Sentiment Analysis

In [1]:
```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file
train = pd.read_csv('train.csv')
test = pd.read_csv('test.csv')
```

In [2]:
```
train.head()
```

Out[2]:

	id	label	tweet
0	1	0	@user when a father is dysfunctional and is s...
1	2	0	@user @user thanks for #lyft credit i can't us...
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in ...
4	5	0	factsguide: society now #motivation

In [3]:
```
test.head()
```

Out[3]:

	id	tweet
0	31963	#studiolife #aislife #requires #passion #dedic...
1	31964	@user #white #supremacists want everyone to s...
2	31965	safe ways to heal your #acne!! #altwaystohe...
3	31966	is the hp and the cursed child book up for res...
4	31967	3rd #bihday to my amazing, hilarious #nephew...

Counting the number of words in each tweets

In [4]:
```
def num_of_words(df):
    df['word_count'] = df['tweet'].apply(lambda x : len(str(x).split(" ")))
    print(df[['tweet','word_count']].head())
```

In [5]:
```
num_of_words(train)
num_of_words(test)
```

```
                                                tweet  word_count
0   @user when a father is dysfunctional and is s...          21
1  @user @user thanks for #lyft credit i can't us...          22
2                                bihday your majesty           5
3  #model   i love u take with u all the time in ...          17
4             factsguide: society now    #motivation           8
                                                tweet  word_count
0  #studiolife #aislife #requires #passion #dedic...          12
1   @user #white #supremacists want everyone to s...          20
2  safe ways to heal your #acne!!    #altwaystohe...          15
3  is the hp and the cursed child book up for res...          24
4    3rd #bihday to my amazing, hilarious #nephew...          18
```

Counting and Removing the Stop Words

Stop Words : A stop word is a commonly used word such as the, a, an, in which are filtered out before or after processing of natural language data (text).

In [38]:
```
import nltk
#the below command use it once and download the packages
#nltk.download()
# a window will open in the select popular under collections
from nltk.corpus import stopwords
stop = stopwords.words('english')
```

In [39]:
```
def stop_words(df):
    df['stopwords'] = df['tweet'].apply(lambda x: len([x for x in x.split(
    print(df[['tweet','stopwords']].head())
```

In [40]:
```
stop_words(train)
stop_words(test)
```

```
                                               tweet  stopwords
0  user father dysfunctional selfish drags kids d...          0
1  user user thanks lyft credit cant use cause do...          0
2                                  bihday majesty           0
3              model love u take u time urð ðððð ððð          0
4                     factsguide society motivation           0
                                               tweet  stopwords
0  studiolife aislife requires passion dedication...          0
1  user white supremacists want everyone see new ...          0
2  safe ways heal acne altwaystoheal healthy healing          0
3  hp cursed child book reservations already yes ...          3
4  3rd bihday amazing hilarious nephew eli ahmir ...          0
```

In [41]:
```
def stop_words_removal(df):
    df['tweet'] = df['tweet'].apply(lambda x: " ".join(x for x in x.split(
    print(df['tweet'].head())
```

In [42]:
```
stop_words_removal(train)
stop_words_removal(test)
```

```
0    user father dysfunctional selfish drags kids d...
1    user user thanks lyft credit cant use cause do...
2                                   bihday majesty
3                model love u take u time urð ðððð ððð
4                       factsguide society motivation
Name: tweet, dtype: object
0    studiolife aislife requires passion dedication...
1    user white supremacists want everyone see new ...
2    safe ways heal acne altwaystoheal healthy healing
3    hp cursed child book reservations already yes ...
4    3rd bihday amazing hilarious nephew eli ahmir ...
Name: tweet, dtype: object
```

Converting tweets to lowercase letter

In [43]:
```
def lower_case(df):
    df['tweet'] = df['tweet'].apply(lambda x: " ".join(x.lower() for x in x
    print(df['tweet'].head())
```

In [44]:
```
lower_case(train)
lower_case(test)
```

```
0    user father dysfunctional selfish drags kids d...
1    user user thanks lyft credit cant use cause do...
2                                   bihday majesty
3                model love u take u time urð ðððð ððð
4                       factsguide society motivation
Name: tweet, dtype: object
0    studiolife aislife requires passion dedication...
1    user white supremacists want everyone see new ...
2    safe ways heal acne altwaystoheal healthy healing
3    hp cursed child book reservations already yes ...
4    3rd bihday amazing hilarious nephew eli ahmir ...
Name: tweet, dtype: object
```

Removing special Characters and Punctuation

In [45]:
```
def punctuation_removal(df):
    df['tweet'] = df['tweet'].str.replace('[^\w\s]','')
    print(df['tweet'].head())
```

In [46]:
```
punctuation_removal(train)
punctuation_removal(test)
```

```
0    user father dysfunctional selfish drags kids d...
1    user user thanks lyft credit cant use cause do...
2                                   bihday majesty
3                model love u take u time urð ðððð ððð
4                       factsguide society motivation
Name: tweet, dtype: object
0    studiolife aislife requires passion dedication...
1    user white supremacists want everyone see new ...
2    safe ways heal acne altwaystoheal healthy healing
3    hp cursed child book reservations already yes ...
4    3rd bihday amazing hilarious nephew eli ahmir ...
Name: tweet, dtype: object
```

Remove the most frequently used words and less frequently used words

In [47]:
```
freq = pd.Series(' '.join(train['tweet']).split()).value_counts()[:10]
freq
```

Out[47]:
```
user     17473
love      2647
ð         2511
day       2199
â         1797
happy     1663
amp       1582
im        1139
u         1136
time      1110
dtype: int64
```

In [48]:
```
freq = list(freq.index)
def frequent_words_removal(df):
    df['tweet'] = df['tweet'].apply(lambda x: " ".join(x for x in x.split(
    print(df['tweet'].head())
```

In [49]:
```
frequent_words_removal(train)
frequent_words_removal(test)
```

```
0    father dysfunctional selfish drags kids dysfun...
1    thanks lyft credit cant use cause dont offer w...
2                                      bihday majesty
3                               model take urð ðððð ðð
4                         factsguide society motivation
Name: tweet, dtype: object
0    studiolife aislife requires passion dedication...
1    white supremacists want everyone see new birds...
2    safe ways heal acne altwaystoheal healthy healing
3    hp cursed child book reservations already yes ...
4    3rd bihday amazing hilarious nephew eli ahmir ...
Name: tweet, dtype: object
```

In [50]:
```
freq = pd.Series(' '.join(train['tweet']).split()).value_counts()[-10:]
freq
```

Out[50]:
```
550                1
flyfishingnation   1
vegetablegarden    1
bigbizgtus         1
flowððfriends      1
saban              1
teenageson         1
overflowingjoy     1
kesãkurpitsa       1
koalamuffins       1
dtype: int64
```

In [51]:
```
freq = list(freq.index)
def rare_words_removal(df):
    df['tweet'] = df['tweet'].apply(lambda x: " ".join(x for x in x.split(
    print(df['tweet'].head())
```

In [53]:
```
rare_words_removal(train)
rare_words_removal(test)
```

```
0    father dysfunctional selfish drags kids dysfun...
1    thanks lyft credit cant use cause dont offer w...
2                                    bihday majesty
3                              model take urð ðððð ððð
4                         factsguide society motivation
Name: tweet, dtype: object
0    studiolife aislife requires passion dedication...
1    white supremacists want everyone see new birds...
2    safe ways heal acne altwaystoheal healthy healing
3    hp cursed child book reservations already yes ...
4    3rd bihday amazing hilarious nephew eli ahmir ...
Name: tweet, dtype: object
```

Spelling Correction

In [65]:
```
pip install textblob
```

```
Collecting textblob
  Downloading textblob-0.15.3-py2.py3-none-any.whl (636 kB)
     |████████████████████████████████| 636 kB 2.5 MB/s eta 0:00:01
Requirement already satisfied: nltk>=3.1 in /opt/anaconda3/lib/python3.8/si
te-packages (from textblob) (3.5)
Requirement already satisfied: click in /opt/anaconda3/lib/python3.8/site-p
ackages (from nltk>=3.1->textblob) (7.1.2)
Requirement already satisfied: regex in /opt/anaconda3/lib/python3.8/site-p
ackages (from nltk>=3.1->textblob) (2020.10.15)
Requirement already satisfied: joblib in /opt/anaconda3/lib/python3.8/site-
packages (from nltk>=3.1->textblob) (0.17.0)
Requirement already satisfied: tqdm in /opt/anaconda3/lib/python3.8/site-pa
ckages (from nltk>=3.1->textblob) (4.50.2)
Installing collected packages: textblob
Successfully installed textblob-0.15.3
Note: you may need to restart the kernel to use updated packages.
```

In [69]:
```
from textblob import TextBlob
```

In [70]:
```
def spell_correction(df):
    return df['tweet'][:5].apply(lambda x: str(TextBlob(x).correct()))
```

In [71]:
```
spell_correction(train)
```

Out[71]:
```
0    father dysfunctional selfish drags kiss dysfun...
1    thanks left credit can use cause dont offer wh...
2                                    midday majesty
3                              model take or ðððð ððð
4                         factsguide society motivation
Name: tweet, dtype: object
```

In [72]:
```
spell_correction(test)
```

Out[72]:
```
0    studiolife dislike requires passion education ...
1    white supremacists want everyone see new birds...
2    safe ways heal acne altwaystoheal healthy healing
3    he cursed child book reservations already yes ...
4    rd midday amazing hilarious nephew epi their u...
Name: tweet, dtype: object
```

Tokenizing

Tokenization refers to dividing the text into a sequence of words or sentences.

In [73]:
```
def tokens(df):
    return TextBlob(df['tweet'][1]).words
```

In [74]:
```
tokens(train)
```

Out[74]:
```
WordList(['thanks', 'lyft', 'credit', 'cant', 'use', 'cause', 'dont', 'offe
r', 'wheelchair', 'vans', 'pdx', 'disapointed', 'getthanked'])
```

In [75]:
```
tokens(test)
```

Out[75]:
```
WordList(['white', 'supremacists', 'want', 'everyone', 'see', 'new', 'birds
â', 'movie', 'hereâs'])
```

Stemming

Stemming refers to the removal of suffices, like “ing”, “ly”, “s”, etc. by a simple rule-based approach.

In [76]:
```
from nltk.stem import PorterStemmer
st = PorterStemmer()
```

In [77]:
```
def stemming(df):
    return df['tweet'][:5].apply(lambda x: " ".join([st.stem(word) for wor
```

In [78]:
```
stemming(train)
```

Out[78]:
```
0        father dysfunct selfish drag kid dysfunct run
1    thank lyft credit cant use caus dont offer whe...
2                                      bihday majesti
3                              model take urð ðððð ðð
4                              factsguid societi motiv
Name: tweet, dtype: object
```

In [79]:
```
stemming(test)
```

Out[79]:
```
0    studiolif aislif requir passion dedic willpow ...
1    white supremacist want everyon see new birdsâ ...
2            safe way heal acn altwaystoh healthi heal
3    hp curs child book reserv alreadi ye ðð harry...
4    3rd bihday amaz hilari nephew eli ahmir uncl d...
Name: tweet, dtype: object
```

Applying Term Frequency – Inverse Document Frequency (TF-IDF

In [81]:
```
tf1 = (train['tweet'][1:2]).apply(lambda x: pd.value_counts(x.split(" ")))
tf1.columns = ['words','tf']
```

In [82]:
```
for i,word in enumerate(tf1['words']):
    tf1.loc[i, 'idf'] = np.log(train.shape[0]/(len(train[train['tweet'].st
```

In [83]:
```
tf1['tfidf'] = tf1['tf'] * tf1['idf']
tf1
```

Out[83]:

	words	tf	idf	tfidf
0	disapointed	1	10.372303	10.372303
1	dont	1	3.745585	3.745585
2	pdx	1	8.762865	8.762865
3	use	1	3.542509	3.542509
4	credit	1	7.327781	7.327781
5	getthanked	1	9.679156	9.679156
6	cause	1	5.690172	5.690172
7	cant	1	3.538194	3.538194
8	vans	1	8.426393	8.426393
9	thanks	1	4.597751	4.597751
10	offer	1	6.522155	6.522155
11	lyft	1	8.762865	8.762865
12	wheelchair	1	9.273691	9.273691

Sentiment Analysis

In [84]:
```
def polarity_subjectivity(df):
    return df['tweet'][:5].apply(lambda x: TextBlob(x).sentiment)
```

In [85]:
```
polarity_subjectivity(train)
```

Out[85]:
```
0    (-0.5, 1.0)
1     (0.2, 0.2)
2     (0.0, 0.0)
3     (0.0, 0.0)
4     (0.0, 0.0)
Name: tweet, dtype: object
```

In [86]:
```
polarity_subjectivity(test)
```

Out[86]:
```
0                                      (0.0, 0.0)
1    (0.06818181818181818, 0.22727272727272727)
2                                      (0.5, 0.5)
3                                      (0.5, 1.0)
4      (0.3666666666666667, 0.6333333333333333)
Name: tweet, dtype: object
```

In [87]:
```
def sentiment_analysis(df):
    df['sentiment'] = df['tweet'].apply(lambda x: TextBlob(x).sentiment[0]
    return df[['tweet','sentiment']].head()
```

In [88]:
```
sentiment_analysis(train)
```

Out[88]:

	tweet	sentiment
0	father dysfunctional selfish drags kids dysfun...	-0.5
1	thanks lyft credit cant use cause dont offer w...	0.2
2	bihday majesty	0.0
3	model take urð ðððð ððð	0.0
4	factsguide society motivation	0.0

In [89]:
```
sentiment_analysis(test)
```

Out[89]:

	tweet	sentiment
0	studiolife aislife requires passion dedication...	0.000000
1	white supremacists want everyone see new birds...	0.068182
2	safe ways heal acne altwaystoheal healthy healing	0.500000
3	hp cursed child book reservations already yes ...	0.500000
4	3rd bihday amazing hilarious nephew eli ahmir ...	0.366667