Homework2

Jiarong Ye

September 13, 2018

```
In [332]: import pandas as pd
          import datascience as ds
          from datascience import *
          import numpy as np
          import re
          import string
```

Read the tagged tweets

Sentiment

- Irrelevant: -1
- Neutral: 0
- Positive: 1
- Negative: 2

```
In [333]: tagged= pd.read_csv('tagged.csv', sep= ',')
          tagged['sentiment'] = tagged['sentiment'].astype(dtype=int)
```

Clean the tweet text and remove nan

```
In [334]: def process_text(data):
              cleaned_text = [
                  re.sub('\s+', ' ',
              re.sub("([^0-9A-Za-z \t])|(\w+:\/\/\S+)|^rt|http.+?", '',
              tweets.lower()).strip(string.punctuation).strip()) for tweets in data
              ]
              return cleaned_text

          tagged['text'] = process_text(tagged['text'])
          tagged = tagged.drop(tagged.columns[0], axis=1).reset_index(drop=True)
```

```
In [335]: not_null_text = 1 ^ pd.isnull(tagged["text"])
          not_null_sentiment = 1 ^ pd.isnull(tagged["sentiment"])
          tagged = tagged.loc[not_null_sentiment & not_null_text, :]
```

```
In [336]: len(tagged)

Out[336]: 4817
```

Remove duplicates

```
In [337]: tagged = tagged.iloc[tagged['text'].drop_duplicates().index]
          tagged = tagged.reset_index(drop=True)
          tagged = tagged.drop(tagged.index[nan_idx.values])
          tagged = tagged.dropna(axis=0)
          tagged = tagged.reset_index(drop=True)
          tagged.to_csv('cleaned_tagged.csv', sep=',')
          tagged.head()
```

```
Out[337]:                user_id                       user_name  \
          0   802657195661742080                Christine Warren   
          1  1039245812230893570                  Trumpservative   
          2            282084840     Darrel Sheldon #MAGAVETERAN   
          3             62315639  Queer Liberal Voting Snowflake   
          4   823307049266245633          don jones #veteran (K)   

                                tweet_time         location  \
          0  Wed Sep 12 01:38:14 +0000 2018       Fremont CA   
          1  Wed Sep 12 01:38:16 +0000 2018         Tulsa OK   
          2  Wed Sep 12 01:38:18 +0000 2018             None   
          3  Wed Sep 12 01:38:18 +0000 2018  Big Sandy Texas   
          4  Wed Sep 12 01:38:19 +0000 2018    United States   

                                                         text  sentiment  
          0  2 american lives were just lost in tennessee p...          0  
          1  realfarmacist realdonaldtrump tuckercarlson ou...          1  
          2  rightgottweets rev out the trumptrain for blue...          0  
          3  tennessee is suffering after shooting takes 2 ...          0  
          4  sassytjoy bcmsolt fightinirish109 dybarb wolfi...         -1  
```

Count the tweets

```
In [338]: tagged = ds.Table.read_table('cleaned_tagged.csv', sep=',')
```

```
In [339]: def count_tweets(table):
              classes = {
                         'irrelevant':-1,
                         'neutral':0,
                         'positive':1,
                         'negative':2
                        }
              for i in classes:
                  c =  table.where('sentiment', are.equal_to(classes[i])).num_rows
                  print('There are {} {} tweets in the data'.format(c, i))
```

```
In [340]: count_tweets(tagged)
```

```
There are 417 neutral tweets in the data
There are 247 negative tweets in the data
There are 415 positive tweets in the data
There are 217 irrelevant tweets in the data
```