

Words_count_csv

January 10, 2019

```
In [10]: import re
list=['#cat', '#dog', 'a#yacht', 'cats']
regex=re.compile(".*(#+).*")
[m.group(0) for l in list for m in [regex.search(l)] if m]
```

```
Out[10]: ['#cat', '#dog', 'a#yacht']
```

```
In [11]: # General:
import tweepy          # To consume Twitter's API
import pandas as pd     # To handle data
import numpy as np      # For number computing
import json
import string
from collections import Counter
from nltk.corpus import stopwords
import nltk
nltk.download('stopwords')
stop = stopwords.words('english')

# For plotting and visualization:
from IPython.display import display
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

```
[nltk_data] Downloading package stopwords to
```

```
[nltk_data]      /home/17170311/nltk_data...
```

```
[nltk_data]   Package stopwords is already up-to-date!
```

```
In [12]: from bs4 import BeautifulSoup
```

```
In [14]: #Here we upload the users and the extract tweets to our DataFrame
```

```
cols = ['', 'Tweets', 'Name']
```

```
df = pd.read_csv("clean_dataset_1.csv", header=None, names=cols, encoding='latin-1')
```

```
df.head(7)
```

```
Out[14]:
```

		Tweets	Name
0	NaN	Tweets	Name

		tweets	Name
1	0.0		
2	1.0	NaN	AkincilarCW
3	2.0	NaN	AkincilarCW
4	3.0	NaN	AkincilarCW
5	4.0	egypt s official news agency mena website hack...	AkincilarCW
6	5.0	NaN	AkincilarCW

In [23]: *#Removing punctuation*

```
#display(result['Tweets'])
```

```
#natural_tweets = df['Tweets'].apply(lambda x: x.lower())
natural_tweets = df['Tweets'].str.lower()
```

```
exclude = set(string.punctuation)
def remove_punctuation(x):
    """
    Helper function to remove punctuation from a string
    x: any string
    """
    try:
        x = ''.join(ch for ch in x if ch not in exclude)
    except:
        pass
    return x
```

```
#Apply the function to the DataFrame
```

```
single_words = ['rt', 'us', '1', '2', 'one', '', 'new']
```

```
fixed_tweets = natural_tweets.apply(remove_punctuation)
stop = stopwords.words('english') + stopwords.words('spanish') + single_words
#df['Tweets'] = fixed_tweets.apply(lambda x: ' '.join([word for word in x.split() if word not in stop]))
#print(fixed_tweets)
display(df['Tweets'])
```

	Tweets
0	
1	tweets
2	NaN
3	NaN
4	NaN
5	egypt s official news agency mena website hack...
6	NaN
7	middle east news agency the state news agency ...
8	NaN
9	NaN
10	NaN
11	t rk hackerlar m s r resmi haber ajans n n sit...

12 t rk hackerlar m s r resmi haber ajans n n sit...
 13 t rk hackerlar m s r resmi haber ajans n n sit...
 14 t rk hackerlar m s r resmi ajans n n sitesini ...
 15 t rk hackerlar m s r resmi haber ajans n n sit...
 16 t rk hacker grubu cyberwarriortim taraf ndan m...
 17 kurban bayram n n lkemize ve milletimize huzur...
 18 kendilerini k rdish hackers olarak tan mlayan ...
 19 bu lkede vatanseverler alimler hatipler as l p...
 20 haziran tarihinde irak kuzeyi kandil b lgesine...
 21 cyberwarriortim akincilar ter r rg t pkk opera...
 22 cyberwarriortim oku ren payla kitap seferberli i
 23 ba ta aziz ehitlerimizi rahmetle gazilerimizi ...
 24 ak nc lar hacker grubu dirili postas na konu t...
 25 selam naleyk m fsm hastanesinde yatan dedem i ...
 26 lg n t rkler akincilar
 27 avrupa da g ndem akincilar
 28 t rk hackerler yunan ld rtt
 29 yunanistan n g ndemi akincilar
 ...
 5192 islamic state warns of black days for saudi ar...
 5193 labienustitus anonopsaudix we dont forget our ...
 5194 israel jails palestinians for facebook comments
 5195 rohingya muslims found on boat rescued by myan...
 5196 rohingya muslims found on boat rescued by myan...
 5197 mass graves found in malaysia close to detenti...
 5198 irfan death of humanity burma says muslims not...
 5199 while syrian rebels fight hezbollah assad in e...
 5200 just out of curiosity have hezbollah ever even...
 5201 richard clarke s on iraq we destroyed iraq dae...
 5202 day you know as whole the world is albeit slow...
 5203 fox news cnn on the right wing deceivers and p...
 5204 presstv iranian government influenced media ru...
 5205 saudi regime involved in attack on qatif mosqu...
 5206 saudi regime involved in attack on qatif mosqu...
 5207 these politicians all the same everything when...
 5208 down g n ration identitaire anti muslim hate g...
 5209 breaking the senate has just passed its tpp fa...
 5210 put israel on theunlist of serious rights viol...
 5211 walts insight thank you for the follow
 5212 opdarrienhunt anonrastaftp ted bingham susaa m...
 5213 a malcolmite the bombing of a mosque will gene...
 5214 america uk are protesting government corruptio...
 5215 they don t want to take any responsibility for...
 5216 then the nd gulf war happened and al qaida in ...
 5217 al qaida was created after the st gulf war in ...
 5218 the media who care a lot more about ratings an...
 5219 us and nato occupations interventions led to t...
 5220 the video next is tough to watch but it s a mu...

```
5221    months before isis was even formed reports of ...  
Name: Tweets, Length: 5222, dtype: object
```