Prepared_data

December 7, 2018

1 Drop nulls and merge labeled DataSets

In [11]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from bs4 import BeautifulSoup
import re
```

1.1 First DataSet

In [48]:
```
#Here we upload the users and the extract tweets to our DataFrame
cols = ['','Tweets','Name','Sentiment','']
dataset = pd.read_excel("CLEANED_DATASET_1_Labeled.xlsx",header=None, names=cols, enc
dataset.head()
```

Out[48]:

		Tweets	Name	Sentiment	
0	NaN	Tweets	Name	Sentiment	NaN
1	0.0	tweets	Name	NaN	NaN
2	1.0	NaN	AkincilarCW	0	NaN
3	2.0	NaN	AkincilarCW	0	NaN
4	3.0	NaN	AkincilarCW	0	NaN

In [50]:
```
# Drop the columns that we don't need
dataset.drop(['',''],axis=1,inplace=True)
dataset[dataset.Tweets == 0].head(5)
```

Out[50]:
```
Empty DataFrame
Columns: [Tweets, Name, Sentiment]
Index: []
```

In [52]:
```
#dealing with missing data
sufficientcolumns = dataset.isnull().sum() < 2
df = dataset.loc[:, sufficientcolumns] # keep only columns that have less than 2 miss
df = dataset.drop(dataset.loc[dataset['Tweets'].isnull()].index) # remove the sample
df.head()
```

Out[52]:

	Tweets	Name	Sentiment
0	Tweets	Name	Sentiment

```
1                                                tweets         Name        NaN
5   egypt s official news agency mena website hack...  AkincilarCW          1
7   middle east news agency the state news agency ...  AkincilarCW          1
11  t rk hackerlar m s r resmi haber ajans n n sit...  AkincilarCW          0
```

In [53]:
```
dataset1 = df.ix[2:]
```

```
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/ipykernel_launcher.py:1: DeprecationWarnin
.ix is deprecated. Please use
.loc for label based indexing or
.iloc for positional indexing

See the documentation here:
http://pandas.pydata.org/pandas-docs/stable/indexing.html#ix-indexer-is-deprecated
  """Entry point for launching an IPython kernel.
```

In [54]:
```
dataset1.head()
```

Out[54]:
```
                                               Tweets         Name Sentiment
5   egypt s official news agency mena website hack...  AkincilarCW         1
7   middle east news agency the state news agency ...  AkincilarCW         1
11  t rk hackerlar m s r resmi haber ajans n n sit...  AkincilarCW         0
12  t rk hackerlar m s r resmi haber ajans n n sit...  AkincilarCW         0
13  t rk hackerlar m s r resmi haber ajans n n sit...  AkincilarCW         0
```

In [55]:
```
#Here we upload the users and the extract tweets to our DataFrame
cols = ['','Tweets','Name','Sentiment','','','','','','']
dataset2 = pd.read_excel("CLEANED_DATASET_2_Labeled.xlsx",header=None, names=cols, en
dataset2.head()
```

Out[55]:
```
     Tweets            Name  \
0  NaN                                                Tweets            Name
1  0.0                                                tweets            Name
2  1.0                                anonymousawake links  jacobwolf420_x
3  2.0                                                 ofc v  jacobwolf420_x
4  3.0  kangna ranaut timesofindia idk so confused is ...  jacobwolf420_x

   Sentiment
0  Sentiment NaN  Related to hacking: 1 NaN NaN NaN NaN
1        NaN NaN               Other: 0 NaN NaN NaN NaN
2          1 NaN                    NaN NaN NaN NaN NaN
3          0 NaN                    NaN NaN NaN NaN NaN
4          0 NaN                    NaN NaN NaN NaN NaN
```

In [56]:
```
# Drop the columns that we don't need
dataset2.drop(['','','','','','','','','','',''],axis=1,inplace=True)
dataset2[dataset2.Tweets == 0].head(5)
```

```
Out[56]: Empty DataFrame
         Columns: [Tweets, Name, Sentiment]
         Index: []

In [57]: #dealing with missing data
         sufficientcolumns = df_1.isnull().sum() < 2
         df = dataset2.loc[:, sufficientcolumns] # keep only columns that have less than 2 mis
         df = dataset2.drop(dataset2.loc[dataset2['Tweets'].isnull()].index) # remove the samp
         df.head()

Out[57]:                                                 Tweets            Name  \
         0                                               Tweets            Name
         1                                               tweets            Name
         2                                 anonymousawake links  jacobwolf420_x
         3                                                ofc v  jacobwolf420_x
         4  kangna ranaut timesofindia idk so confused is ...  jacobwolf420_x

            Sentiment
         0  Sentiment
         1        NaN
         2          1
         3          0
         4          0

In [58]: dataset2=df.ix[2:]

/opt/jupyterhub/anaconda/lib/python3.6/site-packages/ipykernel_launcher.py:1: DeprecationWarni
.ix is deprecated. Please use
.loc for label based indexing or
.iloc for positional indexing

See the documentation here:
http://pandas.pydata.org/pandas-docs/stable/indexing.html#ix-indexer-is-deprecated
  """Entry point for launching an IPython kernel.


In [59]: dataset2.head()

Out[59]:                                                 Tweets            Name Sentiment
         2                                 anonymousawake links  jacobwolf420_x         1
         3                                                ofc v  jacobwolf420_x         0
         4  kangna ranaut timesofindia idk so confused is ...  jacobwolf420_x         0
         8  fr x h hides ip a aims target c crack encrypte...  jacobwolf420_x         1
         9                       anonymousawake provide links  jacobwolf420_x         1

In [61]: Full_dataset = [dataset1,dataset2]
         DataSet = pd.concat(Full_dataset)
         DataSet.head()
```

Out[61]:

	Tweets	Name	Sentiment
5	egypt s official news agency mena website hack...	AkincilarCW	1
7	middle east news agency the state news agency ...	AkincilarCW	1
11	t rk hackerlar m s r resmi haber ajans n n sit...	AkincilarCW	0
12	t rk hackerlar m s r resmi haber ajans n n sit...	AkincilarCW	0
13	t rk hackerlar m s r resmi haber ajans n n sit...	AkincilarCW	0

In [63]:
```
DataSet.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8611 entries, 5 to 3625
Data columns (total 3 columns):
Tweets       8611 non-null object
Name         8611 non-null object
Sentiment    8609 non-null object
dtypes: object(3)
memory usage: 269.1+ KB
```

2 Cleaned Labeled Dataset

In [64]:
```
DataSet.to_csv('Labeled_Dataset.csv',encoding='utf-8')
csv = 'Labeled_Dataset.csv'
my_df = pd.read_csv(csv,index_col=0)
my_df
```

Out[64]:

	Tweets	Name	\
5	egypt s official news agency mena website hack...	AkincilarCW	
7	middle east news agency the state news agency ...	AkincilarCW	
11	t rk hackerlar m s r resmi haber ajans n n sit...	AkincilarCW	
12	t rk hackerlar m s r resmi haber ajans n n sit...	AkincilarCW	
13	t rk hackerlar m s r resmi haber ajans n n sit...	AkincilarCW	
14	t rk hackerlar m s r resmi ajans n n sitesini ...	AkincilarCW	
15	t rk hackerlar m s r resmi haber ajans n n sit...	AkincilarCW	
16	t rk hacker grubu cyberwarriortim taraf ndan m...	AkincilarCW	
17	kurban bayram n n lkemize ve milletimize huzur...	AkincilarCW	
18	kendilerini k rdish hackers olarak tan mlayan ...	AkincilarCW	
19	bu lkede vatanseverler alimler hatipler as l p...	AkincilarCW	
20	haziran tarihinde irak kuzeyi kandil b lgesine...	AkincilarCW	
21	cyberwarriortim akincilar ter r rg t pkk opera...	AkincilarCW	
22	cyberwarriortim oku ren payla kitap seferberli i	AkincilarCW	
23	ba ta aziz ehitlerimizi rahmetle gazilerimizi ...	AkincilarCW	
24	ak nc lar hacker grubu dirili postas na konu t...	AkincilarCW	
25	selam naleyk m fsm hastanesinde yatan dedem i ...	AkincilarCW	
26	lg n t rkler akincilar	AkincilarCW	
27	avrupa da g ndem akincilar	AkincilarCW	
28	t rk hackerler yunan ld rtt	AkincilarCW	
29	yunanistan n g ndemi akincilar	AkincilarCW	

30	siirt eruh da icra edilen hava destekli operas...	AkincilarCW
31	alt yap al malar m z tamamlanarak web sitemize...	AkincilarCW
32	ile saatleri aras nda cyberwarriortim sunucula...	AkincilarCW
33	mesele onlar i in milli birlik ve beraberlik d...	AkincilarCW
34	erdo an bah eli destici k l daro lu karamollao...	AkincilarCW
35	temmuz ha n darbe giri imini yapan nsanlar m z...	AkincilarCW
36	temmuz hain darbe giri imini yapan insanlar m ...	AkincilarCW
37	akincilar ypg operasyonu	AkincilarCW
38	vatan sevgisini s z nde de il z nde ya ayan pl...	AkincilarCW
...	...	...
3596	here s where american tax money is being waste...	Skynet_Central
3597	expect a huge and extremly secretive and impor...	Skynet_Central
3598	knowledge is power dumping of turkish ministry...	Skynet_Central
3599	a reminder for every opresser justice is vigil...	Skynet_Central
3600	database tables of here looks like congress is...	Skynet_Central
3601	judithpoe we will inform you and the public of...	Skynet_Central
3602	marcoromagna anonzeus sorry bro zeus it was fu...	Skynet_Central
3603	penetrated we posses full database available u...	Skynet_Central
3604	marcoromagna anonzeus sorry bro zeus it was fu...	Skynet_Central
3605	judithpoe rt erdogan officialfantomn huntercal...	Skynet_Central
3606	full compressed database here skynetcentral ju...	Skynet_Central
3607	marcoromagna here is this proof enough	Skynet_Central
3608	infferna anonzeus officialfantomn realrebirth ...	Skynet_Central
3609	anyone interested in the database of skynetcen...	Skynet_Central
3610	great work done by skynet central against the ...	Skynet_Central
3611	weekend starts now see you all on monday be sa...	Skynet_Central
3612	vinceinthebay wow ure really butthurt about this	Skynet_Central
3613	vinceinthebay you will be notified when the re...	Skynet_Central
3614	vinceinthebay yep we realize it now but still ...	Skynet_Central
3615	vinceinthebay and thank you for informing us a...	Skynet_Central
3616	vinceinthebay we are sorry this website remain...	Skynet_Central
3617	vinceinthebay we will fix the bug and investig...	Skynet_Central
3618	cyb rgh s thank you	Skynet_Central
3619	skynetcentral succeeded in removing a pro isis...	Skynet_Central
3620	itweb thank you for this much appreciated and ...	Skynet_Central
3621	we deeply believe that attacking turkish gover...	Skynet_Central
3622	gazing at turkish cyber space skynetcentral rt...	Skynet_Central
3623	soon	Skynet_Central
3624	soon	Skynet_Central
3625	judithpoe rt erdogan gh sty op realrebirth res...	Skynet_Central

	Sentiment
5	1
7	1
11	0
12	0
13	0
14	0

```
15          0
16          0
17          0
18          0
19          0
20          0
21          0
22          0
23          0
24          0
25          0
26          0
27          0
28          0
29          0
30          0
31          0
32          0
33          0
34          0
35          0
36          0
37          0
38          0
...       ...
3596        0
3597        0
3598        0
3599        1
3600        0
3601        0
3602        1
3603        0
3604        0
3605        0
3606        1
3607        0
3608        0
3609        1
3610        1
3611        1
3612        0
3613        1
3614        0
3615        1
3616        0
3617        1
3618        0
```

```
3619         0
3620         1
3621         1
3622         0
3623         0
3624         0
3625         0

[8611 rows x 3 columns]
```

In [65]:
```
my_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8611 entries, 5 to 3625
Data columns (total 3 columns):
Tweets       8611 non-null object
Name         8611 non-null object
Sentiment    8609 non-null object
dtypes: object(3)
memory usage: 269.1+ KB
```

In [ ]: