process-twitter-data

October 30, 2015

In [257]:
```
import matplotlib.pyplot as plt
import csv
import pandas as pd
import itertools
import numpy as np
import random
import re
from collections import Counter

%matplotlib inline
```

In [262]:
```
POSITIVE = ['(^_^)', '(^o^)', '(^^)', '(^-^)', 'o', '', '(*^_^*)', '(*‘)', '(**)', '(**)', '(
NEGATIVE = ['(# )', '(_)', '(‘_‘)', '(^-^;', '(-_-;)', '(=_=;)', '(‘)', '(-o-;)', '(^^;)', '(
```

In [263]:
```
# Create data frame for smileys
smiley_df = pd.DataFrame.from_records(
    [[x, 'positive'] for x in POSITIVE] + [[x, 'negative'] for x in NEGATIVE],
    columns=['keyword', 'sentiment'])
smiley_df = smiley_df.drop_duplicates()
```

In [282]:
```
MIN_TEXT_LENGTH = 80

# Load the data
data = pd.read_csv('../data/data.csv', names=['keyword', 'text'])
# Preprocessing: Remove hashtags, URLs and user mentions
data.text = data.text.str.replace('#\S+', '<HASHTAG> ', case=False)
data.text = data.text.str.replace('https?://\S+', '<URL> ', case=False)
data.text = data.text.str.replace('@\S+', '<USER> ', case=False)
# Only consider tweets of certain length
data = data[data['text'].map(lambda x: len(x) > MIN_TEXT_LENGTH)]
# Remove all smileys
for smiley in POSITIVE + NEGATIVE:
    data.text = data.text.str.replace(re.escape(smiley), '<SMILEY> ', case=False)
# Join data with positive/negative sentiment
data = pd.merge(data, smiley_df, on='keyword')
data = data.drop_duplicates()
data = data[['keyword', 'sentiment', 'text']]
```

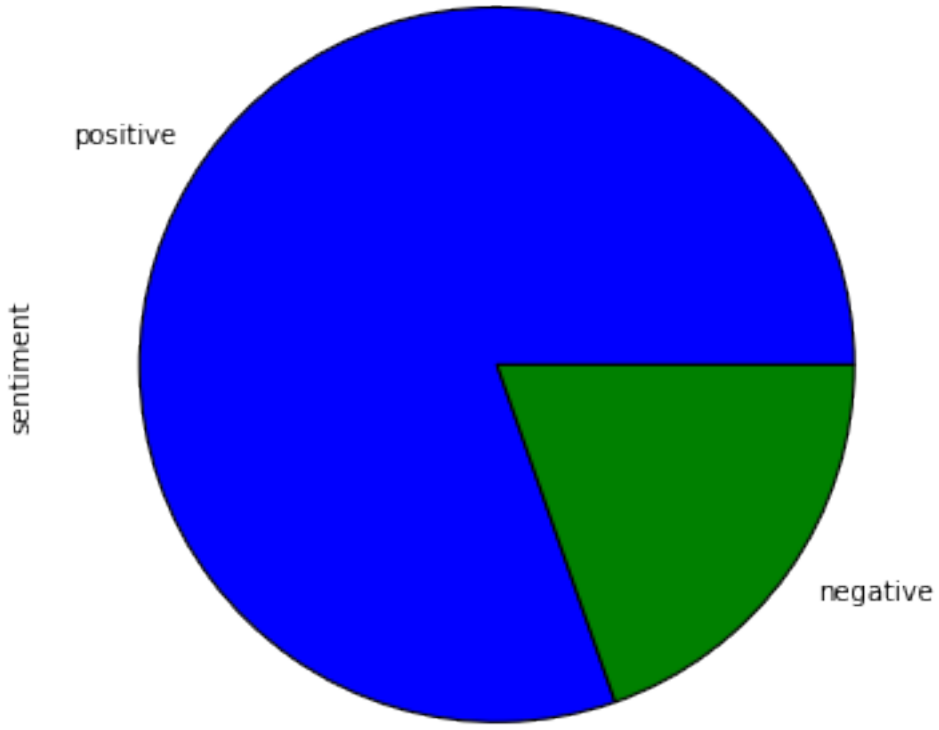
In [283]:
```
data.describe()
```

Out[283]:

	keyword	sentiment	text
count	601	601	601
unique	28	2	601
top	(^o^)	positive	<SMILEY> (^o^) <US...
freq	186	483	1

In [284]:
```
data.sentiment.value_counts().plot(kind='pie', figsize=(6,6), table=True)
```

Out[284]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x10ce0ad68>
```



	positive	negative
sentiment	483	118

In [285]:
```
# Subsample values from each smiley for manual checking
K = 10
filtered_data = data.groupby('keyword').filter(lambda x: len(x) > K)
indices = itertools.chain(*[np.random.choice(v, K, replace=False) for k, v in  filtered_data.g
subsampled_data = data.reindex(indices)
```

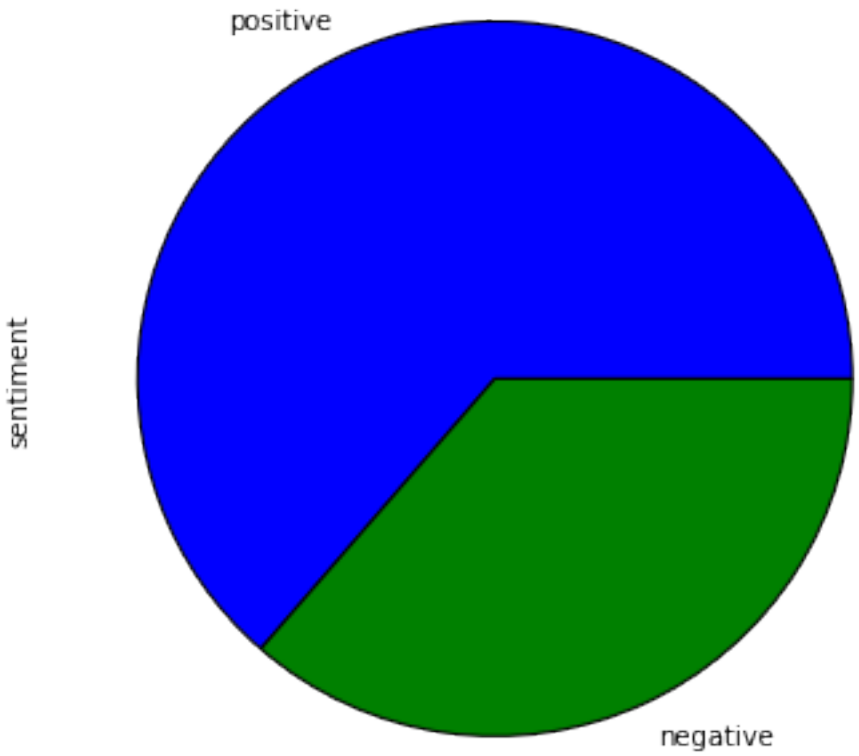
In [286]:
```
subsampled_data.describe()
```

Out[286]:

	keyword	sentiment	text
count	110	110	110
unique	11	2	110
top	(^-^)	positive	(4:02) <URL> <HASHT...
freq	10	70	1

In [287]:
```
# Show positive vs. negative smiley distribution
subsampled_data.sentiment.value_counts().plot(kind='pie', figsize=(6,6), table=True)
```

Out[287]: <matplotlib.axes._subplots.AxesSubplot at 0x10d12a860>



	positive	negative
sentiment	70	40

In [288]:
```
# Write to csv file
subsampled_data.to_csv("../data/samples.csv")
```

In [ ]: