In [1]:

```
import glob
import pandas as pd
import os
import matplotlib.pyplot as plt
import seaborn as sns
from os import path
import collections
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.cm as cm
import seaborn as sns
from scipy.stats import norm
from sklearn.ensemble import RandomForestClassifier
from sklearn.preprocessing import StandardScaler
from sklearn import metrics
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report, confusion_matrix
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
import string
import re
import nltk
from nltk import pos_tag
from nltk.corpus import stopwords
from nltk.tokenize import WhitespaceTokenizer
from nltk.stem import WordNetLemmatizer
from nltk import word_tokenize
from nltk.stem import PorterStemmer
import pandas as pd
from os import path
from PIL import Image
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
```

In [2]:

```
cd Datasets
```

```
C:\Users\bnidh\Desktop\DSPM\Final Project\Datasets
```

In [3]:

```
path = os.getcwd()
path
```

Out[3]:

```
'C:\\Users\\bnidh\\Desktop\\DSPM\\Final Project\\Datasets'
```

In [4]:

```python
all_data = pd.DataFrame()

for f in glob.glob(path+"/*.xlsx"):
    df = pd.read_excel(f)
    all_data = all_data.append(df,ignore_index=True)
```

In [5]:

```python
all_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 463691 entries, 0 to 463690
Data columns (total 41 columns):
Post ID                                   444153 non-null object
Sound Bite Text                           463680 non-null object
Ratings and Scores                        0 non-null float64
Title                                     265883 non-null object
Source Type                               444153 non-null object
Post Type                                 444153 non-null object
Media Type                                444153 non-null object
URL                                       444153 non-null object
Domain                                    444153 non-null object
Published Date (GMT-04:00) New York       463680 non-null object
Author Gender                             444153 non-null object
Author URL                                329066 non-null object
Author Name                               347593 non-null object
Author Handle                             186467 non-null object
Author ID                                 186934 non-null object
Author Location - Country 1               163321 non-null object
Author Location - State/Province 1        38069 non-null object
Author Location - City 1                  32713 non-null object
Author Location - Country 2               405 non-null object
Author Location - State/Province 2        348 non-null object
Author Location - City 2                  311 non-null object
Author Location - Other                   47 non-null object
No. of Followers/Daily Unique Visitors    444153 non-null float64
Professions                               7236 non-null object
Interests                                 12163 non-null object
Positive Objects                          98429 non-null object
Negative Objects                          44639 non-null object
Richness                                  444153 non-null float64
Tags                                      0 non-null float64
Quoted Post                               309 non-null object
Quoted Author Name                        309 non-null object
Quoted Author Handle                      309 non-null object
Total Engagements                         49587 non-null float64
Post Comments                             28915 non-null float64
Post Likes                                47504 non-null float64
Post Shares                               1003 non-null float64
Post Views                                0 non-null float64
Post Dislikes                             0 non-null float64
Product Name                              11355 non-null object
Product Hierarchy                         0 non-null float64
Rating                                    2261 non-null float64
dtypes: float64(12), object(29)
memory usage: 145.0+ MB
```

In [6]:

```
512711-511268
```

Out[6]:

```
1443
```

In [7]:

```
all_data_1 = all_data.drop_duplicates()
##512711 - 511268 =1443
```

In [8]:

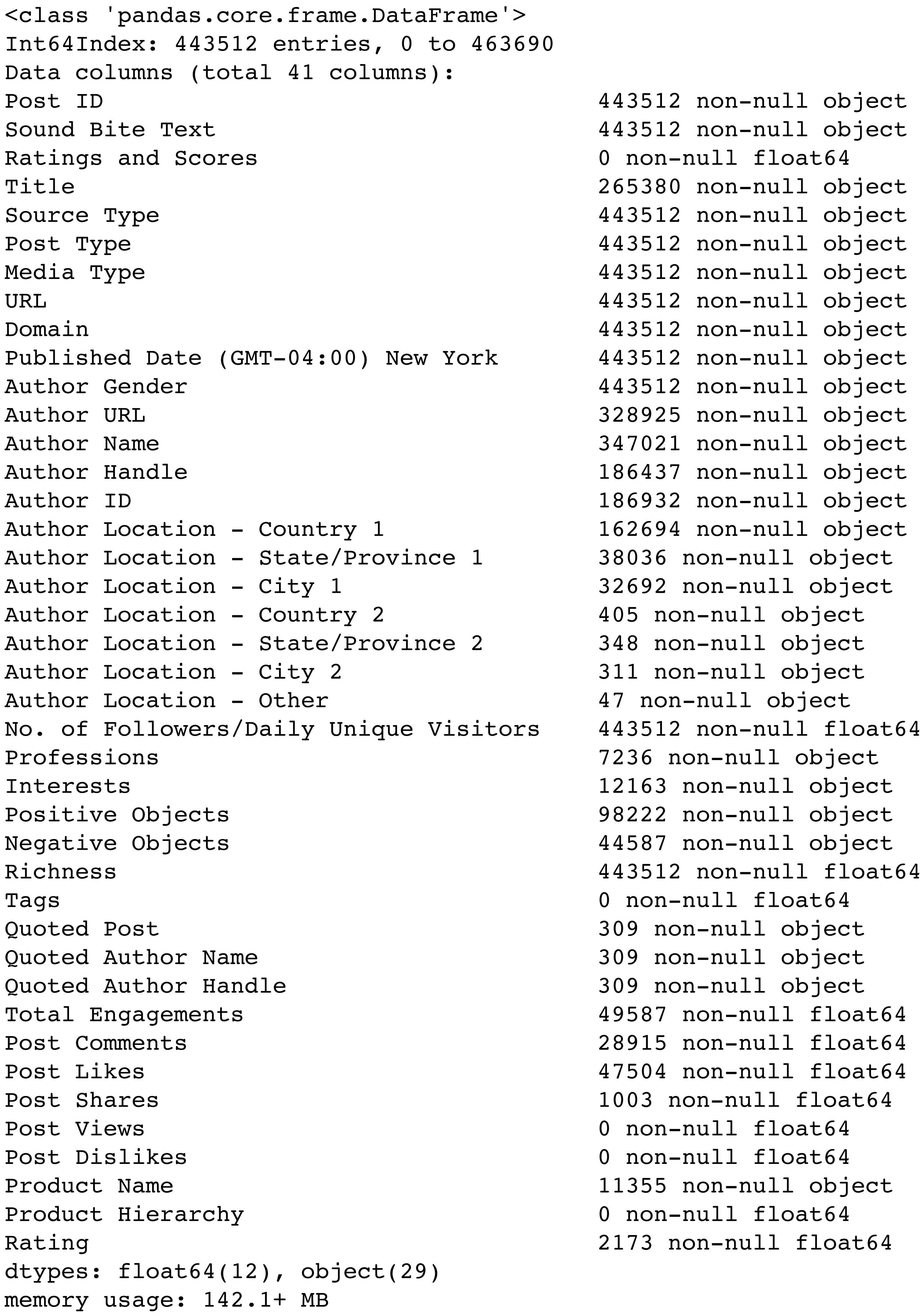
```
pd.set_option('display.max_columns', 500)
pd.set_option('display.max_rows', 500)
all_data.head(3)
```

Out[8]:

	Post ID	Sound Bite Text	Ratings and Scores	Title	Source Type	Post Type	Medi Typ
0	https://cellphoneforums.net/onlineunlocks-com/...	Samsung Galaxy S8 G950U G950P Sprint unlock by...	NaN	Samsung Galaxy S8 G950U G950P Sprint unlock by...	Blogs	Original	N Medi
1	14759701913167960370	According to More, there are moreover circulat...	NaN	IOS 11 Update – Features And Updates leaked	Blogs	Original	Lir
2	13957544623200282820	After a monster reception for the iPhone 6 ser...	NaN	Apple's latest China ad features a very differ...	Blogs	Original	N Medi

In [9]:

```
all_data_2 = all_data_1.dropna(axis=0, subset=['Post ID' , 'Sound Bite Text' , '
Published Date (GMT-04:00) New York' , 'No. of Followers/Daily Unique Visitors']
)
all_data_2.info()
#492518
```



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 443512 entries, 0 to 463690
Data columns (total 41 columns):
Post ID                                   443512 non-null object
Sound Bite Text                           443512 non-null object
Ratings and Scores                        0 non-null float64
Title                                     265380 non-null object
Source Type                               443512 non-null object
Post Type                                 443512 non-null object
Media Type                                443512 non-null object
URL                                       443512 non-null object
Domain                                    443512 non-null object
Published Date (GMT-04:00) New York       443512 non-null object
Author Gender                             443512 non-null object
Author URL                                328925 non-null object
Author Name                               347021 non-null object
Author Handle                             186437 non-null object
Author ID                                 186932 non-null object
Author Location - Country 1               162694 non-null object
Author Location - State/Province 1        38036 non-null object
Author Location - City 1                  32692 non-null object
Author Location - Country 2               405 non-null object
Author Location - State/Province 2        348 non-null object
Author Location - City 2                  311 non-null object
Author Location - Other                   47 non-null object
No. of Followers/Daily Unique Visitors    443512 non-null float64
Professions                               7236 non-null object
Interests                                 12163 non-null object
Positive Objects                          98222 non-null object
Negative Objects                          44587 non-null object
Richness                                  443512 non-null float64
Tags                                      0 non-null float64
Quoted Post                               309 non-null object
Quoted Author Name                        309 non-null object
Quoted Author Handle                      309 non-null object
Total Engagements                         49587 non-null float64
Post Comments                             28915 non-null float64
Post Likes                                47504 non-null float64
Post Shares                               1003 non-null float64
Post Views                                0 non-null float64
Post Dislikes                             0 non-null float64
Product Name                              11355 non-null object
Product Hierarchy                         0 non-null float64
Rating                                    2173 non-null float64
dtypes: float64(12), object(29)
memory usage: 142.1+ MB
```

In [10]:

```
all_data_2.isnull().sum()
```

Out[10]:

```
Post ID                                        0
Sound Bite Text                                0
Ratings and Scores                        443512
Title                                     178132
Source Type                                    0
Post Type                                      0
Media Type                                     0
URL                                            0
Domain                                         0
Published Date (GMT-04:00) New York            0
Author Gender                                  0
Author URL                                114587
Author Name                                96491
Author Handle                             257075
Author ID                                 256580
Author Location - Country 1               280818
Author Location - State/Province 1        405476
Author Location - City 1                  410820
Author Location - Country 2               443107
Author Location - State/Province 2        443164
Author Location - City 2                  443201
Author Location - Other                   443465
No. of Followers/Daily Unique Visitors         0
Professions                               436276
Interests                                 431349
Positive Objects                          345290
Negative Objects                          398925
Richness                                       0
Tags                                      443512
Quoted Post                               443203
Quoted Author Name                        443203
Quoted Author Handle                      443203
Total Engagements                         393925
Post Comments                             414597
Post Likes                                396008
Post Shares                               442509
Post Views                                443512
Post Dislikes                             443512
Product Name                              432157
Product Hierarchy                         443512
Rating                                    441339
dtype: int64
```

In [11]:

```
all_data_3= all_data_2[all_data_2['Post Type'] == 'Original']
all_data_3.info()
## 492518 to 409755
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 364042 entries, 0 to 463688
Data columns (total 41 columns):
Post ID                                   364042 non-null object
Sound Bite Text                           364042 non-null object
Ratings and Scores                        0 non-null float64
Title                                     245942 non-null object
Source Type                               364042 non-null object
Post Type                                 364042 non-null object
Media Type                                364042 non-null object
URL                                       364042 non-null object
Domain                                    364042 non-null object
Published Date (GMT-04:00) New York       364042 non-null object
Author Gender                             364042 non-null object
Author URL                                250723 non-null object
Author Name                               269588 non-null object
Author Handle                             164525 non-null object
Author ID                                 153674 non-null object
Author Location - Country 1               123991 non-null object
Author Location - State/Province 1        24968 non-null object
Author Location - City 1                  22547 non-null object
Author Location - Country 2               185 non-null object
Author Location - State/Province 2        159 non-null object
Author Location - City 2                  148 non-null object
Author Location - Other                   23 non-null object
No. of Followers/Daily Unique Visitors    364042 non-null float64
Professions                               4876 non-null object
Interests                                 8136 non-null object
Positive Objects                          81323 non-null object
Negative Objects                          35799 non-null object
Richness                                  364042 non-null float64
Tags                                      0 non-null float64
Quoted Post                               309 non-null object
Quoted Author Name                        309 non-null object
Quoted Author Handle                      309 non-null object
Total Engagements                         44900 non-null float64
Post Comments                             26612 non-null float64
Post Likes                                44103 non-null float64
Post Shares                               1003 non-null float64
Post Views                                0 non-null float64
Post Dislikes                             0 non-null float64
Product Name                              11355 non-null object
Product Hierarchy                         0 non-null float64
Rating                                    2173 non-null float64
dtypes: float64(12), object(29)
memory usage: 116.7+ MB
```

In [12]:

```
all_data_4= all_data_3[all_data_3['Sound Bite Text'] != 'Post deleted by the author.']
all_data_4.info()
#409755 to 409755
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 364042 entries, 0 to 463688
Data columns (total 41 columns):
Post ID                                   364042 non-null object
Sound Bite Text                           364042 non-null object
Ratings and Scores                        0 non-null float64
Title                                     245942 non-null object
Source Type                               364042 non-null object
Post Type                                 364042 non-null object
Media Type                                364042 non-null object
URL                                       364042 non-null object
Domain                                    364042 non-null object
Published Date (GMT-04:00) New York       364042 non-null object
Author Gender                             364042 non-null object
Author URL                                250723 non-null object
Author Name                               269588 non-null object
Author Handle                             164525 non-null object
Author ID                                 153674 non-null object
Author Location - Country 1               123991 non-null object
Author Location - State/Province 1        24968 non-null object
Author Location - City 1                  22547 non-null object
Author Location - Country 2               185 non-null object
Author Location - State/Province 2        159 non-null object
Author Location - City 2                  148 non-null object
Author Location - Other                   23 non-null object
No. of Followers/Daily Unique Visitors    364042 non-null float64
Professions                               4876 non-null object
Interests                                 8136 non-null object
Positive Objects                          81323 non-null object
Negative Objects                          35799 non-null object
Richness                                  364042 non-null float64
Tags                                      0 non-null float64
Quoted Post                               309 non-null object
Quoted Author Name                        309 non-null object
Quoted Author Handle                      309 non-null object
Total Engagements                         44900 non-null float64
Post Comments                             26612 non-null float64
Post Likes                                44103 non-null float64
Post Shares                               1003 non-null float64
Post Views                                0 non-null float64
Post Dislikes                             0 non-null float64
Product Name                              11355 non-null object
Product Hierarchy                         0 non-null float64
Rating                                    2173 non-null float64
dtypes: float64(12), object(29)
memory usage: 116.7+ MB
```

In [13]:

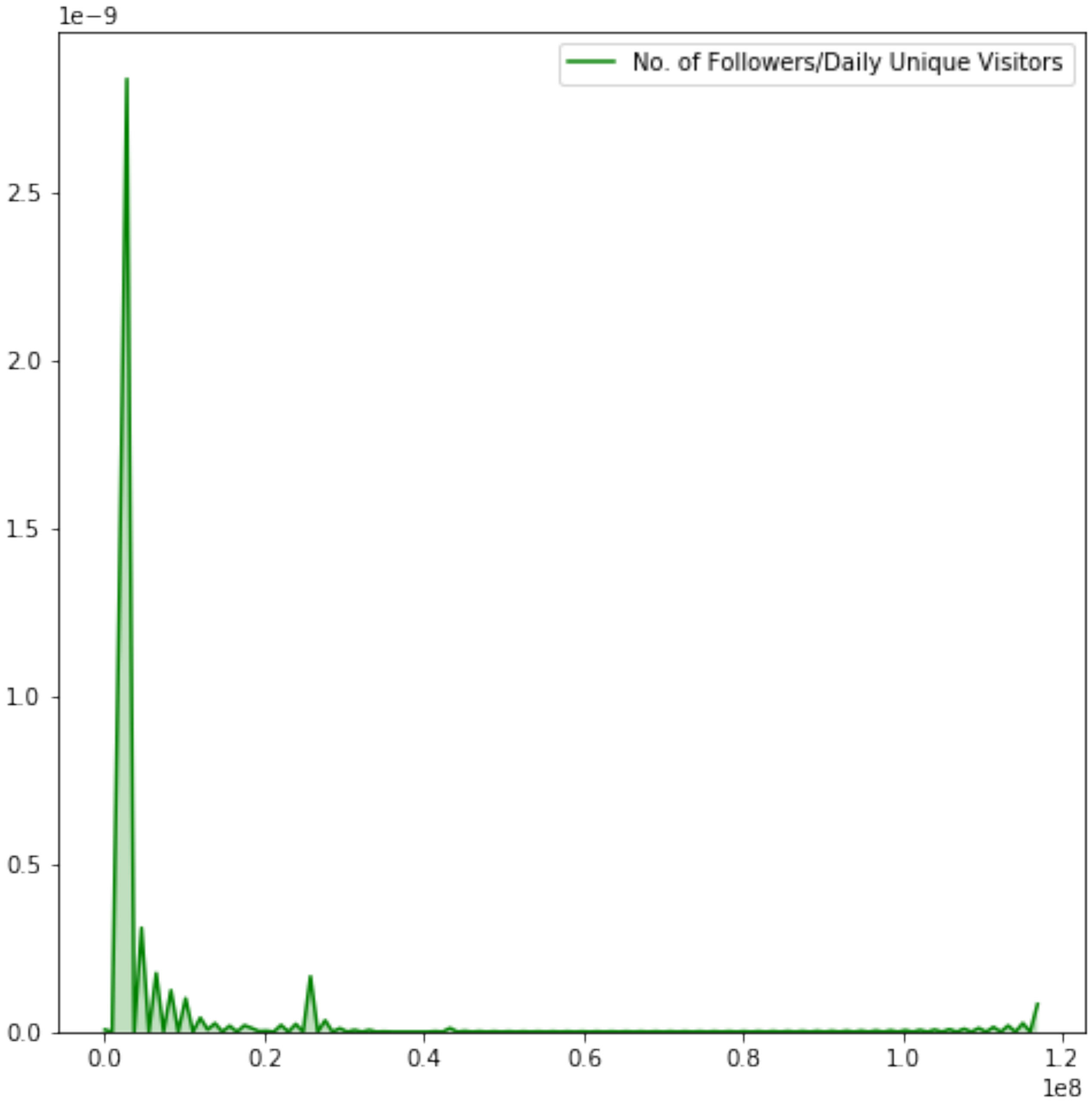
```
all_data_4['No. of Followers/Daily Unique Visitors'].describe()
```

Out[13]:

```
count    3.640420e+05
mean     4.061777e+04
std      9.065021e+05
min      0.000000e+00
25%      0.000000e+00
50%      0.000000e+00
75%      0.000000e+00
max      1.168545e+08
Name: No. of Followers/Daily Unique Visitors, dtype: float64
```

In [14]:

```
plt.figure(figsize = (8,8 ))
sns.kdeplot(all_data_4['No. of Followers/Daily Unique Visitors'], color="green",
shade=True)
plt.show()
```



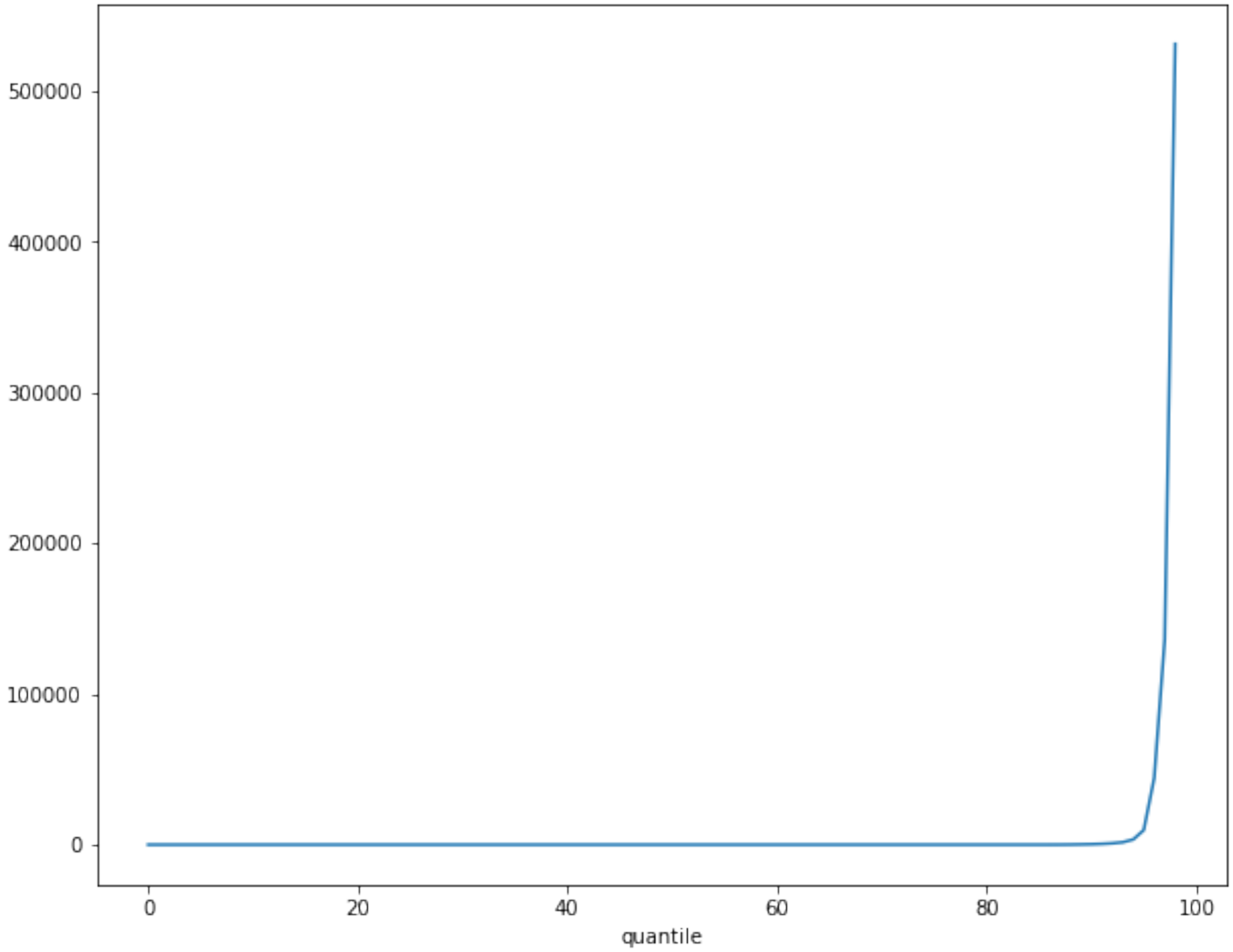
In [15]:

```
values = []
for a in range(0,99) :
    #print(a,"th quantile value :" , all_data_4['No. of Followers/Daily Unique V
isitors'].quantile(a/100))
    values.append(all_data_4['No. of Followers/Daily Unique Visitors'].quantile(
a/100))
print(values)
```

```
[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, 0.0, 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, 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, 0.0, 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, 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, 19.0, 75.0, 179.0, 318.0, 550.0, 931.0, 1
620.0, 3478.539999999979, 9688.0, 44330.0, 136043.580000001, 531195.
0]
```

In [16]:

```
plt.figure(figsize=(10,8))
plt.plot(values)
plt.xlabel('quantile')
plt.show()
```



In [17]:

```
qcap = all_data_4['No. of Followers/Daily Unique Visitors'].quantile(0.95)
qcap
```

Out[17]:

```
9688.0
```

In [18]:

```
all_data_4.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 364042 entries, 0 to 463688
Data columns (total 41 columns):
Post ID                                   364042 non-null object
Sound Bite Text                           364042 non-null object
Ratings and Scores                        0 non-null float64
Title                                     245942 non-null object
Source Type                               364042 non-null object
Post Type                                 364042 non-null object
Media Type                                364042 non-null object
URL                                       364042 non-null object
Domain                                    364042 non-null object
Published Date (GMT-04:00) New York       364042 non-null object
Author Gender                             364042 non-null object
Author URL                                250723 non-null object
Author Name                               269588 non-null object
Author Handle                             164525 non-null object
Author ID                                 153674 non-null object
Author Location - Country 1               123991 non-null object
Author Location - State/Province 1        24968 non-null object
Author Location - City 1                  22547 non-null object
Author Location - Country 2               185 non-null object
Author Location - State/Province 2        159 non-null object
Author Location - City 2                  148 non-null object
Author Location - Other                   23 non-null object
No. of Followers/Daily Unique Visitors    364042 non-null float64
Professions                               4876 non-null object
Interests                                 8136 non-null object
Positive Objects                          81323 non-null object
Negative Objects                          35799 non-null object
Richness                                  364042 non-null float64
Tags                                      0 non-null float64
Quoted Post                               309 non-null object
Quoted Author Name                        309 non-null object
Quoted Author Handle                      309 non-null object
Total Engagements                         44900 non-null float64
Post Comments                             26612 non-null float64
Post Likes                                44103 non-null float64
Post Shares                               1003 non-null float64
Post Views                                0 non-null float64
Post Dislikes                             0 non-null float64
Product Name                              11355 non-null object
Product Hierarchy                         0 non-null float64
Rating                                    2173 non-null float64
dtypes: float64(12), object(29)
memory usage: 116.7+ MB
```

In [19]:

```
all_data_5 = all_data_4.loc[all_data_4['No. of Followers/Daily Unique Visitors']
< qcap]
all_data_5.info()
#409755 to 389185
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 345521 entries, 0 to 463685
Data columns (total 41 columns):
Post ID                                   345521 non-null object
Sound Bite Text                           345521 non-null object
Ratings and Scores                        0 non-null float64
Title                                     236800 non-null object
Source Type                               345521 non-null object
Post Type                                 345521 non-null object
Media Type                                345521 non-null object
URL                                       345521 non-null object
Domain                                    345521 non-null object
Published Date (GMT-04:00) New York       345521 non-null object
Author Gender                             345521 non-null object
Author URL                                242798 non-null object
Author Name                               256175 non-null object
Author Handle                             159164 non-null object
Author ID                                 145470 non-null object
Author Location - Country 1               115302 non-null object
Author Location - State/Province 1        23116 non-null object
Author Location - City 1                  20849 non-null object
Author Location - Country 2               149 non-null object
Author Location - State/Province 2        128 non-null object
Author Location - City 2                  117 non-null object
Author Location - Other                   20 non-null object
No. of Followers/Daily Unique Visitors    345521 non-null float64
Professions                               4095 non-null object
Interests                                 6984 non-null object
Positive Objects                          77168 non-null object
Negative Objects                          33636 non-null object
Richness                                  345521 non-null float64
Tags                                      0 non-null float64
Quoted Post                               272 non-null object
Quoted Author Name                        272 non-null object
Quoted Author Handle                      272 non-null object
Total Engagements                         41599 non-null float64
Post Comments                             23969 non-null float64
Post Likes                                40808 non-null float64
Post Shares                               42 non-null float64
Post Views                                0 non-null float64
Post Dislikes                             0 non-null float64
Product Name                              11345 non-null object
Product Hierarchy                         0 non-null float64
Rating                                    2099 non-null float64
dtypes: float64(12), object(29)
memory usage: 110.7+ MB
```

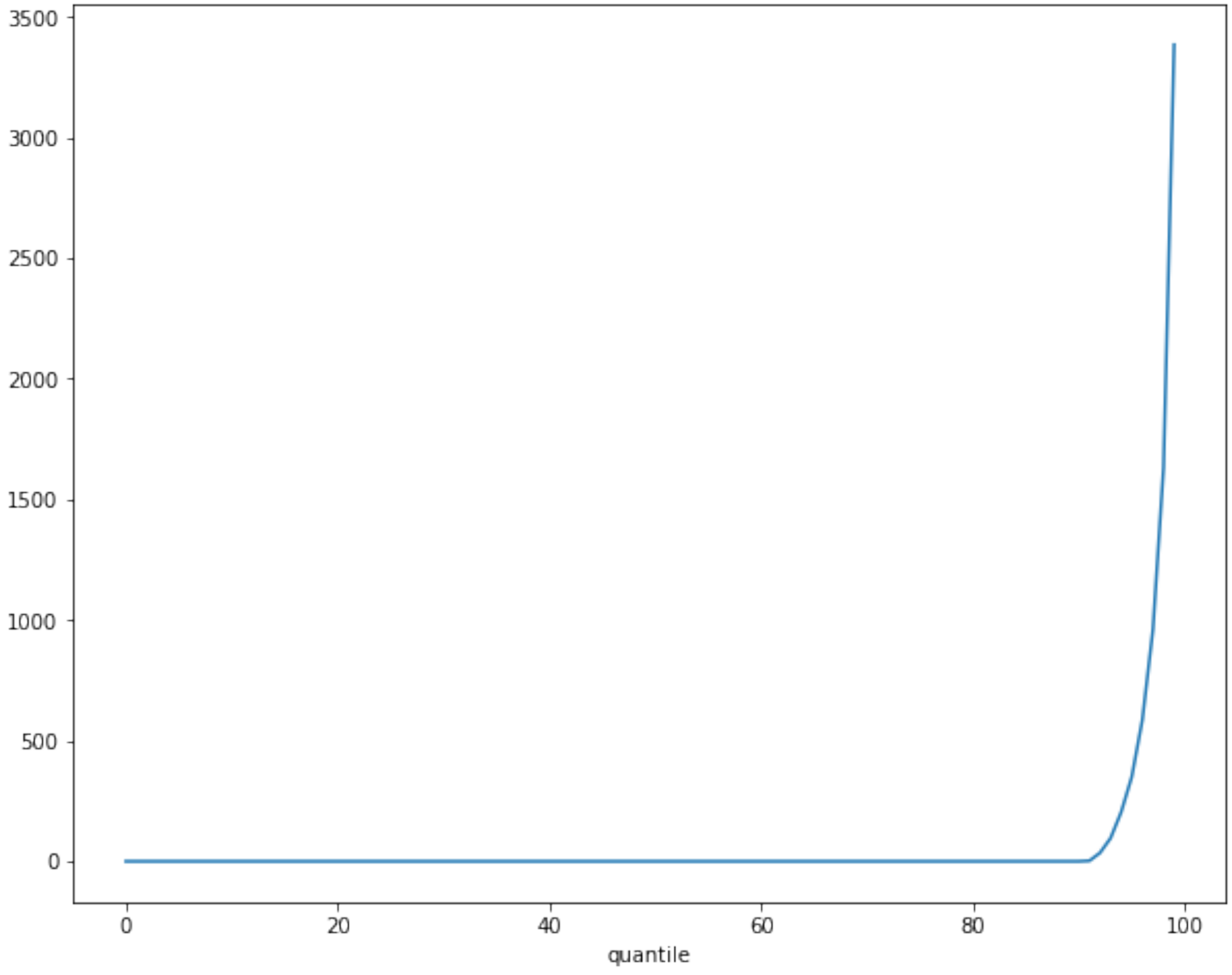
In [20]:

```
values = []
for a in range(0,100) :
    #print(a,"th quantile value :" , all_data_4['No. of Followers/Daily Unique V
isitors'].quantile(a/100))
    values.append(all_data_5['No. of Followers/Daily Unique Visitors'].quantile(
a/100))
print(values)
```

```
[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, 0.0, 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, 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, 0.0, 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, 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, 0.0, 0.0, 0.0, 0.0, 2.0, 35.0, 96.0, 204.
0, 350.0, 586.0, 960.0, 1631.0, 3385.0]
```

In [21]:

```
plt.figure(figsize=(10,8))
plt.plot(values)
plt.xlabel('quantile')
plt.show()
```



In [22]:

```
all_data_5.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 345521 entries, 0 to 463685
Data columns (total 41 columns):
Post ID                                   345521 non-null object
Sound Bite Text                           345521 non-null object
Ratings and Scores                        0 non-null float64
Title                                     236800 non-null object
Source Type                               345521 non-null object
Post Type                                 345521 non-null object
Media Type                                345521 non-null object
URL                                       345521 non-null object
Domain                                    345521 non-null object
Published Date (GMT-04:00) New York       345521 non-null object
Author Gender                             345521 non-null object
Author URL                                242798 non-null object
Author Name                               256175 non-null object
Author Handle                             159164 non-null object
Author ID                                 145470 non-null object
Author Location - Country 1               115302 non-null object
Author Location - State/Province 1        23116 non-null object
Author Location - City 1                  20849 non-null object
Author Location - Country 2               149 non-null object
Author Location - State/Province 2        128 non-null object
Author Location - City 2                  117 non-null object
Author Location - Other                   20 non-null object
No. of Followers/Daily Unique Visitors    345521 non-null float64
Professions                               4095 non-null object
Interests                                 6984 non-null object
Positive Objects                          77168 non-null object
Negative Objects                          33636 non-null object
Richness                                  345521 non-null float64
Tags                                      0 non-null float64
Quoted Post                               272 non-null object
Quoted Author Name                        272 non-null object
Quoted Author Handle                      272 non-null object
Total Engagements                         41599 non-null float64
Post Comments                             23969 non-null float64
Post Likes                                40808 non-null float64
Post Shares                               42 non-null float64
Post Views                                0 non-null float64
Post Dislikes                             0 non-null float64
Product Name                              11345 non-null object
Product Hierarchy                         0 non-null float64
Rating                                    2099 non-null float64
dtypes: float64(12), object(29)
memory usage: 110.7+ MB
```

In [23]:

```
q95 = all_data_5['No. of Followers/Daily Unique Visitors'].quantile(0.95)
q95
```

Out[23]:

```
350.0
```

In [24]:

```
all_data_6 = all_data_5[all_data_5['No. of Followers/Daily Unique Visitors'] < q
95]
all_data_6['Type'] = 'Normal'
professional_df = all_data_5[all_data_5['No. of Followers/Daily Unique Visitors'
] >= q95]
professional_df['Type'] = 'Professional'

temp = [all_data_6 , professional_df]
all_data_7 = pd.concat(temp)
```

```
C:\Users\bnidh\Anaconda3\lib\site-packages\ipykernel_launcher.py:2:
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: http://pandas.pydata.org/panda
s-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-cop
y

C:\Users\bnidh\Anaconda3\lib\site-packages\ipykernel_launcher.py:4:
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: http://pandas.pydata.org/panda
s-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-cop
y
  after removing the cwd from sys.path.
```

In [25]:

```
all_data_7.groupby(['Type']).count()
```

Out[25]:

	Post ID	Sound Bite Text	Ratings and Scores	Title	Source Type	Post Type	Media Type	URL	Domain	Publ ((New
Type										
Normal	328234	328234	0	236641	328234	328234	328234	328234	328234	3:
Professional	17287	17287	0	159	17287	17287	17287	17287	17287	'

In [26]:

```
normal_df=all_data_6
```

In [27]:

```
sources = pd.DataFrame(normal_df.groupby('Source Type')['Sound Bite Text'].count
())
sources = sources.reset_index()
sources = sources.sort_values(by = 'Sound Bite Text', ascending = False)
sources
```

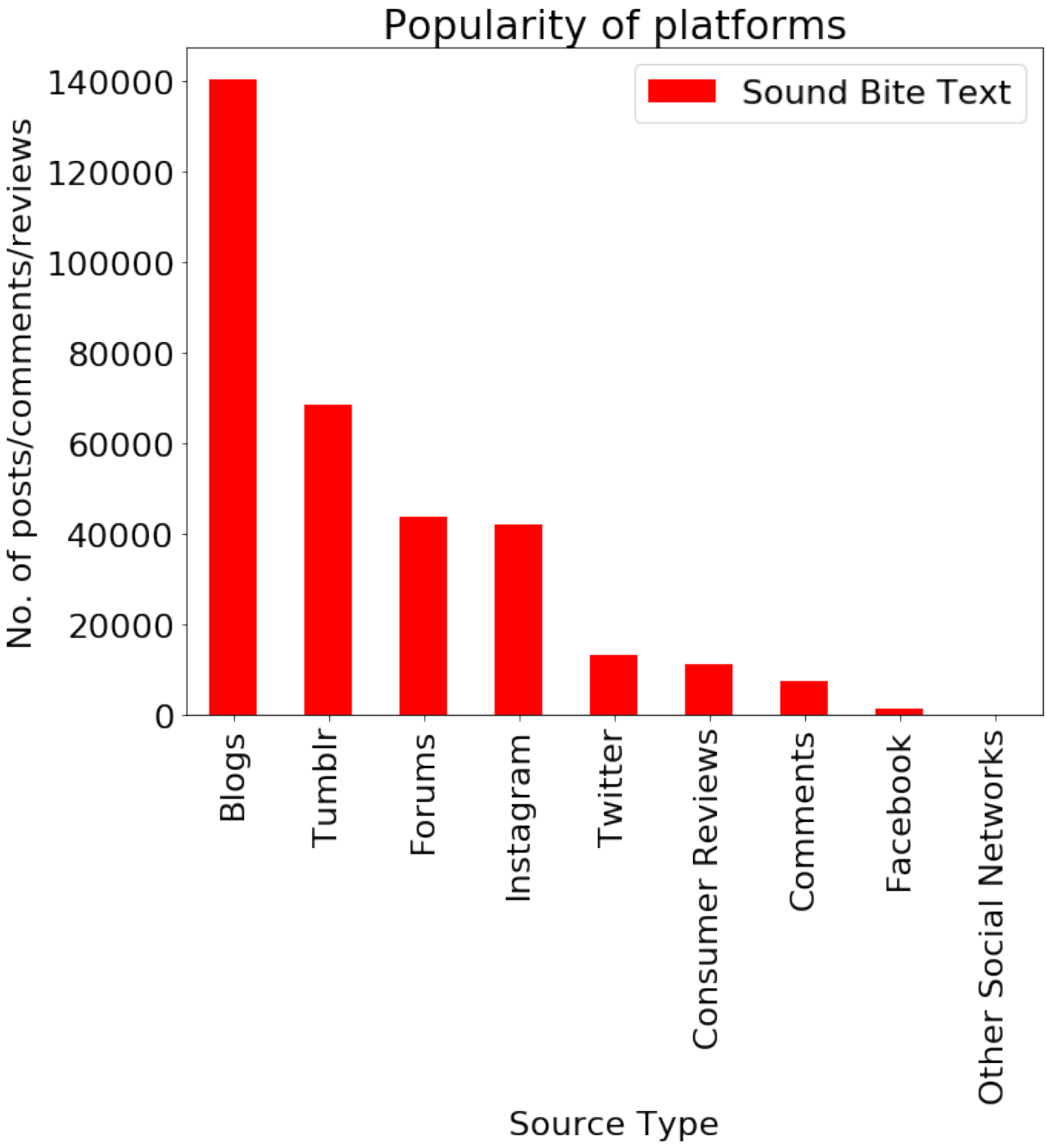
Out[27]:

	Source Type	Sound Bite Text
0	Blogs	140297
7	Tumblr	68589
4	Forums	43790
5	Instagram	42093
8	Twitter	13112
2	Consumer Reviews	11394
1	Comments	7480
3	Facebook	1456
6	Other Social Networks	23

```
In [28]:
```

```python
plt.rcParams['figure.figsize'] = (10,8)
plt.rcParams.update({'font.size': 22})
ax = sources.plot(x="Source Type", y="Sound Bite Text", kind="bar", color = "red")
ax.set_xlabel("Source Type")
ax.set_ylabel("No. of posts/comments/reviews")
ax.set_title("Popularity of platforms")
```

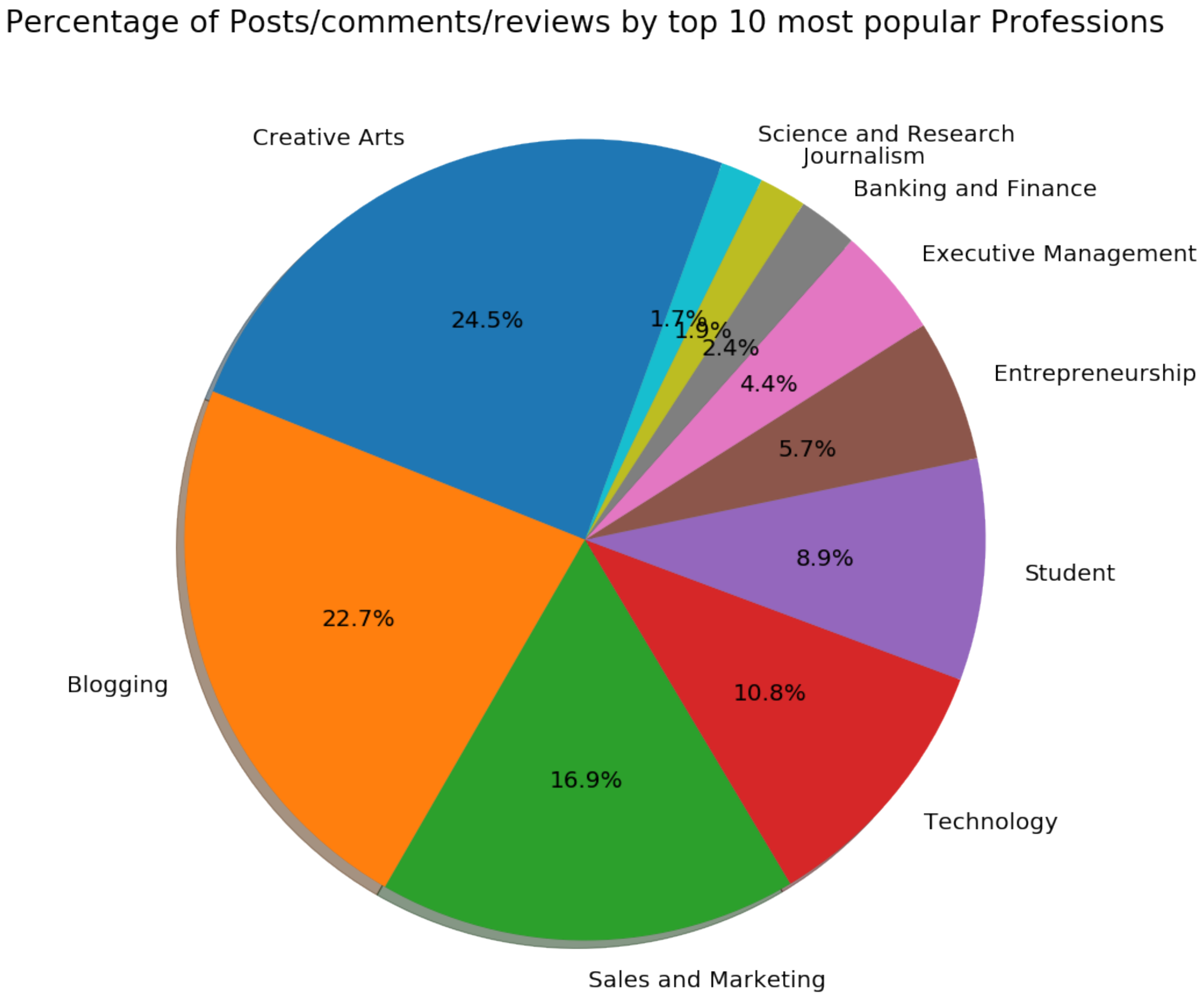
```
Out[28]:

Text(0.5, 1.0, 'Popularity of platforms')
```



In [29]:

```
#analysis on the percentage of comments/posts by top 10 most frequently repeating profession
prof = pd.DataFrame(normal_df.groupby(['Professions'])['Sound Bite Text'].count())
prof = prof.reset_index()
print('Total no. of professions to which people belong to:', prof['Professions'].nunique())
prof = prof.sort_values('Sound Bite Text', ascending = False)
prof = prof.head(10)
plt.figure(figsize = (15,13))
plt.pie(prof['Sound Bite Text'], labels=prof['Professions'], startangle=70, autopct='%1.1f%%', textprops={'fontsize': 20}, shadow=True)
plt.title('Percentage of Posts/comments/reviews by top 10 most popular Professions')
plt.tight_layout()
plt.show()
```

```
Total no. of professions to which people belong to: 78
```



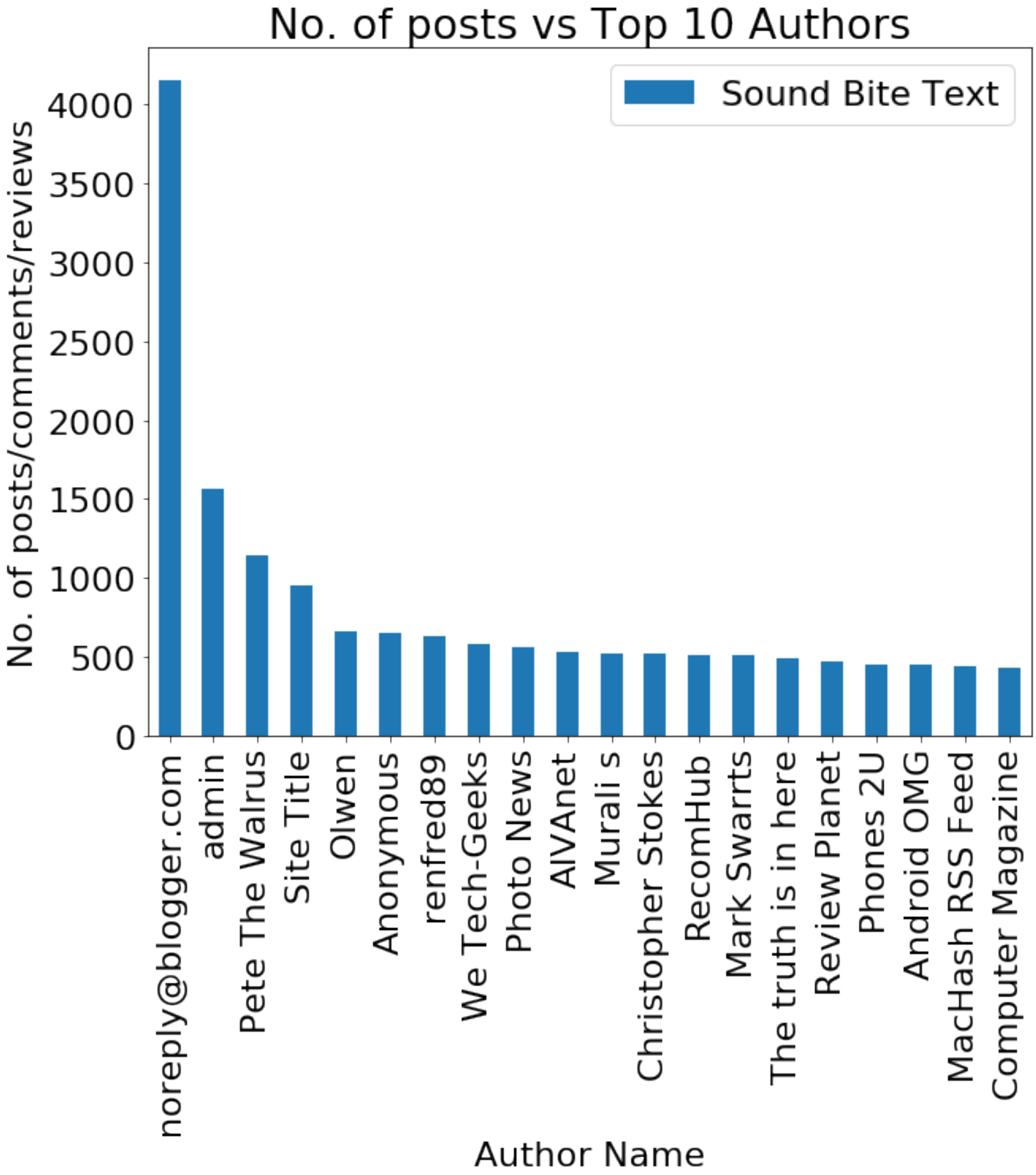
In [30]:

```
#analysis on authors and number of comments/posts by top 20 most frequently posting authors
authors_df = normal_df
print('Total no. of authors posting on twitter & non-twitter sites:',authors_df['Author Name'].nunique())
authors_df = pd.DataFrame(authors_df.groupby(['Author Name','No. of Followers/Daily Unique Visitors'])['Sound Bite Text'].count())
authors_df = authors_df.reset_index()
authors_df = authors_df.sort_values('Sound Bite Text', ascending = False)
authors_df = authors_df.head(20)
ax = authors_df.plot(x="Author Name", y="Sound Bite Text", kind="bar")
ax.set_ylabel('No. of posts/comments/reviews')
ax.set_title('No. of posts vs Top 10 Authors')
```

```
Total no. of authors posting on twitter & non-twitter sites: 80692
```

Out[30]:

```
Text(0.5, 1.0, 'No. of posts vs Top 10 Authors')
```


No. of posts vs Top 10 Authors
Sound Bite Text
No. of posts/comments/reviews
0
500
1000
1500
2000
2500
3000
3500
4000
noreply@blogger.com
admin
Pete The Walrus
Site Title
Olwen
Anonymous
renfred89
We Tech-Geeks
Photo News
AIVAnet
Murali s
Christopher Stokes
RecomHub
Mark Swarrts
The truth is in here
Review Planet
Phones 2U
Android OMG
MacHash RSS Feed
Computer Magazine
Author Name

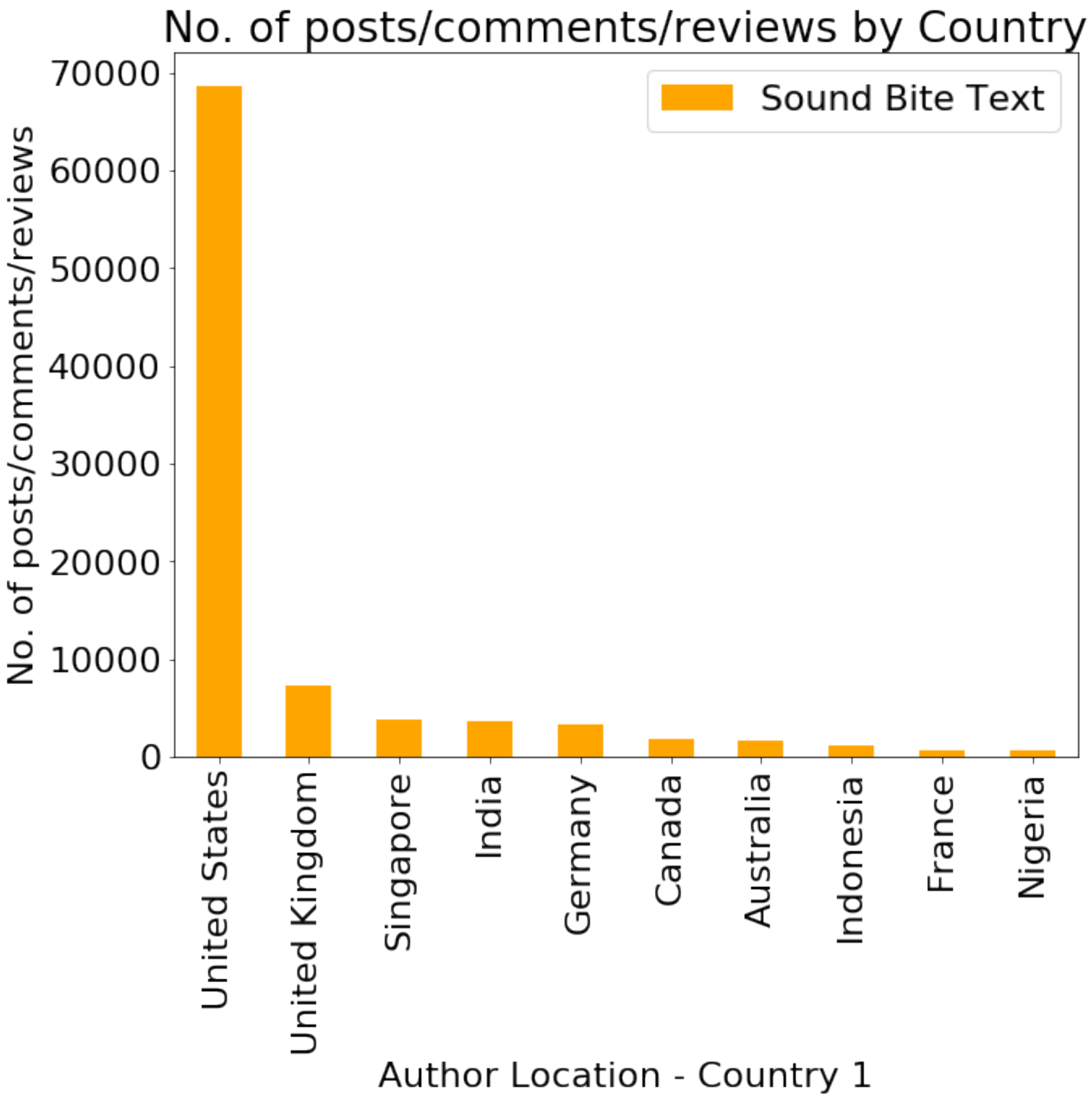
In [31]:

```
print('Total no. of countries to which people who have posted belong to is' , normal_df['Author Location - Country 1'].nunique())
country = pd.DataFrame(normal_df.groupby(['Author Location - Country 1'])['Sound Bite Text'].count())
country = country.reset_index()
country = country.sort_values(by ='Sound Bite Text', ascending = False)
country = country.head(10)
plt.rcParams['figure.figsize'] = (10,8)
ax = country.plot(x="Author Location - Country 1", y="Sound Bite Text", kind="bar", color = "orange")
ax.set_ylabel('No. of posts/comments/reviews')
ax.set_title('No. of posts/comments/reviews by Country')
```

```
Total no. of countries to which people who have posted belong to is
178
```

Out[31]:

```
Text(0.5, 1.0, 'No. of posts/comments/reviews by Country')
```



In [32]:

```
normal_df['Published Date (GMT-04:00) New York']= pd.to_datetime\
(normal_df['Published Date (GMT-04:00) New York']).dt.date
```

```
C:\Users\bnidh\Anaconda3\lib\site-packages\ipykernel_launcher.py:2:
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: http://pandas.pydata.org/panda
s-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-cop
y
```

In [33]:

```
normal_df.sort_values(inplace=True,ascending=True,by='Published Date (GMT-04:00)
New York')
```

```
C:\Users\bnidh\Anaconda3\lib\site-packages\ipykernel_launcher.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-cop
y
  """Entry point for launching an IPython kernel.
```

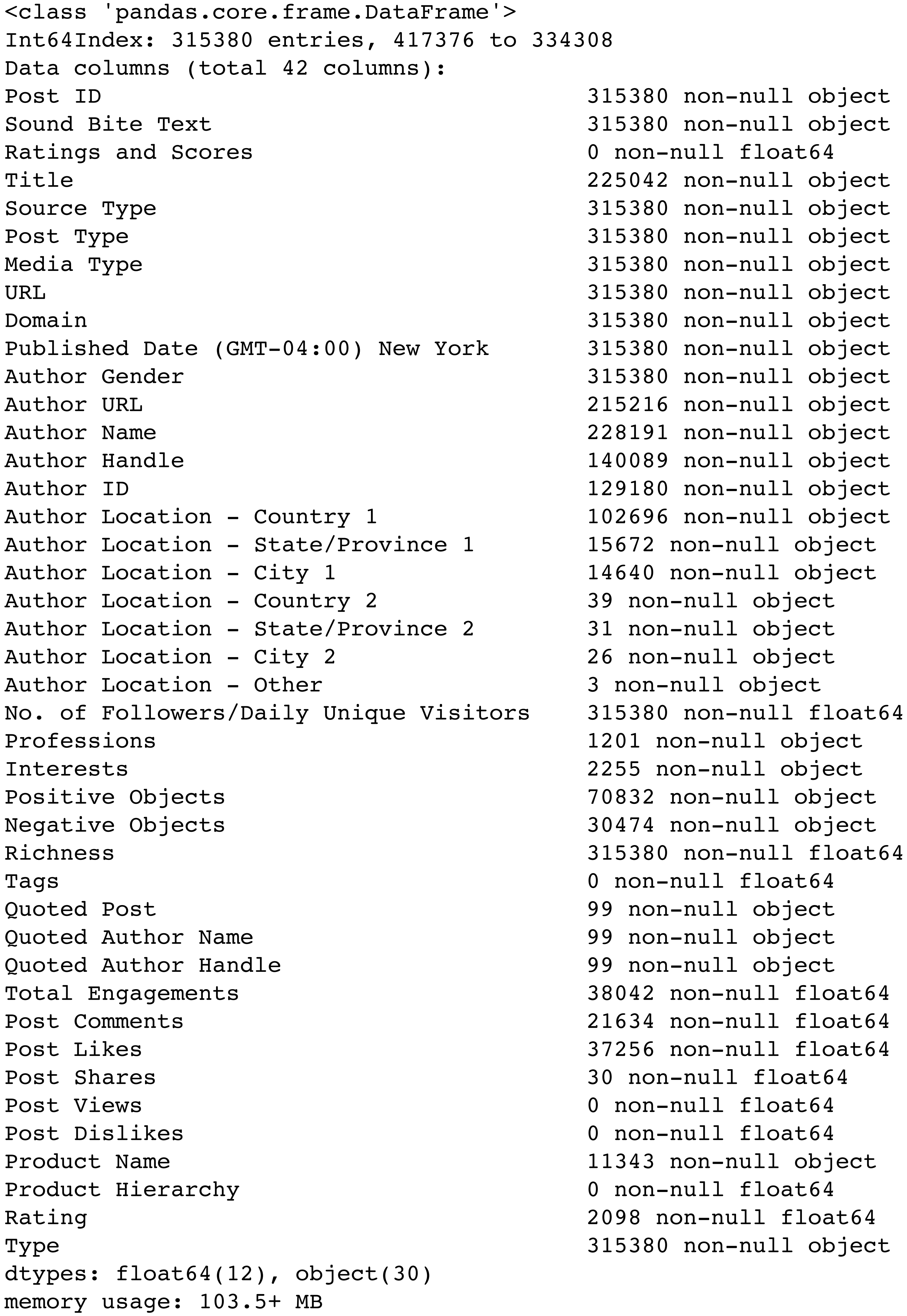
In [34]:

```
normal_df.drop(normal_df[normal_df["Author Name"].str.contains("News|news",na=Fa
lse)].index, inplace = True)
```

```
C:\Users\bnidh\Anaconda3\lib\site-packages\pandas\core\frame.py:4102
: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/panda
s-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-cop
y
  errors=errors,
```

In [35]:

```
normal_df.info()
```



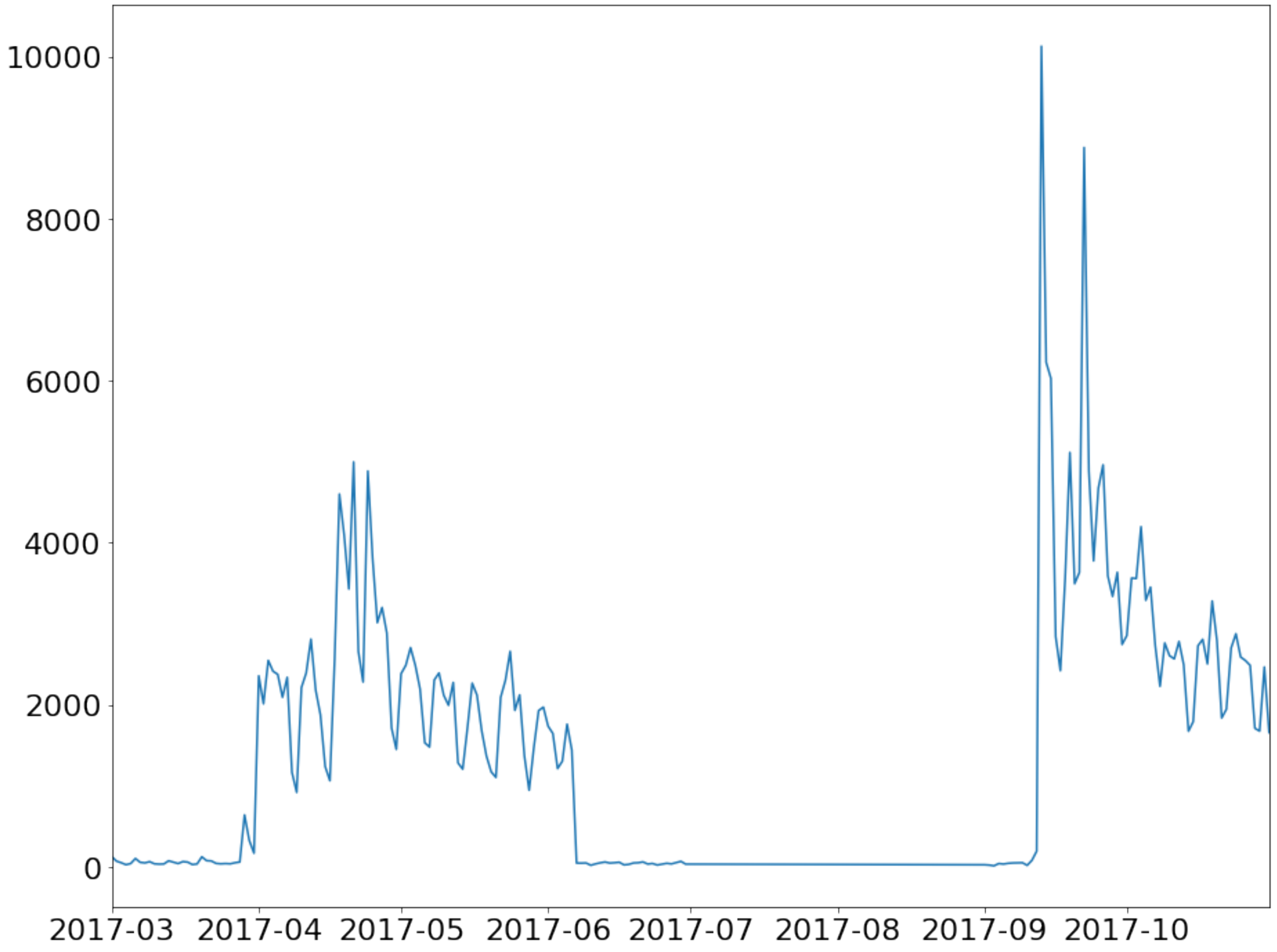
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 315380 entries, 417376 to 334308
Data columns (total 42 columns):
Post ID                                   315380 non-null object
Sound Bite Text                           315380 non-null object
Ratings and Scores                        0 non-null float64
Title                                     225042 non-null object
Source Type                               315380 non-null object
Post Type                                 315380 non-null object
Media Type                                315380 non-null object
URL                                       315380 non-null object
Domain                                    315380 non-null object
Published Date (GMT-04:00) New York       315380 non-null object
Author Gender                             315380 non-null object
Author URL                                215216 non-null object
Author Name                               228191 non-null object
Author Handle                             140089 non-null object
Author ID                                 129180 non-null object
Author Location - Country 1               102696 non-null object
Author Location - State/Province 1        15672 non-null object
Author Location - City 1                  14640 non-null object
Author Location - Country 2               39 non-null object
Author Location - State/Province 2        31 non-null object
Author Location - City 2                  26 non-null object
Author Location - Other                   3 non-null object
No. of Followers/Daily Unique Visitors    315380 non-null float64
Professions                               1201 non-null object
Interests                                 2255 non-null object
Positive Objects                          70832 non-null object
Negative Objects                          30474 non-null object
Richness                                  315380 non-null float64
Tags                                      0 non-null float64
Quoted Post                               99 non-null object
Quoted Author Name                        99 non-null object
Quoted Author Handle                      99 non-null object
Total Engagements                         38042 non-null float64
Post Comments                             21634 non-null float64
Post Likes                                37256 non-null float64
Post Shares                               30 non-null float64
Post Views                                0 non-null float64
Post Dislikes                             0 non-null float64
Product Name                              11343 non-null object
Product Hierarchy                         0 non-null float64
Rating                                    2098 non-null float64
Type                                      315380 non-null object
dtypes: float64(12), object(30)
memory usage: 103.5+ MB
```

In [36]:

```
s = normal_df['Published Date (GMT-04:00) New York'].value_counts()
plt.figure(figsize=(15,12))
s.plot()
```

Out[36]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x266003a1108>
```


10000
8000
6000
4000
2000
0
2017-03
2017-04
2017-05
2017-06
2017-07
2017-08
2017-09
2017-10