```

[10]:
```
df_true = df[df['target'] == True]
```

[11]:
```
df_false = df[df['target'] == False]
```

[12]:
```
print(df_true.length.mean())
print(df_false.length.mean())
```

```
108.11342097217977
95.70681713496084
```

[30]:
```
df['count'] = df.groupby('length').transform('count').text
df
```

[30]:
```
        id                                               text  target  length  \
1882  2703                                            Crushed   False       7
4890  6962                                            Bad day   False       7
5115  7295                                            Err:509   False       7
24      36                                           LOOOOOOL   False       8
30      44                                           The end!   False       8
…      …                                                  …       …       …
6945  9961  @helene_yancey GodsLove &amp; #thankU my siste…    True     148
2718  3904  @UN No more #GujaratRiot &amp; #MumbaiRiot92-9…    True     149
633    915  @HowardU If 90BLKs&amp;8WHTs colluded 2 take W…    True     150
635    919  @cspanwj If 90BLKs&amp;8WHTs colluded 2 take W…    True     150
614    885  @CAgov If 90BLKs&amp;8WHTs colluded 2 take WHT…    True     151

      count
1882      3
4890      3
5115      3
24        4
30        4
…       …
6945      5
2718      1
633       2
635       2
614       1

[7613 rows x 5 columns]
```

[28]:
```
df = df.sort_values(by=['target','length'])
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7613 entries, 1882 to 614
Data columns (total 5 columns):
id        7613 non-null int64
text      7613 non-null object
target    7613 non-null bool
length    7613 non-null int64
count     7613 non-null int64
dtypes: bool(1), int64(3), object(1)
memory usage: 304.8+ KB
```

[34]:

```
df_pyramid = df.loc[:,['count', 'target']]
df_pyramid['length_period'] = pd.cut(df['length'], np.arange(0,180,20),␣
 ↪right=True)
df_pyramid = df_pyramid.groupby(['length_period','target']).agg({'count':
 ↪'count'}).reset_index()
```

[35]:

```
df_pyramid.loc[df_pyramid['target'] == True,'count'] = -df_pyramid['count']
df_pyramid
```

[35]:

```
   length_period  target  count
0        (0, 20]   False     71
1        (0, 20]    True     -7
2       (20, 40]   False    323
3       (20, 40]    True    -73
4       (40, 60]   False    485
5       (40, 60]    True   -174
6       (60, 80]   False    613
7       (60, 80]    True   -354
8      (80, 100]   False    670
9      (80, 100]    True   -624
10    (100, 120]   False    712
11    (100, 120]    True   -591
12    (120, 140]   False   1359
13    (120, 140]    True  -1354
14    (140, 160]   False    109
15    (140, 160]    True    -94
```

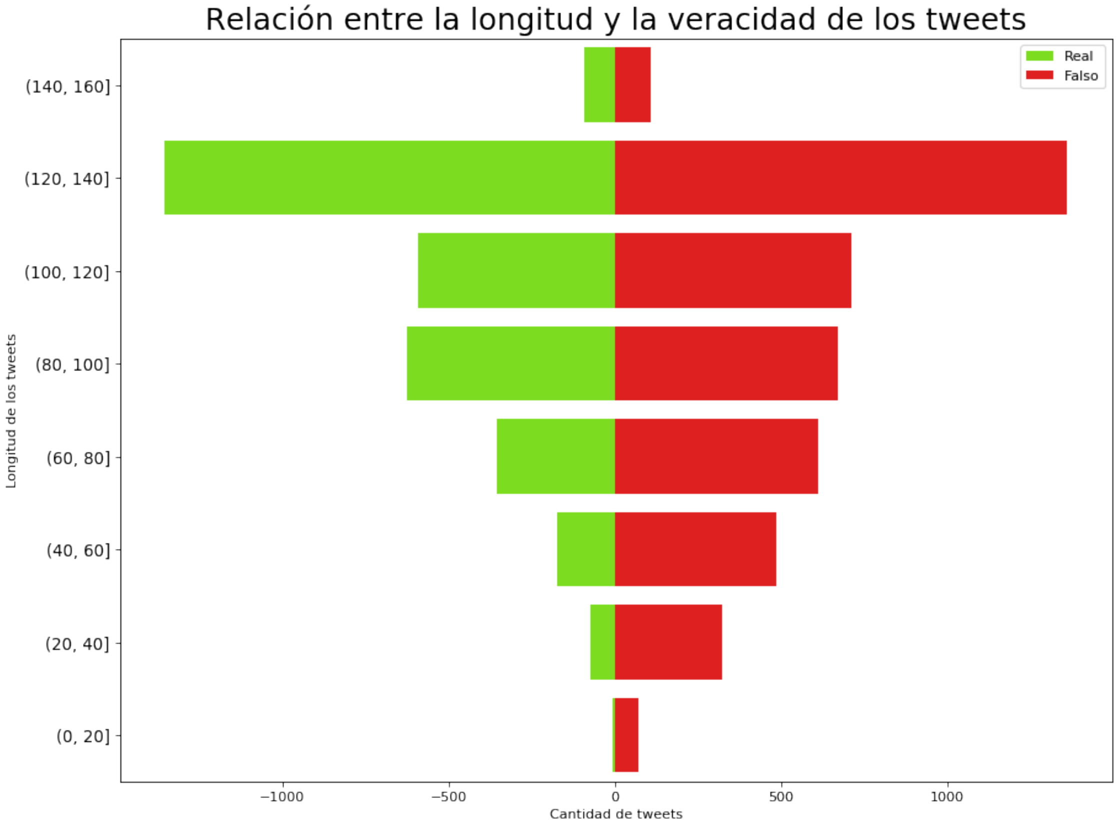
[17]:

```
plt.figure(figsize=(13,10), dpi= 80)
group_col = 'target'
order_of_bars = df_pyramid.length_period.unique()[::-1]

sns.barplot(x='count', y='length_period', data=df_pyramid.
 ↪loc[df_pyramid[group_col] == True, :], order=order_of_bars,␣
 ↪color="lawngreen", label='Real')
```

```
sns.barplot(x='count', y='length_period', data=df_pyramid.
 ↪loc[df_pyramid[group_col] == False, :], order=order_of_bars, color="red",␣
 ↪label='Falso')

plt.xlabel("Cantidad de tweets")
plt.ylabel("Longitud de los tweets")
plt.yticks(fontsize=12)
plt.title("Relación entre la longitud y la veracidad de los tweets",␣
 ↪fontsize=22)
plt.legend()
plt.show()
```



[ ]: