Cleaning_Data_Programme

December 7, 2018

1 Cleaning the Tweets

```
In [2]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        from bs4 import BeautifulSoup
        import re
```

```
In [3]: #Here we upload the users and the extract tweets to our DataFrame
        cols = ['','Tweets','Name','Lenght','ID','Date','Source','Likes','RTs']
        df = pd.read_excel("dataset 1.xlsx",header=None, names=cols, encoding='latin-1')
        df.head()
```

```
Out[3]:                                                    Tweets         Name  \
        0  NaN                                             Tweets         Name
        1  0.0  RT @ajmhashtag:      ...  AkincilarCW
        2  1.0  RT @BBCArabic:      ...  AkincilarCW
        3  2.0  RT @Elsanhory:      ...  AkincilarCW
        4  3.0  RT @BBCMonitoring: Egypt's official news agenc...  AkincilarCW

           Lenght                   ID                 Date              Source  \
        0  Length                   ID                 Date              Source
        1      92  1039757171754446976  2018-09-12 06:06:42  Twitter Web Client
        2     139  1039757123905888000  2018-09-12 06:06:30  Twitter Web Client
        3      75  1039757024937090944  2018-09-12 06:06:07  Twitter Web Client
        4     140  1039756976249556992  2018-09-12 06:05:55  Twitter Web Client

           Likes  RTs
        0  Likes  RTs
        1      0    8
        2      0   19
        3      0    6
        4      0   12
```

```
In [4]: # Drop the columns that we don't need
        df.drop(['','Lenght','ID','Date','Source','Likes','RTs'],axis=1,inplace=True)
        df[df.Tweets == 0].head(10)
```

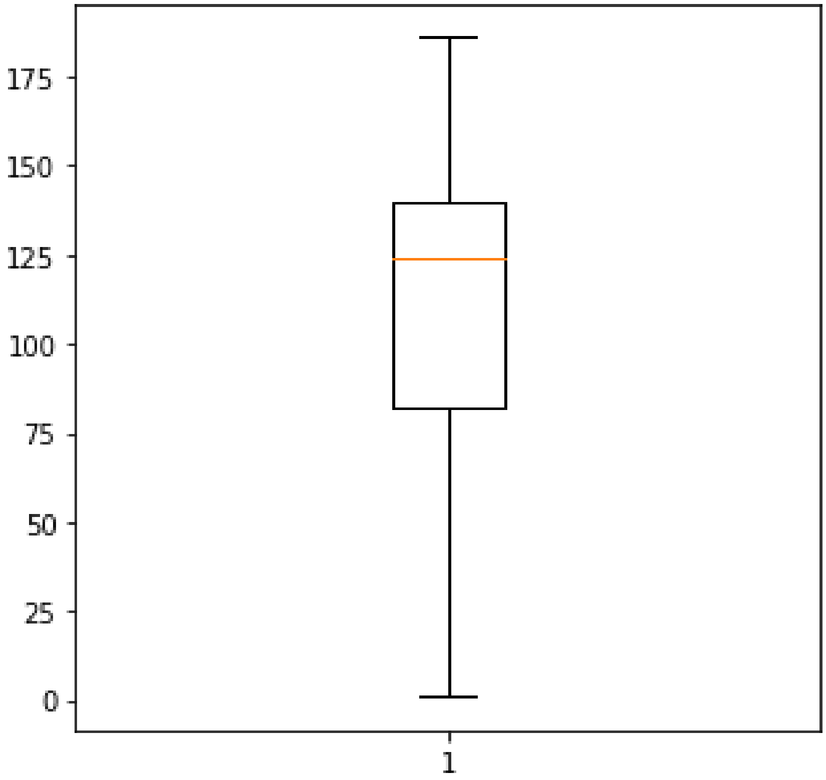
```
Out[4]: Empty DataFrame
        Columns: [Tweets, Name]
        Index: []
```

```
In [5]: #Lets look at the lenght in text column for each entry
        df['pre_clean_len'] = [len(t) for t in df.Tweets]
```

```
In [6]: #First draft of data dictionary
        from pprint import pprint
        data_dict = {
            #'sentiment':{
               # 'type':df.Sentiment.dtype,
               # 'description':'sentiment class - 0:negative, 1:positive'
           # },
            'text':{
                'type':df.Tweets.dtype,
                'description':'tweet text'
            },
            'pre_clean_len':{
                'type':df.pre_clean_len.dtype,
                'description':'Length of the tweet before cleaning'
            },
            'dataset_shape':df.shape
        }
        pprint(data_dict)
```

```
{'dataset_shape': (5223, 3),
 'pre_clean_len': {'description': 'Length of the tweet before cleaning',
                   'type': dtype('int64')},
 'text': {'description': 'tweet text', 'type': dtype('O')}}
```

```
In [7]: #I can see the overall distribution of length of strings in each entry.
        fig, ax = plt.subplots(figsize=(5, 5))
        plt.boxplot(df.pre_clean_len)
        plt.show()
```



```
In [8]: #Here its all the rules to clean the tweets from undesired characters
        from nltk.tokenize import WordPunctTokenizer
        tok = WordPunctTokenizer()
        pat1 = r'RT @[A-Za-z0-9]+'
        pat2 = r'https?://[A-Za-z0-9./]+'
        combined_pat = r'|'.join((pat1, pat2))
        def tweet_cleaner(text):
            soup = BeautifulSoup(text, 'lxml')
            souped = soup.get_text()
            stripped = re.sub(combined_pat, '', souped)
            try:
                clean = stripped.decode("utf-8-sig").replace(u"\ufffd", "?")
            except:
                clean = stripped
            letters_only = re.sub("[^a-zA-Z]", " ", clean)
            lower_case = letters_only.lower()
            # During the letters_only process two lines above, it has created unnecessay white
            # I will tokenize and join together to remove unneccessary white spaces
            words = tok.tokenize(lower_case)
            return (" ".join(words)).strip()
        testing = df.Tweets[:100]
        test_result = []
```

```
         for t in testing:
             test_result.append(tweet_cleaner(t))
         test_result
```

```
Out[8]: ['tweets',
          '',
          '',
          '',
          'egypt s official news agency mena website hacked with message of support for impriso
          '',
          'middle east news agency the state news agency in egypt was hacked by who seems to be
          '',
          '',
          '',
          't rk hackerlar m s r resmi haber ajans n n sitesini hackledi son dakika cnnturk',
          't rk hackerlar m s r resmi haber ajans n n sitesini hackledi haberturk',
          't rk hackerlar m s r resmi haber ajans n n sitesini hackledi',
          't rk hackerlar m s r resmi ajans n n sitesini hackledi',
          't rk hackerlar m s r resmi haber ajans n n sitesini hackledi',
          't rk hacker grubu cyberwarriortim taraf ndan m s r haber ajans web sitesine akincila
          'kurban bayram n n lkemize ve milletimize huzur bar ve mutluluklar getirmesini diler
          'kendilerini k rdish hackers olarak tan mlayan ve t rkiye geneline leak edilmi telefo
          'bu lkede vatanseverler alimler hatipler as l p bitene kadar idam vard ne zaman ki s
          'haziran tarihinde irak kuzeyi kandil b lgesine yap lan hava harek t nda b l c ter r
          'cyberwarriortim akincilar ter r rg t pkk operasyonu intro',
          'cyberwarriortim oku ren payla kitap seferberli i',
          'ba ta aziz ehitlerimizi rahmetle gazilerimizi minnetle an yor onlar n de erli ailele
          'ak nc lar hacker grubu dirili postas na konu tu yunanistan n b t n resmi kurumlar a
          'selam naleyk m fsm hastanesinde yatan dedem i in a negatif kana ihtiya vard r kan gr
          'lg n t rkler akincilar',
          'avrupa da g ndem akincilar',
          't rk hackerler yunan ld rtt',
          'yunanistan n g ndemi akincilar',
          'siirt eruh da icra edilen hava destekli operasyonlarda gri listede yer alan adil k m
          'alt yap al malar m z tamamlanarak web sitemize ula m sa lanmaktad r',
          'ile saatleri aras nda cyberwarriortim sunucular nda alt yap al malar olacakt r t m c
          'mesele onlar i in milli birlik ve beraberlik de il erdo an indirme meselesidir',
          'erdo an bah eli destici k l daro lu karamollao lu ak ener demirta y ksekda be ar esa
          'temmuz ha n darbe giri imini yapan nsanlar m z eh t eden ha n fet ter r rg t yelerin
          'temmuz hain darbe giri imini yapan insanlar m z eh t eden ha n fet ter r rg t yeleri
          'akincilar ypg operasyonu',
          'vatan sevgisini s z nde de il z nde ya ayan platform cyber warrior ilminize kuvvet a
          'rt kirmizicehreli ypg resmi websitesi a cyberwarriortim akincilar taraf ndan bayrak
          'ter r rg t ypg operasyonu',
          'toyota turkiye orkun tmak denen erefsize verdi iniz deste i saklam yorsunuz biz de t
          'toyota turkiye hey toyotamotorcorp unfortunately toyota turkey is using a convicted
          'ypg pkk n n kolu olan ypgrojava nin resm websitesi akincilar taraf ndan m h rlenmi t
          't rkiye nin ilk ve tek siber savunma g c cyber warrior akincilar grubu taraf ndan yp
```

```
'afrin ehitleri an s na',
'ayt sahtekar toplulu undan cyber warrior vatan topraklar na gelen onlarca kardesim va
'y llard r ayt nin yedi i haltlar anlat yoruz kan tlarla sunuyoruz k r k r ne inanan
'ha l zihniyetli yunanistan milletvekili niki founta ya cyberwarriortim akincilar ders
'cuman z mubarek olsun cyberwarriortim',
'aptal g r yorsunuz anlatmaya gerek yok',
'sayg yla an yoruz',
'nam m z kendi an m zdan de il cyber warrior un an ndand r',
'allah bu millete bir daha stiklal mar yazd rmas n mehmetakifersoy',
've bir ses y kseldi veda hutbesinden kad nlar size allah n emanetidir hz muhammed s a
'cyber warrior tim akincilar hollanda seferi intro t rk hacker grubu cyberwarriortim a
'cyber warrior tim sefer bizim zafer allah nd r cyberwarriortim',
'afrin zeytindaliharekati nda ter r rg tlerine kar y r t len operasyonlar esnas nda bi
'ac l bah elievler medical parkta yatmakta olan hastam z i in acil rh negatif kana iht
'vur anl silah nla g n l m lk d zelsin sen l yorken de vururken de g zelsin',
'biz s radan bir site de iliz bizimkisi ayn dava da hizmet',
'biz bu topraklara k k sald k bizden olmayana ektirmeyiz abdulhamid han ubat vefat n
'allah bize yeter',
'allah yard mc n z olsun yi itler y re imiz sizinle cyberwarriortim not cepheden garaj
'st m zden eksilmesin albayra n g lgesi',
'vatan n sanal kalesi cyber warrior tim akincilarafrinde',
'ok k ymetlidir benim ailem sanalda de il g n lde t rkiyenin sessiz kahramanlar bili
'ocak faili mechul cinayete kurban giden ehit ali gaffar okkan unutulmayacaks n aliga
'l g libe llallah millimesele',
'man edenler allah yolunda cenk ederler nkar edenler ise t ut allah n yerine tuttukla
'zeytindal harekat',
'u kopan f rt na t rk ordusudur y rabbi senin u runda len ordu budur y rabbi t ki y ks
'yi itlerimizin gazas m barek olsun rabbim zaferler nasip etsin in allah afrinoperasy
'afrinoperasyonu bismillahirrahmanirrahim',
'medya chp stanbul l ba kan canan kaftanc o lu nun ay nce tkp nin bulu mas na kat ld
'cw samil asim xkatakulli salihzekierdem evetocyber turkcubeg salihzekierdem ne zaman
'cyberwarriortim new lockpos malware analiz',
'cyberwarriortim sunucular m zda yap lan internet alt yap s al malar ndan dolay web s
'alemde er cw de er t kenmez',
't rk hackerlardan srail e ok cyberwarriortim',
't rk hackerlar srail e ait g venlik kameralar sistemini hackledi cyberwarriortim',
'cyber warrior ak nc lar hacker grubu srail e ait g venlik kameralar sistemini hackle
'cyber warrior ak nc lar hacker grubu srail e ait g venlik kameralar sistemini hackle
'ak nc lar cyberwarriortim',
'srail efa amr belediyesi t rk hacker grubu cyber warrior tim akincilar taraf ndan ha
'inject r isimtescilnet inanmam ciddi misin',
'cwf r verdi in adrese kargolad m',
'cwf r adresini yolla kargolayay m kardesim',
'sercanarga ertesi g n f ak nc lar indexi',
't rkiyekud s inayakta abd ve srail siteleri a k hedefimizdir gazam z mubarek olsun',
'a haber i te o yalanlar de ifre ediyor ahabersusturulamaz memlktmeselesi kerimulak h
'metasploitte microsoft office dde g venlik a n n kullan m',
'microsoft office exploit kali linux lojistik grup',
'blueborne kritik bluetooth sald r s cyberwarriortim',
'haydar a milli tak m kadrosuna koy s r tmaz',
'abd lhamidhan tv lerde de il okuyarak renin nerim heyet',
'cuman n nemini bilmeyen zat k ymetini nereden bilsin blackfriday karacuma de il muba
'dualar m zdas n retmenim dogukanilbeyi ruhun ad mekan n cennet olsun',
'ba ta cyberwarriortim b nyesinde bulunan retmenlerimiz olmak zere t m retmenlerimizi
'ufukcc tarikkbjk gol yolun yar s eder',
'tarikkbjk ufukcc negredo i buldu herhalde']
```

```
In [9]: nums = [0,5221]
        #nums = [5222,8845]
```

```
In [10]: print ("Cleaning and parsing the tweets...\n")
         clean_tweet_texts = []
         for i in range(nums[0],nums[1]):
             if( (i+1)%10000 == 0 ):
                 print ("Tweets %d of %d has been processed" % ( i+1, nums[1] ))
             clean_tweet_texts.append(tweet_cleaner(df['Tweets'][i]))
```

```
Cleaning and parsing the tweets...



/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
```

```
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/

https://t.co/ckBBV4KPAh" looks like a URL. Beautiful Soup is not an HTTP client. You should pro
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
/opt/jupyterhub/anaconda/lib/python3.6/site-packages/bs4/__init__.py:336: UserWarning: "https:/
  ' that document to Beautiful Soup.' % decoded_markup
```

```
In [11]: len(clean_tweet_texts)

Out[11]: 5221
```

2 Saving cleaned data in CSV

```
In [12]: clean_df = pd.DataFrame(clean_tweet_texts,columns=['Tweets'])
         clean_df['Name'] = df.Name
         clean_df.head()

Out[12]:                                                 Tweets         Name
         0                                               tweets         Name
         1                                                       AkincilarCW
         2                                                       AkincilarCW
         3                                                       AkincilarCW
         4  egypt s official news agency mena website hack...  AkincilarCW
```

```
In [21]: clean_df.to_csv('clean_dataset_1.csv',encoding='utf-8')
         csv = 'clean_dataset_1.csv'
         my_df = pd.read_csv(csv,index_col=0)
         my_df.head()

Out[21]:                                                 Tweets         Name
         0                                               tweets         Name
         1                                                  NaN  AkincilarCW
         2                                                  NaN  AkincilarCW
         3                                                  NaN  AkincilarCW
         4  egypt s official news agency mena website hack...  AkincilarCW
```

3 Drop Missing Values

```
In [31]: #dealing with missing data
         sufficientcolumns = my_df.isnull().sum() < 2
         df = my_df.loc[:, sufficientcolumns] # keep only columns that have less than 2 missin
         df = my_df.drop(my_df.loc[my_df['Tweets'].isnull()].index) # remove the sample that h
```

```
In [33]: df.head()

Out[33]:                                                  Tweets         Name
         0                                                tweets         Name
         4   egypt s official news agency mena website hack...  AkincilarCW
         6   middle east news agency the state news agency ...  AkincilarCW
         10  t rk hackerlar m s r resmi haber ajans n n sit...  AkincilarCW
         11  t rk hackerlar m s r resmi haber ajans n n sit...  AkincilarCW
```

```
In [34]: df.to_csv('clean_dataset_1.csv',encoding='utf-8')
         csv = 'clean_dataset_1.csv'
         my_df = pd.read_csv(csv,index_col=0)
         my_df.head()
```

```
Out[34]:                                                         Tweets         Name
         0                                                       tweets         Name
         4   egypt s official news agency mena website hack...  AkincilarCW
         6   middle east news agency the state news agency ...  AkincilarCW
         10  t rk hackerlar m s r resmi haber ajans n n sit...  AkincilarCW
         11  t rk hackerlar m s r resmi haber ajans n n sit...  AkincilarCW
```

```
In [ ]:
```