CSC455Assignment5 Email:kying@mail.depaul.edu
Date: 5/31/2017

Problem 1-a:
Create Tables: Tweet , User

In [2]:
```
#Import pandas and numpy library
import pandas as pd
import numpy as np
import urllib.request as urlreq
import lxml.etree as et
import json
import pprint
```

In [12]:
```
#Part1-a
#Create Table Tweet from Assignment 4
tw='''Create Table Tweet(

    created_at DATE,
    id_str VARCHAR(20),
    text VARCHAR(100),
    source VARCHAR(100),
    in_reply_to_user_id VARCHAR(20),
    in_reply_to_screen_name VARCHAR(20),
    in_reply_to_status_id VARCHAR(20),
    retweet_count INTEGER(5),
    contributors VARCHAR(10),
    user_id VARCHAR(20),

    CONSTRAINT Tweet_FK
        FOREIGN KEY (user_id)
        REFERENCES User(id)
)
'''

#Create Table User
user='''Create Table User(
    id VARCHAR(20),
    name VARCHAR(30),
    screen_name VARCHAR(100),
    description VARCHAR(100),
    friends_count INTEGER(5)
)
'''

import sqlite3
from sqlite3 import OperationalError
conn=sqlite3.connect('csc455_hw5.db')
c=conn.cursor()
c.execute('Drop Table if exists Tweet;')
c.execute(tw)
c.execute('Drop Table if exists User;')
c.execute(user)
```

Out[12]:
```
<sqlite3.Cursor at 0x11a6c9f80>
```

CSC455Assignment5 Email:kying@mail.depaul.edu
Date: 5/31/2017

Problem1-b:

In [5]:
```
#Part1-b
#Open assignment5.txt
file = open("assignment5.txt","r")
content=file.read()
content.strip()
lines=content.split('\n')
count=0
err_lst=[]
```

In [6]:
```
#To load the contents from json format
for tweet in lines:
    try:
        obj=json.loads(tweet)
        #pprint.pprint(obj)

        #INSERT to data to Table Tweet
        c.execute("INSERT INTO Tweet Values(?,?,?,?,?,?,?,?,?,?);",
        (obj['created_at'],
        obj['id_str'],
        obj['text'],
        obj['source'],
        obj['in_reply_to_user_id'],
        obj['in_reply_to_screen_name'],
        obj['in_reply_to_status_id'],
        obj['retweet_count'],
        obj['contributors'],
        obj['user']['id']))

        #INSERT data to Table User
        c.execute("INSERT INTO User Values(?,?,?,?,?);",
        (obj['user']['id'],
         obj['user']['name'],
         obj['user']['screen_name'],
         obj['user']['description'],
         obj['user']['friends_count']))

    except ValueError:#Error for damaged tweet
        count+=1 #Record the number of damaged tweet
        #err_lst.append(tweet)
        #Save all the damaged tweet into the Assignment5_errors.txt file
        with open('Assignment5_errors.txt', 'a') as f:
            print(tweet,file=f)
```

Table Tweet includes 9797 Records

In [91]:
```
#Check the table Tweet
data=c.execute("select * from Tweet;").fetchall()

#Check the number of tweet extracted into the Table tweet
print(len(data))

#check the tweet content from the tweet table
for line in data:
    print(line)
```

```
9797
('Tue May 20 00:00:19 +0000 2014', '468541694288207874', 'la asusto a selena me dice es joda te vy a extrañar jajaja
jaja ni m fui pero ta vy a tener tiempo libre y todo wi', '<a href="https://mobile.twitter.com" rel="nofollow">Mobil
e Web (M2)</a>', None, None, None, 0, None, '367361405')
('Tue May 20 00:00:19 +0000 2014', '468541694284017664', '【現在視聴中のアニメ】ソウルイーターノット!/僕らはみんな河合荘/メカク
シティアクターズ/一週間フレンズ。/悪魔のリドル/ハイキュー!! etc…', '<a href="http://makebot.sh" rel="nofollow">ほたるボット</a>
', None, None, None, 0, None, '1605442621')
('Tue May 20 00:00:19 +0000 2014', '468541694279835649', '瑤潤 嬉陽(ようじゅん きひ)先生が待機開始しました', '<a href="http
://admin.pure-c.jp/prog01_test3/tweet_app/twitter_sample.php" rel="nofollow">purely_c tweet app</a>', None, None, No
ne, 0, None, '1685288690')
('Tue May 20 00:00:19 +0000 2014', '468541694292402176', '10 More FREE Pampers Gifts to Grow Points! via http://t.co
/3Sgq2UVq67 - We have a new Pampers Gifts ... http://t.co/DJa59HyRDi', '<a href="http://www.feedblitz.com/f/f.fbz?he
lp/default#twitter" rel="nofollow">FeedBlitz</a>', None, None, None, 0, None, '766497559')
('Tue May 20 00:00:19 +0000 2014', '468541694292410369', '@Tinkonsan まじで！ リアルマネー稼げるって聞いたんだけどどうなん？
？効率いい？', '<a href="https://sites.google.com/site/tweentwitterclient/" rel="nofollow">Tween</a>', '428260826', 'T
inkonsan', '468536777804767200', 0, None, '229690290')
('Tue May 20 00:00:19 +0000 2014', '468541694284009472', 'RT @bijyo_bijin: タイプだったらRT http://t.co/inZzxp17vh',
'<a href="https://twitter.com/shinri_black" rel="nofollow">とびだせツイッターの森</a>', None, None, None, 0, None, '237
4270429')
('Tue May 20 00:00:19 +0000 2014', '468541694317568001', 'ことほのうみED～！', '<a href="http://twitter.com/download/i
```

CSC455Assignment5 Email:kying@mail.depaul.edu
Date: 5/31/2017

Table User includes 9797 Records

In [92]:
```
#Check the table User
data=c.execute("select * from User;").fetchall()

#Check the number of user info extracted into the User Table
print(len(data))

#Check the content of user info from the User Table
for line in data:
    print(line)
```

```
9797
('367361405', 'gugusuarez', 'guadalupesuare4', None, 543)
('1605442621', 'ほたる', 'htr_ruby', '黒バス/青エク/あの花/君僕/進撃/Free! etc.只今黒バス/ハイキュー!!が熱いです 詳しくはツイプロで(ฅ•ω•ฅ)', 117)
('1685288690', '電話占いピュアリ', 'purely_c', '人気・口コミランキングNo.1！当たると評判の電話占い。タレントや芸能人も御用達でテレビや雑誌で話題の一流占い師が多数在籍。復縁や複雑な恋愛相談はお任せ下さい。初回最大4000円分無料！', 0)
('766497559', 'Veronica', 'verodoglover', None, 89)
('229690290', 'こさんつ', 'Cosan2', '(`ω´)\u3000東方やってます。 たまにお絵かき。\u3000天則のことつぶやくことが多いかも\u3000 規制用垢(@Cosantwo) 絵はじめました(24.7/4)', 610)
('2374270429', 'マジ!? 知らなきゃ損する雑学', 'zatugaku_fan', 'そうだったのか！と思わず唸ってしまう究極の雑学をお届け！\u3000知らなかったらRTしてね☆', 200)
('1351058173', 'yui', '_yuuu1_', '成人済 腐', 82)
('141120654', 'GMK', 'gmnyjewel', 'トゥートゥートゥー♪', 17)
('2478849157', 'Chulbul Pandey', 'chulbulpunday', None, 8)
('1297356302', 'Joenalyn Dela Cruz', 'jeandelacruz23', 'Happy to be with someone that makes me happy. Im so a BIG FAN of BRYAN TERMULO.  Certified SBN :)', 2001)
('134321102', 'Thais Ruiz', 'thais_ruiz', None, 92)
('1218018326', 'HweEjr .', 'H_HEJO', 'http://ask.fm/hajer894 .', 112)
('1450442486', 'ハンジさんbot', 'love_sony_bean', '進撃の巨人に出てくる巨人大好きなハンジさん 非公式自動botです。元気いっぱいなハンジさんは兵長を小馬鹿にしています。詳細はURLから。稀に手動稼働。', 759)
```

Assignment5_errors.txt includes 204 damaged tweet

In [17]:
```
#Check the content of damaged tweets in error file
content=open("Assignment5_errors.txt","r").readlines()

#Check the number of lines of damaged tweet
print(len(content))

#Check the content of damaged tweet from the Assignment5_errors.txt
for line in content:
    print(line)
```

```
204
{"created_at":"Tue May 20 00:00:19 +0000 2014","id":468541694288228350,"id_str":"468541694288228352","text":"Can this bitch not rush me!","source":"<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for Android</a>","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":1206334711,"id_str":"1206334711","name":"Cass .♥⚓","screen_name":"CassieSaravia","location":"","url":null,"description":"Devin Lee Leija & Serenity Nevaeh Cabrera are my  ♥","protected":false,"followers_count":534,"friends_count":496,"listed_count":1,"created_at":"Fri Feb 22 00:04:44 +0000 2013","favourites_count":7645,"utc_offset":null,"time_zone":null,"geo_enabled":true,"verified":false,"statuses_count":37275,"lang":"en","contributors_enabled":false,"is_translator":false,"is_translation_enabled":false,"profile_background_color":"C0DEED","profile_background_image_url":"http://abs.twimg.com/images/themes/theme1/bg.png","profile_background_image_url_https":"https://abs.twimg.com/images/themes/theme1/bg.png","profile_background_tile":false,"profile_image_url":"http://pbs.twimg.com/profile_images/457248656605712385/E8ygjqws_normal.jpeg","profile_image_url_https":"https://pbs.twimg.com/profile_images/457248656605712385/E8ygjqws_normal.jpeg","profile_banner_url":"https://pbs.twimg.com/profile_banners/1206334711/1398442668","profile_link_color":"0084B4","profile_sidebar_border_color":"C0DEED","profile_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"default_profile":true,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentions":[]},"favorited":false,"retweeted":false,"lang":"en"}

{"created_at":"Tue May 20 00:00:20 +0000 2014","id":468541698511872000,"id_str":"468541698511872001","text":"Two guy
```

CSC455Assignment5 Email:kying@mail.depaul.edu
Date: 5/31/2017

Problem 2a:

In [17]:
```
#Problem 2-a: Fine users ("id" and "name") with the minimum "friend_count" in the database
data=c.execute("select id,name from User where friends_count<=(select min(friends_count) from User)").fetchall()
for line in data:
    print(line)
```

```
('1685288690', '電話占いピュアリ')
('1640311244', 'تابعو البابو')
('2484449988', 'سعودية')
('469068614', 'おあおい')
('2498024720', '【彼氏のあるべき姿】')
('2508769764', 'سعود السميران')
('2421149844', '翼ばばぁ@ChasyacatFx')
('2505430447', 'AHHQ')
('2379017190', '出雲ハルキ@ダイエット支援bot')
('2400158192', 'スゲェ雑学まとめたった！')
('2554646961', 'Rory Morgan')
('2283418420', 'بيع متابعين خليجين')
('2465678160', '泣ける映画ランキング')
('635069556', 'Jolda Kotowski')
('2467729446', '誰も知らない原価の真実')
('2465494650', 'バスケ:スーパープレイ動画')
('2475504231', '松本潤応援画像bot')
('2508580920', 'アキ')
('2360614350', '@2intenfxもSPAM報告')
('2508797917', 'هند محمد')
```

CSC455Assignment5 Email:kying@mail.depaul.edu
Date: 5/31/2017

Problem 2b:

In [18]:
```
#Problem2-b
#Open assignment5.txt
file = open("assignment5.txt","r")
content=file.read()
content.strip()
lines=content.split('\n')
count=0

lst_created_at=[]
lst_id_str=[]
lst_text=[]
lst_source=[]
lst_in_reply_to_user_id=[]
lst_in_reply_to_screen_name=[]
lst_in_reply_to_status_id=[]
lst_retweet_count=[]
lst_contributors=[]
lst_user_id=[]

lst_user_name=[]
lst_user_screen_name=[]
lst_user_description=[]
lst_user_friends_count=[]

lst_geo=[]
```

In [70]:
```
#Problem2-b
#To load the contents from txt json format from 'lines' which stores all the tweet from Part1-b
for tweet in lines:
    try:
        obj=json.loads(tweet)
        #pprint.pprint(obj)
        lst_created_at.append(obj['created_at'])
        lst_id_str.append(obj['id_str'])
        lst_text.append(obj['text'])
        lst_source.append(obj['source'])
        lst_in_reply_to_user_id.append(obj['in_reply_to_user_id'])
        lst_in_reply_to_screen_name.append(obj['in_reply_to_screen_name']),
        lst_in_reply_to_status_id.append(obj['in_reply_to_status_id']),
        lst_retweet_count.append(obj['retweet_count'],)
        lst_contributors.append(obj['contributors']),
        lst_user_id.append(obj['user']['id'])


        lst_user_name.append(obj['user']['name'])
        lst_user_screen_name.append(obj['user']['screen_name'])
        lst_user_description.append(obj['user']['description'])
        lst_user_friends_count.append(obj['user']['friends_count'])


    except ValueError:#Error for damaged tweet
        count+=1 #Record the number of damaged tweet
        #err_lst.append(tweet)
        #Save all the damaged tweet into the Assignment5_errors.txt file
        with open('Assignment5_errors.txt', 'a') as f:
            print(tweet,file=f)
```

In [71]:
```
#Problem2-b
#Forming Data Frame from the Tweet Record
df_tweet=pd.DataFrame({
    "created_at":lst_created_at,
    "id_str":lst_id_str,
    "source":lst_source,
    "in_reply_to_user_id":lst_in_reply_to_user_id,
    "in_reply_to_screen_name":lst_in_reply_to_screen_name,
    "in_reply_to_status_id":lst_in_reply_to_status_id,
    "retweet_count":lst_retweet_count,
    "contributors":lst_contributors,
    "user_id":lst_user_id})

#Forming Data Frame from the User Record
df_user=pd.DataFrame({
    "user_id":lst_user_id,
    "user_name":lst_user_name,
    "user_screen_name":lst_user_screen_name,
    "user_description":lst_user_description,
    "user_friends_count":lst_user_friends_count})
```

CSC455Assignment5 Email:kying@mail.depaul.edu
Date: 5/31/2017

In [81]:
```
#Substitute all the empty or nan cell by None
df_tweet = df_tweet.where((pd.notnull(df_tweet)), None)
df_user = df_user.where((pd.notnull(df_user)), None)
```

In [82]:
```
#check the size of tweet data frame
df_tweet.shape
```

Out[82]: (9797, 9)

In [83]:
```
#check the size of user data frame
df_user.shape
```

Out[83]: (9797, 5)

In [84]:
```
#Problem2-b
#Check the Tweet Data Frame
df_tweet.head()
```

Out[84]:

in_reply_to_screen_name	in_reply_to_status_id	in_reply_to_user_id	retweet_count	source	user_id
None	None	None	0	<a href="https://mobile.twitter.com" rel="nofo...	367361405
None	None	None	0	<a href="http://makebot.sh" rel="nofollow">ほたる...	1605442621
None	None	None	0	<a href="http://admin.pure-c.jp/prog01_test3/t...	1685288690
None	None	None	0	<a href="http://www.feedblitz.com/f/f.fbz?help...	766497559
Tinkonsan	4.68537e+17	4.28261e+08	0	<a href="https://sites.google.com/site/tweentw...	229690290

In [85]:
```
#Problem2-b
#Check the User Data Frame
df_user.head()
```

Out[85]:

	user_description	user_friends_count	user_id	user_name	user_screen_name
0	None	543	367361405	gugusuarez	guadalupesuare4
1	黒バス/青エク/あの花/君僕/進撃/Free! etc.只今黒バス/ハイキュー!!が熱いです...	117	1605442621	ほたる	htr_ruby
2	人気・口コミランキングNo.1！当たると評判の電話占い・タレントや芸能人も御用達でテレビや雑...	0	1685288690	電話占いピュアリ	purely_c
3	None	89	766497559	Veronica	verodoglover
4	(´ω`) 東方やってます。たまにお絵かき。 天則のことつぶやくことが多いかも 規制用垢...	610	229690290	こさんつ	Cosan2

CSC455Assignment5 Email:kying@mail.depaul.edu
Date: 5/31/2017

Question 2-b: Final Answer

In [86]:
```
#Problem2-b: Show the user with the minimum "friends_count"
df_user[df_user['user_friends_count']==min(df_user['user_friends_count'])][['user_id',
                                                                            'user_name',
                                                                            'user_friends_count']]
```

Out[86]:

	user_id	user_name	user_friends_count
2	1685288690	電話占いピュアリ	0
16	1640311244	تابعو اليلو	0
17	2484449988	سعودية	0
68	469068614	おあおい	0
99	2498024720	【彼氏のあるべき姿】	0
101	2508769764	سعود السميران	0
150	2421149844	糞ばばぁ@ChasyacatFx	0
154	2505430447	Анна	0
252	2379017190	出雲ハルキ@ダイエット支援bot	0
253	2400158192	スゲェ雑学まとめたった！	0

CSC455Assignment5 Email:kying@mail.depaul.edu
Date: 5/31/2017

Problem 2-c: Add ‘geo’ attribute to the Table Tweet

In [23]:
```
#Problem2-c
#add 'geo' attribute to the Table Tweet
tw='''Create Table Tweet(

    created_at DATE,
    id_str VARCHAR(20),
    text VARCHAR(100),
    source VARCHAR(100),
    in_reply_to_user_id VARCHAR(20),
    in_reply_to_screen_name VARCHAR(20),
    in_reply_to_status_id VARCHAR(20),
    retweet_count INTEGER(5),
    contributors VARCHAR(10),
    user_id VARCHAR(20),
    geo integer(100),

    CONSTRAINT Tweet_FK
        FOREIGN KEY (user_id)
        REFERENCES User(id)
)
'''
```

In [25]:
```
#Problem2-c
import sqlite3
from sqlite3 import OperationalError
conn=sqlite3.connect('csc455_hw5.db')
c=conn.cursor()
#c.execute('Drop Table Tweet;')
c.execute(tw)


for tweet in lines:
    try:
        obj=json.loads(tweet)
        #pprint.pprint(obj)
        if obj['geo']==None:
            #INSERT to data to Table Tweet
            c.execute("INSERT INTO Tweet Values(?,?,?,?,?,?,?,?,?,?,?);",
            (obj['created_at'],
            obj['id_str'],
            obj['text'],
            obj['source'],
            obj['in_reply_to_user_id'],
            obj['in_reply_to_screen_name'],
            obj['in_reply_to_status_id'],
            obj['retweet_count'],
            obj['contributors'],
            obj['user']['id'],
            obj['geo']))
        else:
             #INSERT to data to Table Tweet
            c.execute("INSERT INTO Tweet Values(?,?,?,?,?,?,?,?,?,?,?);",
            (obj['created_at'],
            obj['id_str'],
            obj['text'],
            obj['source'],
            obj['in_reply_to_user_id'],
            obj['in_reply_to_screen_name'],
            obj['in_reply_to_status_id'],
            obj['retweet_count'],
            obj['contributors'],
            obj['user']['id'],
            str(obj['geo']['coordinates'])))
    except ValueError:#Error for damaged tweet
        pass
```

In [26]:
```
#Problem2-c
#Check the the content of Table Tweet
data=c.execute("select * from tweet").fetchall()
for line in data:
    print(line)
```

```
('Tue May 20 00:00:19 +0000 2014', '468541694288207874', 'la asusto a selena me dice es joda te vy a extrañar j
ajajajaja ni m fui pero ta vy a tener tiempo libre y todo wi', '<a href="https://mobile.twitter.com" rel="nofol
low">Mobile Web (M2)</a>', None, None, None, 0, None, '367361405', None)
('Tue May 20 00:00:19 +0000 2014', '468541694284017664', '【現在視聴中のアニメ】ソウルイーターノット!/僕らはみんな河合荘/
メカクシティアクターズ/一週間フレンズ。/悪魔のリドル/ハイキュー!! etc…', '<a href="http://makebot.sh" rel="nofollow">ほた
るボット</a>', None, None, None, 0, None, '1605442621', None)
('Tue May 20 00:00:19 +0000 2014', '468541694279835649', '瑤澗 嬉陽(ようじゅん きひ)先生が待機開始しました', '<a href=
"http://admin.pure-c.jp/prog01_test3/tweet_app/twitter_sample.php" rel="nofollow">purely_c tweet app</a>', None
, None, None, 0, None, '1685288690', None)
('Tue May 20 00:00:19 +0000 2014', '468541694292402176', '10 More FREE Pampers Gifts to Grow Points! via http:/
/t.co/3Sgq2UVq67 - We have a new Pampers Gifts ... http://t.co/DJa59HyRDi', '<a href="http://www.feedblitz.com/
f/f.fbz?help/default#twitter" rel="nofollow">FeedBlitz</a>', None, None, None, 0, None, '766497559', None)
('Tue May 20 00:00:19 +0000 2014', '468541694292410369', '@Tinkonsan まじで！ リアルマネー稼げるって聞いたんだけどどう
なん？？効率いい？', '<a href="https://sites.google.com/site/tweentwitterclient/" rel="nofollow">Tween</a>', '4282
60826', 'Tinkonsan', '468536777804767200', 0, None, '229690290', None)
('Tue May 20 00:00:19 +0000 2014', '468541694284009472', 'RT @bijyo_bijin: タイプだったらRT http://t.co/inZzxpl7v
h', '<a href="https://twitter.com/shinri_black" rel="nofollow">とびだせツイッターの森</a>', None, None, None, 0, No
ne, '2374270429', None)
('Tue May 20 00:00:19 +0000 2014', '468541694317568001', 'ことほのうみED～！', '<a href="http://twitter.com/downlo
ad/iphone" rel="nofollow">Twitter for iPhone</a>', None, None, None, 0, None, '1351058173', None)
('Tue May 20 00:00:19 +0000 2014', '468541694296600576', '隣のマンション付近におるし、変な道通って普段より高いし腹立つタク
```

In [27]:
```
#Problem2-c: Display Tweet with "geo" is NULL
data=c.execute("select * from Tweet where geo is NULL;").fetchall()
for line in data:
    print(line)
```

```
'106253406', None)
('Tue May 20 00:00:19 +0000 2014', '468541694280208384', 'RT @TwitWingsAr: هناك أكثر من 3442 صورة تم تداولها من
ضمن 9816 تغريدة في هاشتاق #سامي_الجابر ! http://t.co/5YFzycZ7kA', '<a href="http://twitlamp.com/" rel="nofollow
">GroupinApp</a>', None, None, None, 0, None, '1710706016', None)
('Tue May 20 00:00:19 +0000 2014', '468541694288207872', 'RT @hemmings_tho: Can twitter please stop fucking up
I swear #twitterfuckedupfollowparty', '<a href="http://twitter.com/download/android" rel="nofollow">Twitter for
Android</a>', None, None, None, 0, None, '1959608522', None)
('Tue May 20 00:00:19 +0000 2014', '468541694313779200', 'mas adeus', '<a href="http://twitter.com/download/iph
one" rel="nofollow">Twitter for iPhone</a>', None, None, None, 0, None, '2425496601', None)
('Tue May 20 00:00:19 +0000 2014', '468541694305370113', 'RT @0ntiverosCande: Hola agu diosa', '<a href="https:
//mobile.twitter.com" rel="nofollow">Mobile Web (M2)</a>', None, None, None, 0, None, '1587544904', None)
('Tue May 20 00:00:19 +0000 2014', '468541694292815872', '#LHHATL', '<a href="http://twitter.com/download/andro
id" rel="nofollow">Twitter for Android</a>', None, None, None, 0, None, '1978399038', None)
('Tue May 20 00:00:19 +0000 2014', '468541694296985601', 'Não aguento mais limpar, limpei toda a casa. Foda é p
ensar que não vai demorar muito pra ela ficar toda suja.', '<a href="http://twitter.com/download/android" rel="
nofollow">Twitter for Android</a>', None, None, None, 0, None, '183066547', None)
('Tue May 20 00:00:19 +0000 2014', '468541694288592898', "If you're one of those people that sends the same sna
pchat out personally and on your story I hate you.", '<a href="http://twitter.com/download/iphone" rel="nofollo
w">Twitter for iPhone</a>', None, None, None, 0, None, '1167821629', None)
('Tue May 20 00:00:19 +0000 2014', '468541694280204288', 'That awkward moment when you snapchat someone only to
```

CSC455Assignment5 Email:kying@mail.depaul.edu
Date: 5/31/2017

Problem 2-d: Add “geo” attribute to the Data Frame

In [1]:
```
#Import pandas and numpy library
import pandas as pd
import numpy as np
import urllib.request as urlreq
import lxml.etree as et
import json
import pprint
```

In [4]:
```
#Open assignment5.txt
file = open("assignment5.txt","r")
content=file.read()
content.strip()
lines=content.split('\n')
count=0
err_lst=[]
```

In [28]:
```
lst_created_at=[]
lst_id_str=[]
lst_text=[]
lst_source=[]
lst_in_reply_to_user_id=[]
lst_in_reply_to_screen_name=[]
lst_in_reply_to_status_id=[]
lst_retweet_count=[]
lst_contributors=[]
lst_user_id=[]

lst_user_name=[]
lst_user_screen_name=[]
lst_user_description=[]
lst_user_friends_count=[]

lst_geo=[]
```

Create the Data Frame with ‘geo’ attribute:

In [4]:
```
#Problem2-d: Display Tweet with "geo" is NULL
#All the tweet stores in "lines" from Part 1-b

for tweet in lines:
    try:
        obj=json.loads(tweet)
        #pprint.pprint(obj)
        lst_created_at.append(obj['created_at'])
        lst_id_str.append(obj['id_str'])
        lst_text.append(obj['text'])
        lst_source.append(obj['source'])
        lst_in_reply_to_user_id.append(obj['in_reply_to_user_id'])
        lst_in_reply_to_screen_name.append(obj['in_reply_to_screen_name']),
        lst_in_reply_to_status_id.append(obj['in_reply_to_status_id']),
        lst_retweet_count.append(obj['retweet_count'],)
        lst_contributors.append(obj['contributors']),
        lst_user_id.append(obj['user']['id'])
        #Store the Coordinates to the geo cell if 'geo' is NOT None
        if obj['geo']!=None:
            lst_geo.append(str(obj['geo']['coordinates']))
        else:
            lst_geo.append(obj['geo'])

    except ValueError:#Error for damaged tweet
        pass
```

In [5]:
```
#Forming Data Frame from the Tweet Record
df_tweet=pd.DataFrame({
    "created_at":lst_created_at,
    "id_str":lst_id_str,
    "text":lst_text,
    "source":lst_source,
    "in_reply_to_user_id":lst_in_reply_to_user_id,
    "in_reply_to_screen_name":lst_in_reply_to_screen_name,
    "in_reply_to_status_id":lst_in_reply_to_status_id,
    "retweet_count":lst_retweet_count,
    "contributors":lst_contributors,
    "user_id":lst_user_id,
    "geo":lst_geo})
```

In [6]:
```
#Replace null type to None
df_tweet2=df_tweet.where((pd.notnull(df_tweet)), None)
#Check the None type
type(df_tweet2.ix[1][0])
```

Out[6]: NoneType

In [8]:
```
#Display the tweet for 'geo' is NOT None
df_tweet2[df_tweet2['geo'].notnull()]
```

Out[8]:

	contributors	created_at	geo	id_str	in_reply_to_screen_name	in_reply_to_status_id	in_reply_to_user_id	retwe
49	None	Tue May 20 00:00:19 +0000 2014	[-22.9708306, -42.0213279]	468541694313771008	None	None	None	0
105	None	Tue May 20 00:00:20 +0000 2014	[-33.9152902, -60.5670637]	468541698474524672	juanfrahoms	4.68541e+17	6.21562e+08	0
219	None	Tue May 20 00:00:22 +0000	[41.5471493, -81.606624]	468541706892484608	None	None	None	0

Problem 2-d: Final Answer

In [10]:
```
#Question 2-d: Final Answer - Display the tweet with
df_tweet2[df_tweet2['geo'].isnull()][['id_str',
                                      'geo',
                                      'text',
                                      'user_id']]
```

Out[10]:

	id_str	geo	text	user_id
0	468541694288207874	None	la asusto a selena me dice es joda te vy a ext...	367361405
1	468541694284017664	None	【現在視聴中のアニメ】ソウルイーターノット!/僕らはみんな河合荘/メカクシティアクターズ/一...	1605442621
2	468541694279835649	None	瑤澍 嬉陽(ようじゅん きひ)先生が待機開始しました	1685288690
3	468541694292402176	None	10 More FREE Pampers Gifts to Grow Points! via...	766497559
4	468541694292410369	None	@Tinkonsan まじで！ リアルマネー稼げるって聞いたんだけどどうなん？？効率いい？	229690290
5	468541694284009472	None	RT @bijyo_bijin: タイプだったらRT http://t.co/inZzxp17vh	2374270429
6	468541694317568001	None	ことほのうみED～！	1351058173
7	468541694296600576	None	隣のマンション付近におるし、変な道通って普段より高いし腹立つタクシー	141120654
8	468541694292414465	None	Phoonk - http://t.co/GLclulDbLu	2478849157
9	468541694279819264	None	Just posted a photo http://t.co/GAnutMRQJZ	1297356302

CSC455Assignment5 Email:kying@mail.depaul.edu
Date: 5/31/2017

Problem 3-a:
Longest Text in Tweet

```
In [59]: #Problem 3-a
         #Check the length of Text in Tweet in Descending Order
         c.execute("select text from Tweet order by length(text) DESC").fetchall()
```

```
Out[59]: [('RT @Rileymonroeee: When you start catching feelings , becoming attached , wanting more &lt;&lt;&gt;&gt;&gt;&
         lt;&lt;&lt;&lt;&gt;&lt;&gt;&gt;&gt;&lt;&lt;&gt;&gt; 😩😍🙇💞',),
          ('RT @damnhxrold: &gt; rt for this perfect human being \n\n&gt; follow me\n\n&gt; follow whoeved rts\n\n&gt; g
         ain\n\n#twitterfuckedupfollowparty http://t.co/VPGE…',),
          ('"@___NotoriousE: pizza hut pizza from target is so much better then the one in the store!"farreal, the targe
         t by that walmart shit be Al&gt;&gt;&gt;&gt;',),
          ('I stand against Keystone XL &amp; w/ farmers &amp; tribal members in DC today fighting for their land, water
         &amp; air. #NoKXL  http://t.co/neKR8wm9RO',),
          ('RT @kusuma_putri99: &gt;&gt;&gt; "@adhiet_youdie: apa krn takut kalau molor dr jadwal yg tetapkan di dlm kon
         trak? Trus kalau molor dr jadwal apa kon…',),
          ('Cheer up! Tuesday is almost here &amp; your Fish n Chips, New England Clam Chowder &amp; a Ice Cold Beer wil
         l be ready for you to get starting @11am',),
          ("RT @jamesmaslow: Freestyle &amp; Tango tonight on the #dwts finale! Might even see some of my friends stop b
         y...don't forget to watch &amp; vote! @…",),
          ('RT @SwesterrMai: Happy Birthday to my baby @_imaji enjoy ur day. I love u to the Galaxy &amp;&amp; Beyond. N
         ow you can go clubbing with us. Will ma…',),
          ('RT @Cimorelliband: Follow @TheNickRadio &amp; @Maxwellshouse to hear our chat about @awesomenesstv TONIGHT a
         t 7 PM (EST) here! &gt; http://t.co/9I…',),
          ('RT @DenjenDaphne: Diduga Berbuat Mesum, Caleg Golkar Batanghari Ditangkap #GalaSV10 http://t.co/YqSHpFQXvA
         VIDEO download di &gt;&gt;http://t.co/…',),
          ('RT @hanitweet: "모든 것을 경제 성장으로 환원시킨 결과, 우리는 지금 '근대의 비극'을 겪고 있다"\n김덕영 동일 카셀대 교수가 &lt;환원근대&
```

```
In [57]: #Find all the longest Tweet Text
         c.execute("select text from Tweet where length(text)>=(select max(length(text)) from Tweet);").fetchall()
```

```
Out[57]: [('RT @Rileymonroeee: When you start catching feelings , becoming attached , wanting more &lt;&lt;&gt;&gt;&gt;&
         lt;&lt;&lt;&lt;&gt;&lt;&gt;&gt;&gt;&lt;&lt;&gt;&gt; 😩😍🙇💞',)]
```

Shortest Text in Tweet:

```
In [58]: #Problem 3-a
         #Check the length of text in the Tweet in Ascending order
         c.execute("select text from Tweet order by length(text) ASC").fetchall()
```

```
Out[58]: [('😈',),
          ('😊',),
          ('.',),
          ('.',),
          ('😏',),
          ('ﾌ',),
          ('😻',),
          ('🍔',),
          ('😁',),
          ('ﾉ',),
          ('😔',),
          ('😔😔',),
          ('On',),
          ('～♪',),
          ('••••',),
          ('子種',),
          ('確定',),
          ('21',),
          ('出勤',),
          ('😷😔',),
          ('🚗💨',),
```

```
In [56]: #Find all the Shortest Tweet Text
         c.execute("select text from Tweet where length(text)<=(select min(length(text)) from Tweet);").fetchall()
```

```
Out[56]: [('😈',),
          ('😊',),
          ('.',),
          ('.',),
          ('😏',),
          ('ﾌ',),
          ('😻',),
          ('🍔',),
          ('😁',),
          ('ﾉ',),
          ('😔',)]
```

CSC455Assignment5 Email:kying@mail.depaul.edu
Date: 5/31/2017

Problem 3-b: Find the top 5 most frequent terms from the 'text' of the tweets
Note: The original tweet2 data frame is formed from Problem 2-d

In [206]:
```
#Problem 3-b: the df_tweet2 data frame is built from part 2-d
#Display the text string from the dataframe's text column
for line in df_tweet2['text']:
    print(line)
```

```
[自動]【時報】09:00をお知らせします
GO HABS GO 💙🤍❤️
なんか提督でもアニオタでもフォロワーでもない人からRT来た、しかもいかにもスパムの予感...
ただいまの時刻は9時です。
@fawazAlRuwaili جداااا طيوف شرايح بسابق فواز البديد @tooshafx 😅
RT @wadai_free_app: 200万DLを突破した無料マンガアプリ♪
毎日無料で漫画読み放題ｗｗｗ
これで漫画買う必要ないｗ

iPhoneもAndroidもこちら
→http://t.co/wF4c3RYEuX http://t.co/5A7q87yJmY
I miss my mom 😝
ポケットモンスター

理系選択 / 文系選択 / 留年(マイナーチェンジ)
BBCSって何？
あーーー最近めっちゃだるいしきついし、もぅゃだ。。。リスカしょ。。。状態だったけど（笑）
やーっと浮上できた！とりあえず存在無視してた課題やって溜まってるDVDみて
今日の2限は遅刻しよう。
明日から遅刻しない笑
```

In [207]:
```
#Problem 3-b
#Form a short article ignore list
ignore_lst=['an','a','the','this','that','with','on','those','they','them','http','&amp;','your','have','our'
            ,'hers','when','para','just','don\'t','will','know','about','from','what']
text_dict={}

#for loop to count the length of text tweet
for line in df_tweet2['text']:
    line=line.split(' ')
    #for loop to count the words the tweet
    for word in line:
        #Filter out the word is in the ignore list
        if word not in ignore_lst:
            #filter out words' length less or equal to 4
            if (len(word) >=4):
                #Increment the count for every word found in the tweet, except those in the ignore lost
                if word in text_dict:
                    text_dict[word] +=1
                else:
                    #If the word is new found from tweet, start that word count from ZERO
                    text_dict[word] = 0
```

In [209]:
```
#Problem 3-b
#find the top 5 most Frequency appearing in the text

#Store the dict value to a list
lst_value=list(set(text_dict.values()))
#Put the list in Descending order
lst_value.sort(reverse=True)
#Extract the top 5 frequency into a new list
top5=lst_value[0:5]


#use for loop to find out the top 5 KEY word
for word, length in text_dict.items():
    #print the top five terms and frequency of the terms
    if length in top5:
        print("The most frequent word is: ",word)
        print("The Frequency is: ",str(length))
```

```
The most frequent word is:  like
The Frequency is:  349
The most frequent word is:  love
The Frequency is:  309
The most frequent word is:  como
The Frequency is:  151
The most frequent word is:  🔥
The Frequency is:  177
The most frequent word is:  FOLLOW
The Frequency is:  221
```