```
import numpy as np
import pandas as pd
```

```
data = pd.read_csv("/content/training.1600000.processed.noemoticon.csv.zip",encoding='latin-1')
data.head()
```

	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ;D
0	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...
1	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...
2	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
3	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all....
4	0	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf	@Kwesidei not the whole crew

```
DATASET_COLUMNS = ["target", "ids", "date", "flag", "user", "TweetText"]
data.columns = DATASET_COLUMNS
```

```
data.describe(include='all')
```

```
/usr/local/lib/python3.10/dist-packages/google/colab/_dataframe_summarizer.py:88: FutureWarning: Parsed string "Mon Jun 15 12:53:14
  cast_date_col = pd.to_datetime(column, errors="coerce")
```

	target	ids	date	flag	user	TweetText
count	1.599999e+06	1.599999e+06	1599999	1599999	1599999	1599999
unique	NaN	NaN	774362	1	659775	1581465
top	NaN	NaN	Mon Jun 15 12:53:14 PDT 2009	NO_QUERY	lost_dog	isPlayer Has Died! Sorry
freq	NaN	NaN	20	1599999	549	210
mean	2.000001e+00	1.998818e+09	NaN	NaN	NaN	NaN
std	2.000001e+00	1.935757e+08	NaN	NaN	NaN	NaN
min	0.000000e+00	1.467811e+09	NaN	NaN	NaN	NaN
25%	0.000000e+00	1.956916e+09	NaN	NaN	NaN	NaN
50%	4.000000e+00	2.002102e+09	NaN	NaN	NaN	NaN
75%	4.000000e+00	2.177059e+09	NaN	NaN	NaN	NaN
max	4.000000e+00	2.329206e+09	NaN	NaN	NaN	NaN

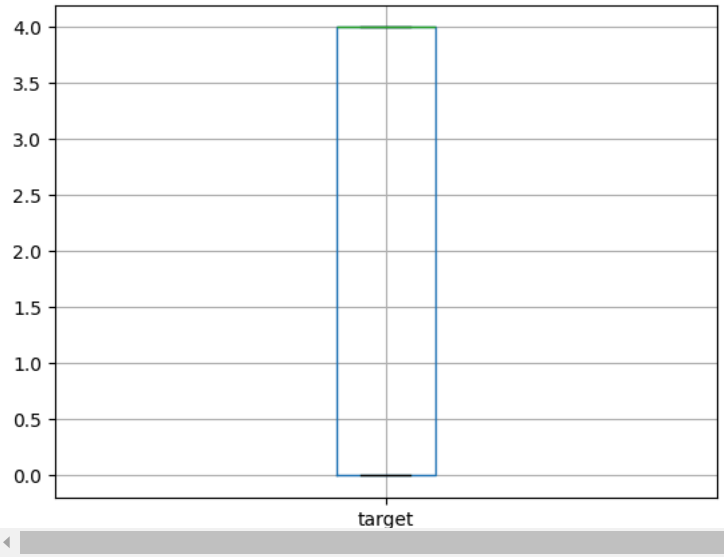
```
data.dtypes
```

	0
target	int64
ids	int64
date	object
flag	object
user	object
TweetText	object

```
import copy
data_ = copy.deepcopy(data)

positif_data = data_[data_.target==4].iloc[:80000,:]
negative_data = data_[data_.target==0].iloc[:80000,:]

sub_data = pd.concat([positif_data,negative_data],axis=0)
```

```
data.boxplot(column='target')
```

```
<Axes: >
```




```
data_target=data.groupby('target')
```

```
data['target'].value_counts()
```

target	count
4	800000
0	799999

```
data.head()
```

	target	ids	date	flag	user	TweetText
0	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...
1	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...
2	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
3	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all....
4	0	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf	@Kwesidei not the whole crew

```
data_ = {'target': data['target'], 'date': data['date']}
df = pd.DataFrame(data_)
df.head()
```

	target	date
0	0	Mon Apr 06 22:19:49 PDT 2009
1	0	Mon Apr 06 22:19:53 PDT 2009
2	0	Mon Apr 06 22:19:57 PDT 2009
3	0	Mon Apr 06 22:19:57 PDT 2009
4	0	Mon Apr 06 22:20:00 PDT 2009

```
df['date'] = pd.to_datetime(df['date'])
```

```
<ipython-input-14-e8d2d516eb0e>:1: FutureWarning: Parsed string "Mon Apr 06 22:19:49 PDT 2009" included an un-recognized timezone "P
  df['date'] = pd.to_datetime(df['date'])
```

```
hour = [ df['date'][i].hour for i in range(len(df['date'])) ]
df['hour'] = hour
df.head()
```

	target	date	hour
0	0	2009-04-06 22:19:49	22
1	0	2009-04-06 22:19:53	22
2	0	2009-04-06 22:19:57	22
3	0	2009-04-06 22:19:57	22
4	0	2009-04-06 22:20:00	22

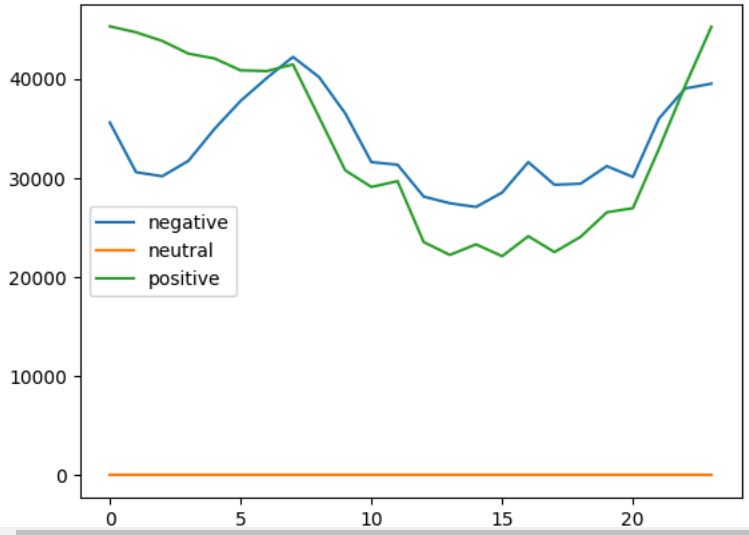
```
hour_data = {'0': [0]*24, '2': [0]*24, '4': [0]*24}
for i in range(len(df['hour'])):
    target = str(df['target'][i])
    hour = int(df['hour'][i])
    hour_data[target][hour] += 1
```

```
hour_data = [hour_data['0'], hour_data['2'], hour_data['4']]
# Transpose
hour_data = list(map(list,zip(*hour_data)))
```

```
df1 = pd.DataFrame(hour_data,index = [i for i in range(24)],columns=['negative', 'neutral', 'positive'])
```

```
df1.plot()
```

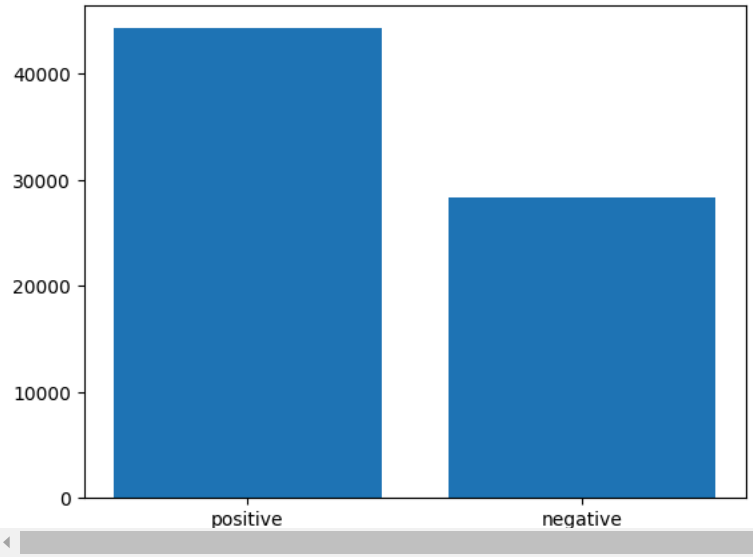
```
<Axes: >
```



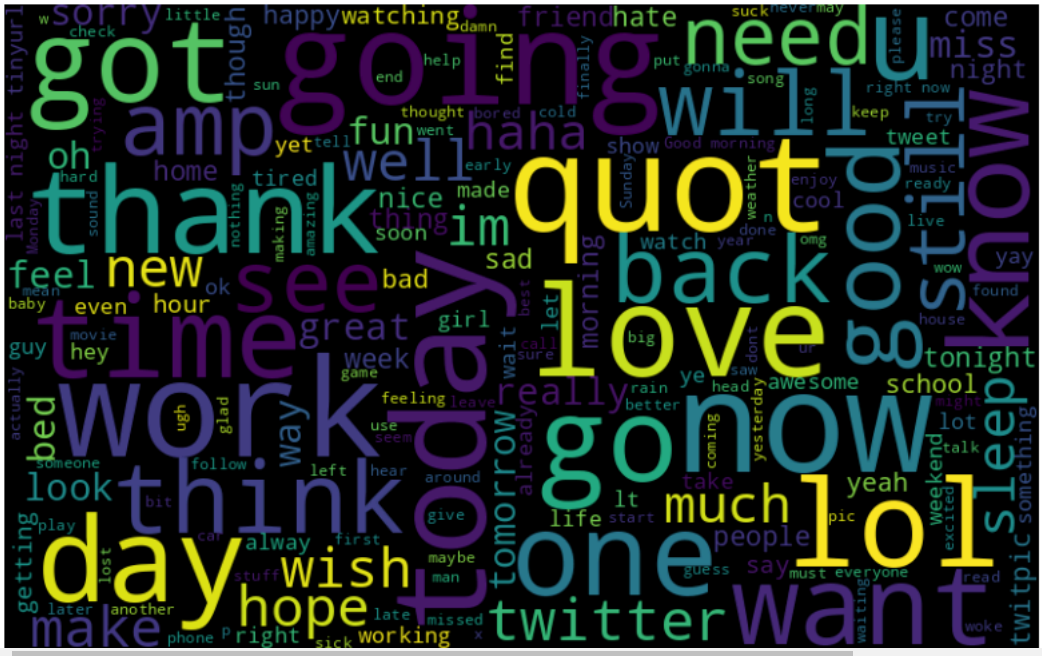

```
positive_at_count = 0
negative_at_count = 0
TweetTextList = list(sub_data['TweetText'])
targetList = list(sub_data['target'])
for i in range(len(sub_data['TweetText'])):
    if TweetTextList[i].find('@') != -1:
        if targetList[i] == 4:
            positive_at_count += 1
        else:
            negative_at_count += 1
at_counts = [positive_at_count, negative_at_count]
```

```
import matplotlib.pyplot as plt
names = ['positive', 'negative']
values = [positive_at_count, negative_at_count]
plt.bar(names, values)
```

```
<BarContainer object of 2 artists>
```



```
import copy
newdata = copy.deepcopy(sub_data)
newdata.drop(['ids','date','flag','user'],axis = 1,inplace = True)


import wordcloud
from wordcloud import WordCloud
import matplotlib.pyplot as plt
%matplotlib inline

import re
import nltk
import string
import warnings

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import os



all_words = ' '.join([text for text in newdata['TweetText']])
wordcloud = WordCloud(width=800,height=500,random_state=21,max_font_size=110).generate(all_words)
plt.figure(figsize=(10,7))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.show()
```



```
positive_words = ' '.join([text for text in data['TweetText'][data['target']==4]])
wordcloud = WordCloud(width=800,height=500,random_state=21,max_font_size=110).generate(positive_words)
plt.figure(figsize=(10,7))
plt.imshow(wordcloud,interpolation="bilinear")
plt.axis('off')
plt.show()
```

```
negative_words = ' '.join([text for text in data['TweetText'][data['target']==0]])
wordcloud = WordCloud(width=800,height=500,random_state=21,max_font_size=110).generate(negative_words)
plt.figure(figsize=(10,7))
plt.imshow(wordcloud,interpolation="bilinear")
plt.axis('off')
plt.show()
```

```
positif_data = data[data.target==4].iloc[:10000,:]
print(positif_data.shape)
negative_data = data[data.target==0].iloc[:10000,:]
print(negative_data.shape)
data = pd.concat([positif_data,negative_data],axis = 0)
print(data.shape)
data.head()
```

```
(10000, 6)
(10000, 6)
(20000, 6)
```

	target	ids	date	flag	user	TweetText
799999	4	1467822272	Mon Apr 06 22:22:45 PDT 2009	NO_QUERY	ersle	I LOVE @Health4UandPets u guys r the best!!
800000	4	1467822273	Mon Apr 06 22:22:45 PDT 2009	NO_QUERY	becca210	im meeting up with one of my besties tonight! ...
800001	4	1467822283	Mon Apr 06 22:22:46 PDT 2009	NO_QUERY	Wingman29	@DaRealSunisaKim Thanks for the Twitter add, S...
800002	4	1467822287	Mon Apr 06 22:22:46 PDT 2009	NO_QUERY	katarinka	Being sick can be really cheap when it hurts t...
800003	4	1467822293	Mon Apr 06 22:22:46 PDT 2009	NO_QUERY	EmilyYoung	@LovesBrooklyn2 he has that effect on everyone

Next steps: Generate code with `data` View recommended plots New interactive sheet