bot-tweets

December 5, 2016

1 Bot Tweets

(at long last)

Import our libraries needed for the data handling.

```
In [1]: import pandas as pd
        import numpy as np
        import json
        import glob
```

```
In [2]: #Set PANDAS to show all columns in DataFrame
        pd.set_option('display.max_columns', None)
```

Libraries for stupid text encoding

```
In [3]: from urllib2 import quote
        # Unicode strings
        from __future__ import unicode_literals
```

Import libraries needed for visualization.

```
In [4]: import matplotlib
        import matplotlib.pyplot as plt
        # Within notebook viewing
        %matplotlib inline

        print (matplotlib.__version__)
```

```
1.5.3
```

```
In [5]: # Import for axes, color, etc
        from pylab import *
```

Natural Language Processing

```
In [6]: %run twokenize.py
        import nltk
```

1.0.1 Directories

```
In [7]: testDir = '../../data/external/trump-bots/'
        botDir = '../../data/external/botresults/'
        outDir = '../../data/processed/bot-tweets/'
```

Read in the data files by combining the extracted files.

```
In [8]: # Crudely combine
        process = []
        for f in glob.glob((botDir + "*.txt")):
            with open(f, "rb") as infile:
                for line in infile:
                    process.append(json.loads(line))
        raw = pd.DataFrame.from_records(process)

        del process

        print (raw.shape)
```

```
(77722, 33)
```

Helper functions

```
In [56]: def tknz(text):
             tokens = tokenizeRawTweetText(text)
             filt = [x for x in tokens if not (x.startswith('RT')
                                               or x.startswith('@') or x.startswith('
                                               or x.startswith('http://') or x.startsw
                                              or x.startswith('-') or x.startswith('ar
                                                or x.startswith('.') or x.startswith('
                                                 or '')]
             return filt

         def hsh(tokens):
             # reads a list of tokens
             tuped = tuple(tokens)
             hashed = hash(tuped)
             return hashed
```

```
In [44]: def extractInfo(tweet):
             # User variables
             userID = tweet['user']['id_str']
             screenName = tweet['user']['screen_name']
             # Tweet Variables
             text = tweet['text']
             tokens = tknz(text)
             hashed = hsh(tokens)
```

```
        timestamp = tweet['timestamp_ms']

        return {'userID':userID, 'screenName':screenName, 'timestamp':timestam
                'text':text, 'tokens':tokens, 'hash':hashed}

        #### below is a horrible rendition of network variables

    """
    # RT, reply, quoting, or none?
        try:
            if (tweet['retweeted_status'] == True):
                method = 'retweet'
                otherID = tweet['retweeted_status']['user']['id_str']

            if (type(tweet['in_reply_to_user_id_str']) != np.float64):
                method = 'replyUser'
                otherID = tweet['in_reply_to_user_id_str']
            elif (tweet['in_reply_to_status_id_str']):
                method = 'replyStatus'
                otherID = None
            else:
                method = None
                otherID = None

            return {'userID':userID, 'screenName':screenName, 'text':text, 'ti
                    'method':method, 'otherID':otherID}
        except:
            print tweet
    """
```

Out[44]:
```
u"\n# RT, reply, quoting, or none?\n    try:\n        if (tweet['retweeted
```

In [45]:
```
extracted = pd.DataFrame.from_records(raw.apply(lambda x: extractInfo(x),
extracted['timestamp'] = pd.to_datetime(extracted['timestamp'], unit='ms')

extracted.head()
```

Out[45]:

	hash	screenName	text	timesta
0	6280506488202206663	WDYL2016	Donald Trump calls for 'civil' debate - CNN ht...	2015-08-06 00:10:34.7
1	-239533655759035841	WDYL2016	Dump Donald Trump and let Rick Perry debate - ...	2015-08-06 00:10:44.8
2	-2242831579107199855	azblonde2015		
3	-5395197686682151618	iVoteForBest		
4	8035591063244529552	TBackers		

```
2  Contribute Today! :: Donald J. Trump for Presi... 2015-08-06 00:15:09.5
3  #ModiMinistry As Republicans take the debate s... 2015-08-06 00:15:45.3
4  As Republicans take the debate stage, all eyes... 2015-08-06 00:16:52.4

                                                 tokens      userID
0  [Donald, Trump, calls, for, ', civil, ', debat...  2414927882
1  [Dump, Donald, Trump, and, let, Rick, Perry, d...  2414927882
2  [Contribute, Today, !, Donald, J, Trump, for, ...  3271255423
3  [#ModiMinistry, As, Republicans, take, the, de...  2425268995
4  [As, Republicans, take, the, debate, stage, al...  3044768595
```

In [46]:
```
hashes = extracted[['screenName', 'text', 'tokens', 'hash']].set_index('ha
hashes.head()
```

Out[46]:
```
                          screenName  \
hash
 6280506488202206663      WDYL2016
-239533655759035841       WDYL2016
-2242831579107199855  azblonde2015
-5395197686682151618  iVoteForBest
 8035591063244529552      TBackers

                                                                text  \
hash
 6280506488202206663  Donald Trump calls for 'civil' debate - CNN ht...
-239533655759035841   Dump Donald Trump and let Rick Perry debate - ...
-2242831579107199855  Contribute Today! :: Donald J. Trump for Presi...
-5395197686682151618  #ModiMinistry As Republicans take the debate s...
 8035591063244529552  As Republicans take the debate stage, all eyes...

                                                              tokens
hash
 6280506488202206663  [Donald, Trump, calls, for, ', civil, ', debat...
-239533655759035841   [Dump, Donald, Trump, and, let, Rick, Perry, d...
-2242831579107199855  [Contribute, Today, !, Donald, J, Trump, for, ...
-5395197686682151618  [#ModiMinistry, As, Republicans, take, the, de...
 8035591063244529552  [As, Republicans, take, the, debate, stage, al...
```

In [72]:
```
# Get most popular
print ('Most popular tweet hashes.')

popularTweets = extracted['hash'].value_counts()
#popularTweets = popularTweets.reset_index()
popularTweets.columns = ['frequency']

popularTweets.head(10)
```

```
Most popular tweet hashes.
```

```
Out[72]:  717126146592199451     1574
          5162929327475364277    1362
         -4501194672645902651     807
         -2692256532814125541     624
          3527539                 609
          7510515497875988663     520
         -4907917567665423769     484
          2314862951965282667     176
          8969368713954905093     146
         -7003498234340419358     142
         Name: hash, dtype: int64
```

```
In [48]: tweetsByUsers = extracted[['userID', 'hash', 'text']].groupby('hash')\
             .agg({'userID': pd.Series.nunique})\
             .sort_values(by='userID', ascending=False)

         print ('Tweets with multiple (unique) users.')
         tweetsByUsers.head()
```

```
Tweets with multiple (unique) users.
```

```
Out[48]:                       userID
         hash
         -2928380469476387809      50
          8969368713954905093      50
          191154071824568122       48
          2378718859638583145      47
         -2296336973128567635      19
```

```
In [68]: tweets = raw[raw['retweeted'] == False]
         retweets = raw[raw['retweeted'] == True]
```

```
In [ ]:
```