Sathish_Kumar_Rajendiran_Week8_8.3_Twitter_bball_Activity

August 19, 2020

Name: Sathish Kumar Rajendiran Task: Week8: 8.3 MongoDB & Tweets Collection Date: 8/19/2020

[ ]:
```
#pip install pymongo library
!pip install pymongo
!pip install tweepy
```

[1]:
```
#import libraries

# standard library
import os
import sys
from datetime import datetime
import time
import re

# csv, xls, pandas & json
import pandas as pd
import json
import csv
import xlrd

#twitter libraries
import tweepy
from tweepy import StreamListener
from tweepy import Stream
import preprocessor as p
# from tweet-preprocessor import clean,tokenize,parse

#MongoDB libraries
import pymongo
from pymongo import MongoClient

#visualization
import matplotlib.pyplot as plt
from wordcloud import WordCloud, STOPWORDS
import seaborn as sns
%matplotlib inline
```

```
import timeit
# import random

os.getcwd()
```

[1]:
```
'/Users/sathishrajendiran/ist652-python'
```

[38]:
```
# Connecting to the database
# Connection to Mongo DB
try:
    client = MongoClient('localhost', 27017)
    print ('Authentication OK - You''re now connected to the MongoDB.\n')
# use database named usgs or create it if not there already
    db = client.bball
    # create collection named earthquakes or create it if not there already
    coll = db.bbcoll

#   print('MongoDB: Tweets db with tweets collection has been created')
    print('MongoDB database: ' + str(db))
    print('MongoDB collection:' + str(coll))

except pymongo.errors.ConnectionFailure as e:
    print ('Could not connect to MongoDB: %s' % e )
```

```
Authentication OK - Youre now connected to the MongoDB.

MongoDB database: Database(MongoClient(host=['localhost:27017'],
document_class=dict, tz_aware=False, connect=True), 'bball')
MongoDB collection:Collection(Database(MongoClient(host=['localhost:27017'],
document_class=dict, tz_aware=False, connect=True), 'bball'), 'bbcoll')
```

[58]:
```
#list database defined
client.list_database_names
```

[58]:
```
<bound method MongoClient.list_database_names of
MongoClient(host=['localhost:27017'], document_class=dict, tz_aware=False,
connect=True)>
```

[59]:
```
#search the first item from the collection
coll.find_one()
```

[59]:
```
{'_id': ObjectId('58d992de64a4f3e56d2db3c7'),
 'user': {'profile_background_tile': True,
  'friends_count': 217,
  'profile_sidebar_fill_color': 'EFEFEF',
  'id_str': '513196438',
```

```
  'is_translation_enabled': False,
  'profile_link_color': 'F70808',
  'followers_count': 256,
  'location': '',
  'protected': False,
  'default_profile_image': False,
  'contributors_enabled': False,
  'favourites_count': 2187,
  'profile_background_color': 'BFD0D9',
  'statuses_count': 1104,
  'id': 513196438,
  'profile_banner_url':
'https://pbs.twimg.com/profile_banners/513196438/1437359097',
  'created_at': 'Sat Mar 03 13:54:37 +0000 2012',
  'profile_image_url_https':
'https://pbs.twimg.com/profile_images/732514428051181568/OkOVa8Ia_normal.jpg',
  'time_zone': 'Eastern Time (US & Canada)',
  'follow_request_sent': None,
  'listed_count': 1,
  'utc_offset': -14400,
  'lang': 'en',
  'is_translator': False,
  'name': 'Will',
  'description': "Do not worry about tomorrow, for tomorrow will worry about
itself.-Matthew 6:34 HSC'20",
  'profile_use_background_image': True,
  'verified': False,
  'geo_enabled': True,
  'profile_text_color': '333333',
  'profile_image_url':
'http://pbs.twimg.com/profile_images/732514428051181568/OkOVa8Ia_normal.jpg',
  'entities': {'description': {'urls': []}},
  'notifications': None,
  'url': None,
  'translator_type': 'none',
  'has_extended_profile': True,
  'default_profile': False,
  'screen_name': 'frenchythe1st',
  'following': None,
  'profile_background_image_url': 'http://pbs.twimg.com/profile_background_image
s/850934042/26e0a43c1f821ac098571fb3de80944c.jpeg',
  'profile_sidebar_border_color': 'FFFFFF',
  'profile_background_image_url_https': 'https://pbs.twimg.com/profile_backgroun
d_images/850934042/26e0a43c1f821ac098571fb3de80944c.jpeg'},
 'metadata': {'iso_language_code': 'en', 'result_type': 'recent'},
 'favorited': False,
 'in_reply_to_user_id': None,
```

```
 'text': 'RT @HowardWKYT: The final seconds of the Kentucky-North Carolina game
was an emotional roller coaster for Big Blue fans. #marchmadness #WKY…',
 'retweet_count': 9739,
 'id': 8.464896333870899e+17,
 'in_reply_to_screen_name': None,
 'created_at': 'Mon Mar 27 22:30:30 +0000 2017',
 'is_quote_status': False,
 'retweeted_status': {'user': {'profile_background_tile': False,
   'friends_count': 573,
   'profile_sidebar_fill_color': 'C0DFEC',
   'id_str': '27059989',
   'is_translation_enabled': False,
   'profile_link_color': '0084B4',
   'followers_count': 3330,
   'location': 'Lexington, KY',
   'protected': False,
   'default_profile_image': False,
   'contributors_enabled': False,
   'favourites_count': 1231,
   'profile_background_color': '000000',
   'statuses_count': 9463,
   'id': 27059989,
   'profile_banner_url':
'https://pbs.twimg.com/profile_banners/27059989/1431999643',
   'created_at': 'Fri Mar 27 18:09:16 +0000 2009',
   'profile_image_url_https':
'https://pbs.twimg.com/profile_images/575376840304386049/atONJG3G_normal.jpeg',
   'time_zone': 'Eastern Time (US & Canada)',
   'follow_request_sent': None,
   'listed_count': 77,
   'utc_offset': -14400,
   'lang': 'en',
   'is_translator': False,
   'name': 'Lee K. Howard',
   'description': 'Sports Anchor/Reporter for CBS/FOX in Lexington Kentucky,
providing sports news and my random rants and chants!',
   'profile_use_background_image': True,
   'verified': True,
   'geo_enabled': True,
   'profile_text_color': '333333',
   'profile_image_url':
'http://pbs.twimg.com/profile_images/575376840304386049/atONJG3G_normal.jpeg',
   'entities': {'description': {'urls': []},
    'url': {'urls': [{'expanded_url':
'http://www.facebook.com/profile.php?id=100003241678454',
       'display_url': 'facebook.com/profile.php?id…',
       'indices': [0, 22],
```

```
       'url': 'http://t.co/C2UFfLkjB1'}]}},
   'notifications': None,
   'url': 'http://t.co/C2UFfLkjB1',
   'translator_type': 'none',
   'has_extended_profile': False,
   'default_profile': False,
   'screen_name': 'HowardWKYT',
   'following': None,
   'profile_background_image_url': 'http://pbs.twimg.com/profile_background_imag
es/596008542/6qfqqg0jndpp3su53ay5.jpeg',
   'profile_sidebar_border_color': 'A8C7F7',
   'profile_background_image_url_https': 'https://pbs.twimg.com/profile_backgrou
nd_images/596008542/6qfqqg0jndpp3su53ay5.jpeg'},
  'metadata': {'iso_language_code': 'en', 'result_type': 'recent'},
  'favorited': False,
  'in_reply_to_user_id': None,
  'text': 'The final seconds of the Kentucky-North Carolina game was an
emotional roller coaster for Big Blue fans.… https://t.co/TPZ6PuXHxH',
  'possibly_sensitive': False,
  'retweet_count': 9739,
  'id': 8.462142300261786e+17,
  'in_reply_to_screen_name': None,
  'created_at': 'Mon Mar 27 04:16:08 +0000 2017',
  'is_quote_status': False,
  'favorite_count': 12590,
  'contributors': None,
  'lang': 'en',
  'in_reply_to_status_id_str': None,
  'source': '<a href="http://twitter.com" rel="nofollow">Twitter Web
Client</a>',
  'in_reply_to_status_id': None,
  'place': {'country': 'United States',
   'contained_within': [],
   'id': '6ffcf3b0b904bbcb',
   'country_code': 'US',
   'bounding_box': {'type': 'Polygon',
    'coordinates': [[[-89.57151, 36.497129],
      [-81.964971, 36.497129],
      [-81.964971, 39.147359],
      [-89.57151, 39.147359]]]},
   'place_type': 'admin',
   'full_name': 'Kentucky, USA',
   'url': 'https://api.twitter.com/1.1/geo/id/6ffcf3b0b904bbcb.json',
   'name': 'Kentucky',
   'attributes': {}},
  'entities': {'hashtags': [],
   'urls': [{'expanded_url':
```

```
'https://twitter.com/i/web/status/846214230026178564',
     'display_url': 'twitter.com/i/web/status/8…',
     'indices': [106, 129],
     'url': 'https://t.co/TPZ6PuXHxH'}],
   'symbols': [],
   'user_mentions': []},
  'geo': None,
  'truncated': True,
  'coordinates': None,
  'in_reply_to_user_id_str': None,
  'retweeted': False,
  'id_str': '846214230026178564'},
 'favorite_count': 0,
 'contributors': None,
 'lang': 'en',
 'in_reply_to_status_id_str': None,
 'source': '<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter
for iPhone</a>',
 'in_reply_to_status_id': None,
 'place': None,
 'entities': {'hashtags': [{'indices': [121, 134], 'text': 'marchmadness'}],
  'urls': [],
  'symbols': [],
  'user_mentions': [{'indices': [3, 14],
    'name': 'Lee K. Howard',
    'id': 27059989,
    'id_str': '27059989',
    'screen_name': 'HowardWKYT'}]},
 'geo': None,
 'truncated': False,
 'coordinates': None,
 'in_reply_to_user_id_str': None,
 'retweeted': False,
 'id_str': '846489633387089920'}
```

[60]:
```
docs = coll.find()
print('docs cursor ready!')
```

```
docs cursor ready!
```

[35]:
```
# def print_tweet_data(tweets):
#     for tweet in tweets:
#         print('\nDate:', tweet['created_at'])
#         print('From:', tweet['user']['name'])
#         print('Screen Name:', tweet['user']['screen_name'])
#         print('Message', tweet['text'])
```

[37]:
```
# print_tweet_data(doclist[:5])
```

[67]:
```
docs = coll.find()
tw_list =[]
doclist = [tweet for tweet in docs]
for tweet in doclist:
    id = tweet['id']
    user = tweet['user']['name']
    friends_count = tweet['user']['friends_count']
    followers = tweet['user']['followers_count']
    retweets = tweet['retweet_count']
    bg_color = tweet['user']['profile_background_color']
    datets = tweet['created_at'] # select unix timestamp in milliseconds
    text = tweet['text']
    tw_list.
 ↪append([id,user,friends_count,followers,retweets,bg_color,datets,text])
print('Number of docs: ',len(tw_list))
```

```
Number of docs:  2000
```

[68]:
```
#define column names
ColNames =␣
 ↪['id','user','friends_count','followers','retweets','bg_color','datets','text']

# Show all columns and do not truncate in the data frame
pd.set_option('display.max_columns', None)
pd.set_option('display.max_colwidth', None)

tweetsDF = pd.DataFrame(tw_list,columns=ColNames)

print('Total Number of rows Processed: ',len(tweetsDF))
```

```
Total Number of rows Processed:  2000
```

[69]:
```
tweetsDF.head()
```

[69]:

	id	user	friends_count	followers	retweets	bg_color	datets
0	8.464896e+17	Will	217	256	9739	BFD0D9	Mon Mar 27 22:30:30 +0000 2017
1	8.464896e+17	Kahlen Donatell	468	386	71	131516	Mon Mar 27 22:30:22 +0000 2017
2	8.464896e+17	Jesús	156	262	9739	C0DEED	Mon Mar 27 22:30:21 +0000 2017
3	8.464896e+17	plug,	55	23	13493	000000	
4	8.464896e+17	Joni Dickerson	381	273	12	EBEBEB	

```
3  Mon Mar 27 22:30:14 +0000 2017
4  Mon Mar 27 22:30:11 +0000 2017

                                                                text
0  RT @HowardWKYT: The final seconds of the Kentucky-North Carolina game was an
emotional roller coaster for Big Blue fans. #marchmadness #WKY…
1           RT @WhistleSports: When you perfectly time the #UNC buzzer beater
#MarchMadness\n\n(via:@SamuelGrubbs1 ) https://t.co/Ol2ibpZjB4
2  RT @HowardWKYT: The final seconds of the Kentucky-North Carolina game was an
emotional roller coaster for Big Blue fans. #marchmadness #WKY…
3                                                          RT @BleacherReport:
And then there were four… #MarchMadness https://t.co/OMbxpgAuUC
4                      RT @mycarolinastdnt: RT if you'll be cheering on
@GamecockWBB and @dawnstaley tonight. Let's Go Gamecocks! #MarchMadness
```

[83]:
```
##### datetime conversion
tweetsDF['datets'] = tweetsDF['datets'].astype('datetime64[ns]')
tweetsDF['date'] = tweetsDF['datets'].dt.date
tweetsDF['month'] = tweetsDF['datets'].dt.month
tweetsDF['monthday'] = tweetsDF['datets'].dt.day
tweetsDF['weekday'] = tweetsDF['datets'].dt.weekday
tweetsDF['dayname'] = tweetsDF['datets'].dt.day_name()
tweetsDF['monthname'] = tweetsDF['datets'].dt.month_name()
tweetsDF['hour'] = tweetsDF['datets'].dt.hour
tweetsDF['minute'] = tweetsDF['datets'].dt.minute
tweetsDF['secs'] = tweetsDF['datets'].dt.second
```

[84]:
```
#validate Number of retweets by values

retweets = tweetsDF[tweetsDF['text'].str.startswith('RT')==True]
print('Number of retweets: ',len(retweets))
print('Percentage of retweets {}%'.format(round((len(retweets))/
 ↪len(tweetsDF['text'])*100,2)))
```

```
Number of retweets:  1499
Percentage of retweets 74.95%
```

[85]:
```
#validate Number of retweets by values

act_tweets = tweetsDF[tweetsDF['text'].str.startswith('RT')==False]
print('Number of actual tweets: ',len(act_tweets))
print('Percentage of actualtweets {}%'.format(round((len(act_tweets))/
 ↪len(tweetsDF['text'])*100,2)))
```

```
Number of actual tweets:  501
Percentage of actualtweets 25.05%
```

[86]:
```
act_tweets
```

[86]:

```
                 id                   user  friends_count  followers  retweets  \
7     8.464895e+17   GooglePlayNewsstand            198       1151         0
8     8.464895e+17             2 3 GANG            785       1038         0
9     8.464895e+17            allAboutYOU            142         54         0
10    8.464895e+17  IndustryNightInsider            525       2034         0
19    8.464895e+17            faith yates             83         98         0
…              …                     …              …          …         …
1980  8.464640e+17          alenasbdesign            208        509         0
1986  8.464640e+17         CollegeBB News          65210      65171        18
1991  8.464639e+17           Steve Clarke           1594        316         0
1992  8.464639e+17            Aniket Goel            224        139         0
1995  8.464639e+17           Josie Degler            796        550         0

     bg_color              datets  \
7      F5F8FA 2017-03-27 22:30:09
8      FFF04D 2017-03-27 22:30:07
9      F5F8FA 2017-03-27 22:30:07
10     131516 2017-03-27 22:30:02
19     000000 2017-03-27 22:29:50
…       …                   …
1980   EBEBEB 2017-03-27 20:48:50
1986   C0DEED 2017-03-27 20:48:36
1991   C0DEED 2017-03-27 20:48:25
1992   000000 2017-03-27 20:48:23
1995   DBE9ED 2017-03-27 20:48:15

                                                                    text  \
7                                                  "It’s OK to believe in the
hot hand." https://t.co/yJOqhlRz8L h/t @ConversationUS #MarchMadness #NCAA
8
It's gonna be a Civil War for the finals! #MarchMadness
9                                    Beach ready? #trend #vip #marketing
#marchmadness #UnnecessaryConfessions #younotfromhouston…
https://t.co/AtoeIuQuXk
10                     Industry Monday! @ShakerChicago #mondaymotivation
#internationalwhiskyday #MarchMadness #spanishpaelladay… https://t.co/rTw25qvjRS
19                                 i won by default   but i'm so proud of @hhn4334
and aiden for our first #marchmadness together https://t.co/XPO3jztjdG
…
…
1980                    25% OFF SITEWIDE\n#CouponCode:
MARCHMADNEZZ\nhttps://t.co/VWjBGlTZfv\n#marchmadness #supersale #onsalenow
https://t.co/DDz84pffQy
1986  Vegas odds to win the National Championship\n\n1. Gonzaga 3/2\n\n2. North
Carolina 7/5\n\n3. Oregon 9/2\n\n4. South Carolina 15/2\n\n#MarchMadness
```

```
1991
@CandaceBurnsTV got the best #marchmadness commercials
1992                                                  How Much Does the NCAA Make
off March Madness? | #Investopedia #MarchMadness https://t.co/I8CNi1zLS4
1995
wot in tar heel nation #marchmadness #UKvsUNC https://t.co/rG2vO79eja

            date  month  monthday  weekday dayname monthname  hour  minute  \
7     2017-03-27      3        27        0  Monday     March    22      30
8     2017-03-27      3        27        0  Monday     March    22      30
9     2017-03-27      3        27        0  Monday     March    22      30
10    2017-03-27      3        27        0  Monday     March    22      30
19    2017-03-27      3        27        0  Monday     March    22      29
…           …    …        …       …      …         …    …      …
1980  2017-03-27      3        27        0  Monday     March    20      48
1986  2017-03-27      3        27        0  Monday     March    20      48
1991  2017-03-27      3        27        0  Monday     March    20      48
1992  2017-03-27      3        27        0  Monday     March    20      48
1995  2017-03-27      3        27        0  Monday     March    20      48

      secs
7        9
8        7
9        7
10       2
19      50
…    …
1980    50
1986    36
1991    25
1992    23
1995    15

[501 rows x 17 columns]
```

[87]:
```
#number of tweets by date
by_date = act_tweets.groupby(['date']).size().reset_index(name='counts')
by_date = by_date.set_index('date')
by_date.reset_index(level=0,inplace = True,drop=False)
by_date.index += 1
print('Tweets by calender date: ')
by_date
```

```
Tweets by calender date:
```

[87]:

	date	counts
1	2017-03-27	501

[88]:
```
#number of tweets by hour of day
by_day_hour = act_tweets.groupby(['dayname','hour']).size().
 ↪reset_index(name='counts')
by_day_hour = by_day_hour.set_index('dayname')
by_day_hour.reset_index(level=0,inplace = True,drop=False)
by_day_hour.index += 1
print('Tweets by hour of the day: ')
by_day_hour
```

```
Tweets by hour of the day:
```

[88]:

	dayname	hour	counts
1	Monday	20	55
2	Monday	21	296
3	Monday	22	150