Esempio Modulo text-PreProcessing

May 4, 2022

1 Esempi di Utilizzo del MODULO TEXT-PREPROCESSING della libreria IntelligenzaArtificiale

1.1 Installare la libreria

[ ]:
```
#Per installare la libreria sul tuo computer puoi usare il comando :
pip3 install intelligenzaartificiale

#se invece utilizzi google colab puoi usare:
!pip install intelligenzaartificiale
```

1.2 Importare la libreria

[2]:
```
from intelligenzaartificiale import dataset as dt
from intelligenzaartificiale import textpreprocessing as tpp
```

[3]:
```
# importiamo un Set di dati
il_mio_dataset = dt.leggi_csv("twitt.csv")

print(il_mio_dataset)
```

```
Tempo impiegato per leggere il file:  0.33322882652282715
                    tweet_id airline_sentiment             name  \
0      5,70306133677761E+017           neutral          cairdin
1      5,70301130888122E+017          positive         jnardino
2      5,70301083672814E+017           neutral       yvonnalynn
3      5,70301031407624E+017          negative         jnardino
4      5,70300817074463E+017          negative         jnardino
…                        …                 …                …
14635  5,69587686496825E+017          positive  KristenReenders
14636  5,69587371693355E+017          negative         itsropes
14637  5,69587242672398E+017           neutral         sanyabun
14638  5,69587188687634E+017          negative       SraJackson
14639  5,69587140490867E+017           neutral        daviddtwu

       retweet_count  \
0                  0
1                  0
```

```
2                   0
3                   0
4                   0
…                 …
14635               0
14636               0
14637               0
14638               0
14639               0

             text  \
0                                                          @VirginAmerica What
@dhepburn said.
1                        @VirginAmerica plus you've added commercials to the
experience… tacky.
2                         @VirginAmerica I didn't today… Must mean I need to
take another trip!
3      @VirginAmerica it's really aggressive to blast obnoxious "entertainment"
in your guest…
4                                        @VirginAmerica and it's a really big
bad thing about it
…
…
14635                             @AmericanAir thank you we got on a different
flight to Chicago.
14636  @AmericanAir leaving over 20 minutes Late Flight. No warnings or
communication until w…
14637                                @AmericanAir Please bring American Airlines
to #BlackBerry10
14638  @AmericanAir you have my money, you change my flight, and don't answer
your phones! An…
14639  @AmericanAir we have 8 ppl so we need 2 know how many seats are on the
next flight. Pl…

                    tweet_created
0      2015-02-24 11:35:52 -0800
1      2015-02-24 11:15:59 -0800
2      2015-02-24 11:15:48 -0800
3      2015-02-24 11:15:36 -0800
4      2015-02-24 11:14:45 -0800
…                            …
14635  2015-02-22 12:01:01 -0800
14636  2015-02-22 11:59:46 -0800
14637  2015-02-22 11:59:15 -0800
14638  2015-02-22 11:59:02 -0800
14639  2015-02-22 11:58:51 -0800

[14640 rows x 6 columns]
```

1.3 Fare Pulizia di Base

[4]:
```
#pulire l'intera colonna con una riga
il_mio_dataset["textPulito"] = tpp.pulisci_testo(il_mio_dataset,"text")

#trasforma in minuscolo il tetso
il_mio_dataset["nameMinuscolo"] = tpp.trasforma_in_minuscolo(il_mio_dataset,␣
 ↪"name")

#rimuovi caratteri speciali e cifre !"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~0123456789
il_mio_dataset["textPulito"] = tpp.
 ↪rimuovi_caratteri_speciali_e_cifre(il_mio_dataset,"textPulito")

#rimuovi caratteri speciali !"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~
il_mio_dataset["textPulito"] = tpp.
 ↪rimuovi_caratteri_speciali(il_mio_dataset,"textPulito")

#rimuovi stopwords
il_mio_dataset["CleanText"] = tpp.
 ↪rimuovi_stopwords(il_mio_dataset,"textPulito", "english")

print(il_mio_dataset)
```

```
                    tweet_id airline_sentiment             name  \
0      5,70306133677761E+017           neutral          cairdin
1      5,70301130888122E+017          positive         jnardino
2      5,70301083672814E+017           neutral       yvonnalynn
3      5,70301031407624E+017          negative         jnardino
4      5,70300817074463E+017          negative         jnardino
…                        …                 …              …
14635  5,69587686496825E+017          positive  KristenReenders
14636  5,69587371693355E+017          negative         itsropes
14637  5,69587242672398E+017           neutral         sanyabun
14638  5,69587188687634E+017          negative       SraJackson
14639  5,69587140490867E+017           neutral        daviddtwu

       retweet_count  \
0                  0
1                  0
2                  0
3                  0
4                  0
…                …
14635              0
14636              0
14637              0
14638              0
14639              0
```

```
            text  \
0                                                    @VirginAmerica What
@dhepburn said.
1                        @VirginAmerica plus you've added commercials to the
experience… tacky.
2                         @VirginAmerica I didn't today… Must mean I need to
take another trip!
3      @VirginAmerica it's really aggressive to blast obnoxious "entertainment"
in your guest…
4                                          @VirginAmerica and it's a really big
bad thing about it
…
…
14635                             @AmericanAir thank you we got on a different
flight to Chicago.
14636  @AmericanAir leaving over 20 minutes Late Flight. No warnings or
communication until w…
14637                                @AmericanAir Please bring American Airlines
to #BlackBerry10
14638  @AmericanAir you have my money, you change my flight, and don't answer
your phones! An…
14639  @AmericanAir we have 8 ppl so we need 2 know how many seats are on the
next flight. Pl…

                   tweet_created  \
0      2015-02-24 11:35:52 -0800
1      2015-02-24 11:15:59 -0800
2      2015-02-24 11:15:48 -0800
3      2015-02-24 11:15:36 -0800
4      2015-02-24 11:14:45 -0800
…                            …
14635  2015-02-22 12:01:01 -0800
14636  2015-02-22 11:59:46 -0800
14637  2015-02-22 11:59:15 -0800
14638  2015-02-22 11:59:02 -0800
14639  2015-02-22 11:58:51 -0800

      textPulito  \
0
virginamerica dhepburn said
1                                          virginamerica plus added commercials
experience tacky
2                                           virginamerica today must mean need
take another trip
3      virginamerica really aggressive blast obnoxious entertainment guests
faces amp little …
4                                                              virginamerica
```

```
really big bad thing
…
…
14635                                                americanair thank got
different flight chicago
14636  americanair leaving minutes late flight warnings communication minutes
late flight cal…
14637                                        americanair please bring american
airlines blackberry
14638                  americanair money change flight answer phones suggestions
make commitment
14639     americanair ppl need know many seats next flight plz put us standby
people next flight

         nameMinuscolo  \
0               cairdin
1             jnardino
2           yvonnalynn
3             jnardino
4             jnardino
…                  …
14635  kristenreenders
14636         itsropes
14637         sanyabun
14638       srajackson
14639        daviddtwu

       CleanText
0
virginamerica dhepburn said
1                                          virginamerica plus added commercials
experience tacky
2                                           virginamerica today must mean need
take another trip
3      virginamerica really aggressive blast obnoxious entertainment guests
faces amp little …
4                                                              virginamerica
really big bad thing
…
…
14635                                                americanair thank got
different flight chicago
14636  americanair leaving minutes late flight warnings communication minutes
late flight cal…
14637                                        americanair please bring american
airlines blackberry
14638                  americanair money change flight answer phones suggestions
make commitment
```

```
14639     americanair ppl need know many seats next flight plz put us standby
people next flight

[14640 rows x 9 columns]
```

1.4 Tokenizzazione e vettorizzazione del testo

[5]:
```
#vettorizzare il testo (tfidf)
il_mio_dataset["testoVet"] = tpp.vettorizza_testo(il_mio_dataset,"textPulito")

#tokenizzare il testo
il_mio_dataset["testoTok"] = tpp.tokenizza_testo(il_mio_dataset,"textPulito")


print(il_mio_dataset)
```

```
                   tweet_id airline_sentiment             name  \
0      5,70306133677761E+017           neutral          cairdin
1      5,70301130888122E+017          positive         jnardino
2      5,70301083672814E+017           neutral       yvonnalynn
3      5,70301031407624E+017          negative         jnardino
4      5,70300817074463E+017          negative         jnardino
…                        …                 …              …
14635  5,69587686496825E+017          positive  KristenReenders
14636  5,69587371693355E+017          negative         itsropes
14637  5,69587242672398E+017           neutral         sanyabun
14638  5,69587188687634E+017          negative       SraJackson
14639  5,69587140490867E+017           neutral        daviddtwu

       retweet_count  \
0                  0
1                  0
2                  0
3                  0
4                  0
…                …
14635              0
14636              0
14637              0
14638              0
14639              0

            text  \
0                                                  @VirginAmerica What
@dhepburn said.
1                       @VirginAmerica plus you've added commercials to the
experience… tacky.
2                        @VirginAmerica I didn't today… Must mean I need to
```

```
take another trip!
3      @VirginAmerica it's really aggressive to blast obnoxious "entertainment"
in your guest…
4                                          @VirginAmerica and it's a really big
bad thing about it
…
…
14635                             @AmericanAir thank you we got on a different
flight to Chicago.
14636  @AmericanAir leaving over 20 minutes Late Flight. No warnings or
communication until w…
14637                                 @AmericanAir Please bring American Airlines
to #BlackBerry10
14638  @AmericanAir you have my money, you change my flight, and don't answer
your phones! An…
14639  @AmericanAir we have 8 ppl so we need 2 know how many seats are on the
next flight. Pl…

                   tweet_created  \
0      2015-02-24 11:35:52 -0800
1      2015-02-24 11:15:59 -0800
2      2015-02-24 11:15:48 -0800
3      2015-02-24 11:15:36 -0800
4      2015-02-24 11:14:45 -0800
…                            …
14635  2015-02-22 12:01:01 -0800
14636  2015-02-22 11:59:46 -0800
14637  2015-02-22 11:59:15 -0800
14638  2015-02-22 11:59:02 -0800
14639  2015-02-22 11:58:51 -0800

      textPulito  \
0
virginamerica dhepburn said
1                                           virginamerica plus added commercials
experience tacky
2                                            virginamerica today must mean need
take another trip
3      virginamerica really aggressive blast obnoxious entertainment guests
faces amp little …
4                                                                 virginamerica
really big bad thing
…
…
14635                                                 americanair thank got
different flight chicago
14636  americanair leaving minutes late flight warnings communication minutes
late flight cal…
```

```
14637                                        americanair please bring american
airlines blackberry
14638                  americanair money change flight answer phones suggestions
make commitment
14639     americanair ppl need know many seats next flight plz put us standby
people next flight

         nameMinuscolo  \
0               cairdin
1              jnardino
2            yvonnalynn
3              jnardino
4              jnardino
…                 …
14635  kristenreenders
14636          itsropes
14637          sanyabun
14638       srajackson
14639        daviddtwu

       CleanText  \
0
virginamerica dhepburn said
1                                           virginamerica plus added commercials
experience tacky
2                                            virginamerica today must mean need
take another trip
3      virginamerica really aggressive blast obnoxious entertainment guests
faces amp little …
4                                                                 virginamerica
really big bad thing
…
…
14635                                              americanair thank got
different flight chicago
14636  americanair leaving minutes late flight warnings communication minutes
late flight cal…
14637                                       americanair please bring american
airlines blackberry
14638                  americanair money change flight answer phones suggestions
make commitment
14639     americanair ppl need know many seats next flight plz put us standby
people next flight

        testoVet  \
0      [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, …
1      [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
```

```
0.0, 0.0, 0.0, …
2      [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, …
3      [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, …
4      [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, …
…
…
14635  [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, …
14636  [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, …
14637  [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, …
14638  [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, …
14639  [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, …

        testoTok
0                                                              [virginamerica,
dhepburn, said]
1                                    [virginamerica, plus, added, commercials,
experience, tacky]
2                                   [virginamerica, today, must, mean, need,
take, another, trip]
3      [virginamerica, really, aggressive, blast, obnoxious, entertainment,
guests, faces, am…
4                                                        [virginamerica, really,
big, bad, thing]
…
…
14635                                        [americanair, thank, got, different,
flight, chicago]
14636  [americanair, leaving, minutes, late, flight, warnings, communication,
minutes, late, …
14637                                    [americanair, please, bring, american,
airlines, blackberry]
14638        [americanair, money, change, flight, answer, phones, suggestions,
make, commitment]
14639  [americanair, ppl, need, know, many, seats, next, flight, plz, put, us,
standby, peopl…

[14640 rows x 11 columns]
```

[8]:
```
#analisi componenti principali
il_mio_dataset["pca"] = tpp.componenti_principali(il_mio_dataset,"testoVet")
il_mio_dataset
```

[8]:

```
                    tweet_id airline_sentiment             name  \
0      5,70306133677761E+017           neutral          cairdin
1      5,70301130888122E+017          positive         jnardino
2      5,70301083672814E+017           neutral       yvonnalynn
3      5,70301031407624E+017          negative         jnardino
4      5,70300817074463E+017          negative         jnardino
…                        …                 …                …
14635  5,69587686496825E+017          positive  KristenReenders
14636  5,69587371693355E+017          negative         itsropes
14637  5,69587242672398E+017           neutral         sanyabun
14638  5,69587188687634E+017          negative       SraJackson
14639  5,69587140490867E+017           neutral        daviddtwu

       retweet_count  \
0                  0
1                  0
2                  0
3                  0
4                  0
…                …
14635              0
14636              0
14637              0
14638              0
14639              0

            text  \
0                                                       @VirginAmerica What
@dhepburn said.
1                        @VirginAmerica plus you've added commercials to the
experience… tacky.
2                         @VirginAmerica I didn't today… Must mean I need to
take another trip!
3      @VirginAmerica it's really aggressive to blast obnoxious "entertainment"
in your guest…
4                                          @VirginAmerica and it's a really big
bad thing about it
…
…
14635                             @AmericanAir thank you we got on a different
flight to Chicago.
14636  @AmericanAir leaving over 20 minutes Late Flight. No warnings or
communication until w…
```

```
14637                                @AmericanAir Please bring American Airlines
to #BlackBerry10
14638  @AmericanAir you have my money, you change my flight, and don't answer
your phones! An…
14639  @AmericanAir we have 8 ppl so we need 2 know how many seats are on the
next flight. Pl…

                   tweet_created  \
0      2015-02-24 11:35:52 -0800
1      2015-02-24 11:15:59 -0800
2      2015-02-24 11:15:48 -0800
3      2015-02-24 11:15:36 -0800
4      2015-02-24 11:14:45 -0800
…                           …
14635  2015-02-22 12:01:01 -0800
14636  2015-02-22 11:59:46 -0800
14637  2015-02-22 11:59:15 -0800
14638  2015-02-22 11:59:02 -0800
14639  2015-02-22 11:58:51 -0800

      textPulito  \
0
virginamerica dhepburn said
1                                          virginamerica plus added commercials
experience tacky
2                                           virginamerica today must mean need
take another trip
3      virginamerica really aggressive blast obnoxious entertainment guests
faces amp little …
4                                                              virginamerica
really big bad thing
…
…
14635                                              americanair thank got
different flight chicago
14636  americanair leaving minutes late flight warnings communication minutes
late flight cal…
14637                                          americanair please bring american
airlines blackberry
14638                  americanair money change flight answer phones suggestions
make commitment
14639     americanair ppl need know many seats next flight plz put us standby
people next flight

         nameMinuscolo  \
0               cairdin
1              jnardino
```

```
2           yvonnalynn
3             jnardino
4             jnardino
…                …
14635  kristenreenders
14636         itsropes
14637         sanyabun
14638       srajackson
14639        daviddtwu

       CleanText  \
0
virginamerica dhepburn said
1                                          virginamerica plus added commercials
experience tacky
2                                           virginamerica today must mean need
take another trip
3      virginamerica really aggressive blast obnoxious entertainment guests
faces amp little …
4                                                                 virginamerica
really big bad thing
…
…
14635                                             americanair thank got
different flight chicago
14636  americanair leaving minutes late flight warnings communication minutes
late flight cal…
14637                                          americanair please bring american
airlines blackberry
14638                  americanair money change flight answer phones suggestions
make commitment
14639     americanair ppl need know many seats next flight plz put us standby
people next flight

        testoVet  \
0      [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, …
1      [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, …
2      [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, …
3      [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, …
4      [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, …
…
…
```

```
14635  [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, …
14636  [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, …
14637  [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, …
14638  [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, …
14639  [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, …

        testoTok  \
0                                                                  [virginamerica,
dhepburn, said]
1                                    [virginamerica, plus, added, commercials,
experience, tacky]
2                                   [virginamerica, today, must, mean, need,
take, another, trip]
3      [virginamerica, really, aggressive, blast, obnoxious, entertainment,
guests, faces, am…
4                                                    [virginamerica, really,
big, bad, thing]
…
…
14635                                      [americanair, thank, got, different,
flight, chicago]
14636  [americanair, leaving, minutes, late, flight, warnings, communication,
minutes, late, …
14637                                    [americanair, please, bring, american,
airlines, blackberry]
14638        [americanair, money, change, flight, answer, phones, suggestions,
make, commitment]
14639  [americanair, ppl, need, know, many, seats, next, flight, plz, put, us,
standby, peopl…

                                                  pca
0      [0.014566264823679266, -0.025938871928369953]
1      [0.016928967912752686, -0.021069497648900988]
2      [-0.02313053313685442, -0.015965084456002206]
3      [0.009051688797034052, -0.022324893356076975]
4       [0.017000910511536654, -0.02830333153448086]
…                                                 …
14635      [0.05000793745431668, 0.3021836620421646]
14636    [-0.13907160823308673, 0.01701213136826489]
14637  [-0.018922608800627034, 0.004674295237504221]
14638   [-0.08594018581317102, 0.007906086738102844]
14639   [-0.13291477966503168, 0.008777565903708585]
```

```
[14640 rows x 12 columns]
```

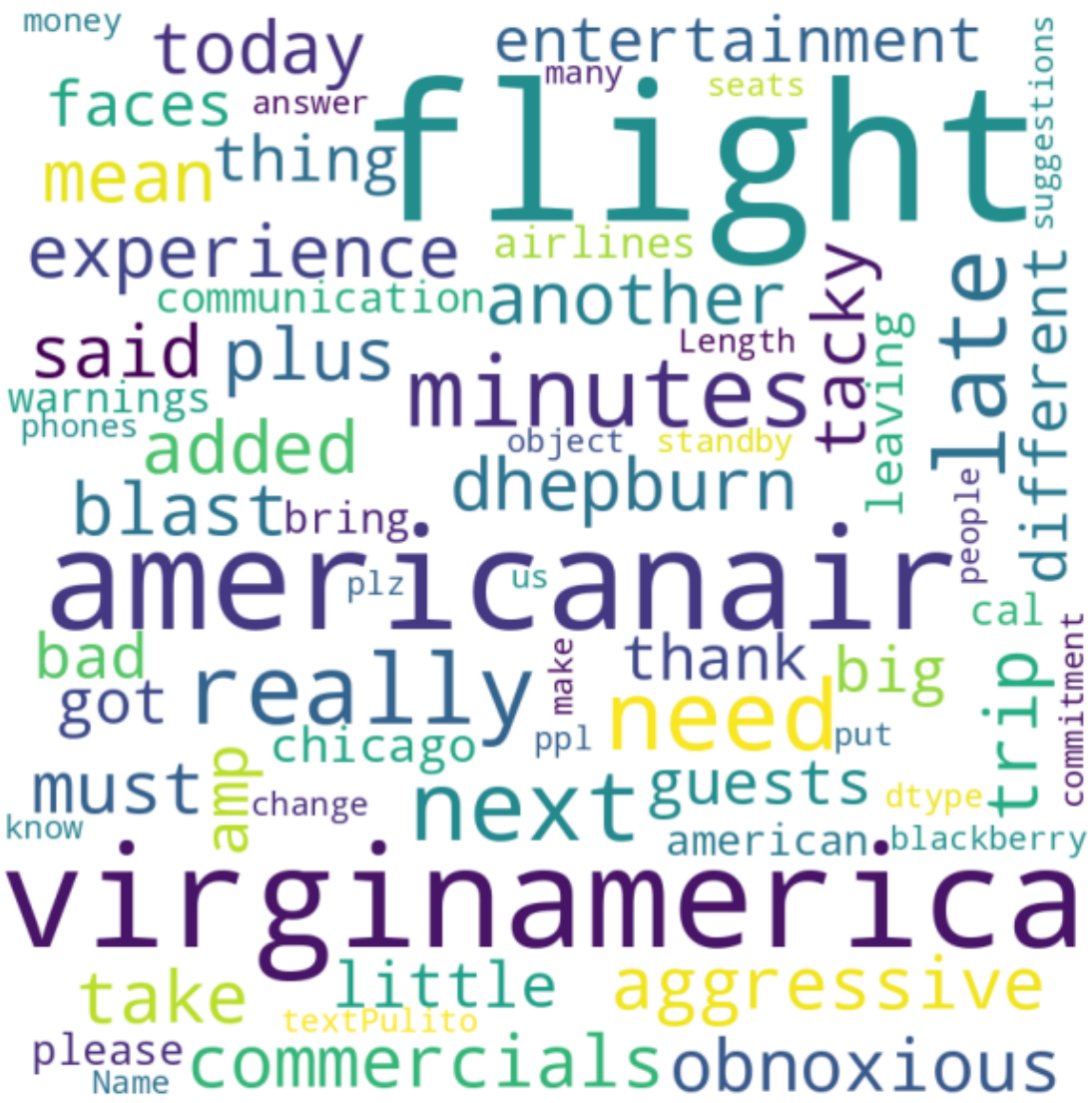
1.5 Altre Funzioni

[9]:
```
il_mio_dataset["wordbags"] = tpp.bag_of_words(il_mio_dataset,"textPulito")

print(il_mio_dataset["wordbags"])

#genera grafico words cloud
tpp.crea_wordcloud(il_mio_dataset,"textPulito")
```

```
0          (0, 2944)\t1\n  (0, 9808)\t1\n  (0, 12263)\t1\n  (1, 140)\t1\n  (1,
2165)\t1\n  (1, …
1          (0, 2944)\t1\n  (0, 9808)\t1\n  (0, 12263)\t1\n  (1, 140)\t1\n  (1,
2165)\t1\n  (1, …
2          (0, 2944)\t1\n  (0, 9808)\t1\n  (0, 12263)\t1\n  (1, 140)\t1\n  (1,
2165)\t1\n  (1, …
3          (0, 2944)\t1\n  (0, 9808)\t1\n  (0, 12263)\t1\n  (1, 140)\t1\n  (1,
2165)\t1\n  (1, …
4          (0, 2944)\t1\n  (0, 9808)\t1\n  (0, 12263)\t1\n  (1, 140)\t1\n  (1,
2165)\t1\n  (1, …
                                                    …
14635      (0, 2944)\t1\n  (0, 9808)\t1\n  (0, 12263)\t1\n  (1, 140)\t1\n  (1,
2165)\t1\n  (1, …
14636      (0, 2944)\t1\n  (0, 9808)\t1\n  (0, 12263)\t1\n  (1, 140)\t1\n  (1,
2165)\t1\n  (1, …
14637      (0, 2944)\t1\n  (0, 9808)\t1\n  (0, 12263)\t1\n  (1, 140)\t1\n  (1,
2165)\t1\n  (1, …
14638      (0, 2944)\t1\n  (0, 9808)\t1\n  (0, 12263)\t1\n  (1, 140)\t1\n  (1,
2165)\t1\n  (1, …
14639      (0, 2944)\t1\n  (0, 9808)\t1\n  (0, 12263)\t1\n  (1, 140)\t1\n  (1,
2165)\t1\n  (1, …
Name: wordbags, Length: 14640, dtype: object
```



1.6 Altre risorse

- **Documentazione Ufficiale**

- **Blog Ufficiale**

- **Corsi Gratis**

- **Ebook Gratis**

- **Progetti Python Open Source**

- **Dataset Pubblici**

- **Editor Python Online per il M.L.**

2 Per favore citaci se usi la Libreria.