process_data

August 6, 2020

[7]:
```
import pandas as pd
import re
from nltk.corpus import stopwords
import ast
from datetime import datetime as dt
import numpy as np

df_trump=pd.read_csv("/home/davide/Scrivania/DAVIDE_PIU_WAAT_00106/00106_piu/
 ↪miosito/sentiment/data/tweets_donald_trump.csv",sep=",")#legge csv dei tweet␣
 ↪di trump e crea data frame
df_biden=pd.read_csv("/home/davide/Scrivania/DAVIDE_PIU_WAAT_00106/00106_piu/
 ↪miosito/sentiment/data/tweets_joe_biden.csv",sep=",")#legge csv dei tweet di␣
 ↪biden e crea data frame
df_trump=df_trump.dropna(thresh=2)#rimuove le righe con all'interno almeno 2 NA
df_biden=df_biden.dropna(thresh=2)#rimuove le righe con all'interno almeno 2 NA

def clean_tweet(tweet): #sostituisce con uno spazio vuoto tramite regular␣
 ↪exxpression i pattern all'interno delle parentesi
      return ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t]) |(\w+:\/\/
 ↪\S+)", " ", tweet).split())

def prepr(df):
    df=df[df['text']!="text"]#aggiungendo i dati al csv venivano aggiunti␣
 ↪nuovamente i nomi delle colonne, questa riga di codice mi permette di␣
 ↪rimuoverle
    df['text'] = df['text'].apply(clean_tweet)#applico funzione per la pulizia␣
 ↪dei tweet
    df.drop(df.columns[7:], axis=1, inplace=True)#cancello colonne inutili.␣
 ↪problema causato dal fatto che inizialmente stavo salvando i tweet in modo␣
 ↪diverso
    df=df.loc[2:]
    df=df.dropna(thresh=3)#rimuove le righe con all'interno almeno 3 NA
    df=df[df['sentiment'].str.startswith('{')]#scelgo per la colonna sentiment␣
 ↪solo le righe che iniziano con "{" perché tweepy restituisce un dizionario
    df['sentiment']=df['sentiment'].apply(ast.literal_eval)#trasforma una␣
 ↪stringa contenente un dizionario, in un dizionario
```

```
    df=pd.concat([df.drop(['sentiment'], axis=1), df['sentiment'].apply(pd.
↪Series)], axis=1)#divide il dizionario in delle colonne che hanno come␣
↪etichetta la chiave del dizionario e all'interno dei campi della colonna␣
↪vengono inseriti i valori del dizionario
    return df




df_trump=prepr(df_trump)#applico funzione appena creata
df_biden=prepr(df_biden)#applico funzione appena creata
```

```
<ipython-input-7-6378671cf59b>:18: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df['text'] = df['text'].apply(clean_tweet)#applico funzione per la pulizia dei
tweet
```

[8]:
```
"""aggiungo la colonna candidato, mi servirà successivamente quando creerò un␣
↪unico dataframe dei tweet dei due candidati
"""
df_biden['candidato'] = 'Joe Biden'
df_trump['candidato'] = 'Donald Trump'
```

[9]:
```
a=pd.DataFrame(df_biden.groupby('sign').size()/df_biden['sign'].
↪count()*100)#percentuali tweet negativi, positivi, neutrali biden
```

[10]:
```
a['trump']=df_trump.groupby('sign').size()/df_trump['sign'].count()*100 #␣
↪aggiunta colonna con percentuali tweet negativi, positivi, neutrali  trump
```

[11]:
```
a.columns=['biden','trump']
a.to_csv('data_for_pie.csv')#salvo i dati in un csv che utilizzerò per i grafici
```

[12]:
```
df=pd.concat([df_trump, df_biden], ignore_index=True)#concateno i dataframe dei␣
↪tweet di biden e di trump
```

[13]:
```
df['new_date_column'] = pd.to_datetime(df['date tweet'],errors='coerce').dt.
↪date#converto la data, non ho bisogno dell'orario
```

[14]:
```
g = df.groupby(["candidato", "new_date_column"])#raggruppo per data e candidato
```

[15]:
```
daily_averages = g.aggregate({"polarity":np.mean})#calcolo la media giornaliera␣
↪di polarity giornaliera per candidato
```

[16]:
```
daily_averages=pd.DataFrame(daily_averages)#converto in un dataframe
daily_averages.to_csv('media_polarity.csv', index = True)
```

[17]:
```
df=pd.read_csv("media_polarity.csv")
```

[18]:
```
"""rimodello il dataset creato, l'indice è la data, le colonne sono i nomi dei␣
 ↪candidati e all'interno
delle celle troviamo la media giornaliera di polarity"""
df=df.pivot(index='new_date_column', columns='candidato', values='polarity')
```

[19]:
```
df = df.iloc[1:]#rimosso prima riga perchè appare NA nei valori di biden
df.to_csv('media_polarity.csv', index = True)
df=pd.read_csv("media_polarity.csv")
```

[20]:
```
df#visualizzo media polarity
```

[20]:

	new_date_column	Donald Trump	Joe Biden
0	2020-07-12	0.035279	0.028989
1	2020-07-13	0.005570	0.121669
2	2020-07-14	0.019764	0.076297
3	2020-07-15	0.020163	0.030099
4	2020-07-18	0.027876	0.082130
5	2020-07-19	0.092829	0.082654
6	2020-07-20	-0.028699	0.144445
7	2020-07-21	0.045359	0.072526
8	2020-07-22	0.054454	-0.019329
9	2020-07-23	0.049804	0.073473
10	2020-07-24	0.067248	0.102774
11	2020-07-25	0.064810	0.045631
12	2020-07-27	0.000232	0.036802
13	2020-07-28	0.024645	0.053352
14	2020-07-30	0.001956	-0.007719
15	2020-07-31	0.064027	0.052404
16	2020-08-01	0.013873	0.015232
17	2020-08-03	0.000772	0.106399
18	2020-08-04	0.067548	0.020286
19	2020-08-05	0.003899	0.131294
20	2020-08-06	-0.026762	0.009479

[21]:
```
a
```

[21]:

sign	biden	trump
negative	22.375935	23.013841
neutral	38.664174	39.815268
positive	38.959891	37.170892

[28]: