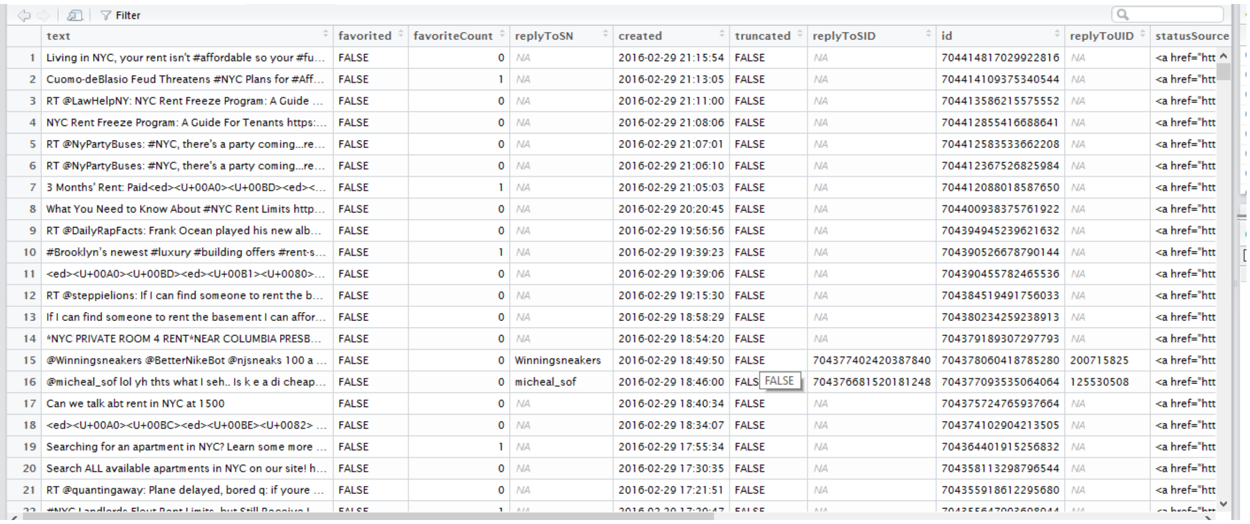
# Problem 1: Data acquisition

Problem Statement: Data collection is an important step that often is quite time consuming and constrained by regulations and privacy and security issues. For instance in order to collect data from patients in a clinical trial, a researcher in an organization has to get prior approval and certification from Internal Review Board (IRB) of the organization. There are many regulatory laws governing what data you collect about common citizens. Students ought to be sensitive to all these when collecting data. The other challenge in data collection is that data comes in different formats (text, html, txt, csv, json etc.) and feature widely varying access methods (web URL, api, hdfs, sqldb etc.). We will learn about this by working on few representative methods for data acquisition given in the handout: http://www.cse.buffalo.edu/~bina/cse487/spring2016/Lectures/RHandout1.pdf. To that list of methods add an approach for reading json data. For example twitter data is published as JSON objects.

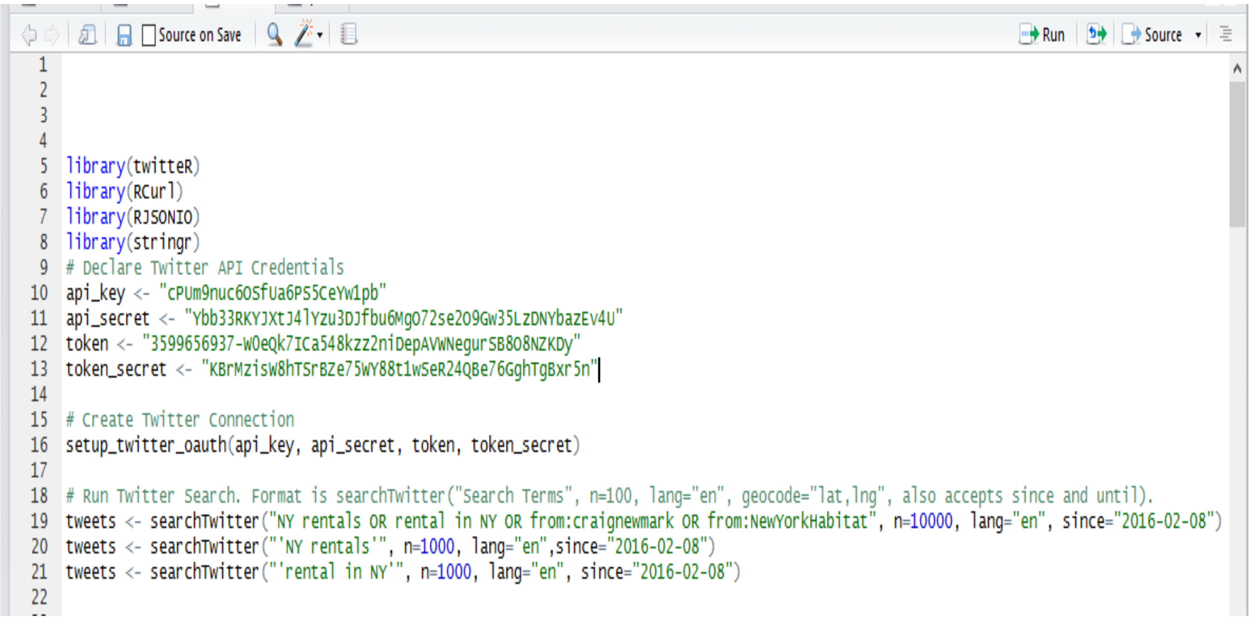
In this project, I have collected data from "**twitter**" on the topic "**Rental Apartments**". I choose "**Twitter**" because of the abundant amount of data that could be acquired on any topic. The topic **"Rental Apartments"** was chosen because it can be used as the data for Problem 4 and 5. The data was collected for a period of approximately **1 month**. This gives us enough data (tweets) using which appropriate analysis can be made.



| | text | favorited | favoriteCount | replyToSN | created | truncated | replyToSID | id | replyToUID | statusSource |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Living in NYC, your rent isn't #affordable so your #fu... | FALSE | 0 | NA | 2016-02-29 21:15:54 | FALSE | NA | 704414817029922816 | NA | <a href="htt |
| 2 | Cuomo-deBlasio Feud Threatens #NYC Plans for #Aff... | FALSE | 1 | NA | 2016-02-29 21:13:05 | FALSE | NA | 704414109375340544 | NA | <a href="htt |
| 3 | RT @LawHelpNY: NYC Rent Freeze Program: A Guide ... | FALSE | 0 | NA | 2016-02-29 21:11:00 | FALSE | NA | 704413586215575552 | NA | <a href="htt |
| 4 | NYC Rent Freeze Program: A Guide For Tenants https:... | FALSE | 0 | NA | 2016-02-29 21:08:06 | FALSE | NA | 704412855416688641 | NA | <a href="htt |
| 5 | RT @NyPartyBuses: #NYC, there's a party coming...re... | FALSE | 0 | NA | 2016-02-29 21:07:01 | FALSE | NA | 704412583533662208 | NA | <a href="htt |
| 6 | RT @NyPartyBuses: #NYC, there's a party coming...re... | FALSE | 0 | NA | 2016-02-29 21:06:10 | FALSE | NA | 704412367526825984 | NA | <a href="htt |
| 7 | 3 Months' Rent: Paid<ed><U+00A0><U+00BD><ed><... | FALSE | 1 | NA | 2016-02-29 21:05:03 | FALSE | NA | 704412088018587650 | NA | <a href="htt |
| 8 | What You Need to Know About #NYC Rent Limits http... | FALSE | 0 | NA | 2016-02-29 20:20:45 | FALSE | NA | 704400938375761922 | NA | <a href="htt |
| 9 | RT @DailyRapFacts: Frank Ocean played his new alb... | FALSE | 0 | NA | 2016-02-29 19:56:56 | FALSE | NA | 704394945239621632 | NA | <a href="htt |
| 10 | #Brooklyn's newest #luxury #building offers #rent-s... | FALSE | 1 | NA | 2016-02-29 19:39:23 | FALSE | NA | 704390526678790144 | NA | <a href="htt |
| 11 | <ed><U+00A0><U+00BD><ed><U+00B1><U+0080>... | FALSE | 0 | NA | 2016-02-29 19:39:06 | FALSE | NA | 704390455782465536 | NA | <a href="htt |
| 12 | RT @steppielions: If I can find someone to rent the b... | FALSE | 0 | NA | 2016-02-29 19:15:30 | FALSE | NA | 704384519491756033 | NA | <a href="htt |
| 13 | If I can find someone to rent the basement I can affor... | FALSE | 0 | NA | 2016-02-29 18:58:29 | FALSE | NA | 704380234259238913 | NA | <a href="htt |
| 14 | *NYC PRIVATE ROOM 4 RENT*NEAR COLUMBIA PRESB... | FALSE | 0 | NA | 2016-02-29 18:54:20 | FALSE | NA | 704379189307297793 | NA | <a href="htt |
| 15 | @Winningsneakers @BetterNikeBot @njsneaks 100 a ... | FALSE | 0 | Winningsneakers | 2016-02-29 18:49:50 | FALSE | 704377402420387840 | 704378060418785280 | 200715825 | <a href="htt |
| 16 | @micheal_sof lol yh thts what I seh.. Is k e a di cheap... | FALSE | 0 | micheal_sof | 2016-02-29 18:46:00 | FALSE | 704376681520181248 | 704377093535064064 | 125530508 | <a href="htt |
| 17 | Can we talk abt rent in NYC at 1500 | FALSE | 0 | NA | 2016-02-29 18:40:34 | FALSE | NA | 704375724765937664 | NA | <a href="htt |
| 18 | <ed><U+00A0><U+00BC><ed><U+00BE><U+0082> ... | FALSE | 0 | NA | 2016-02-29 18:34:07 | FALSE | NA | 704374102904213505 | NA | <a href="htt |
| 19 | Searching for an apartment in NYC? Learn some more ... | FALSE | 1 | NA | 2016-02-29 17:55:34 | FALSE | NA | 704364401915256832 | NA | <a href="htt |
| 20 | Search ALL available apartments in NYC on our site! h... | FALSE | 0 | NA | 2016-02-29 17:30:35 | FALSE | NA | 704358113298796544 | NA | <a href="htt |
| 21 | RT @quantingaway: Plane delayed, bored q: if youre ... | FALSE | 0 | NA | 2016-02-29 17:21:51 | FALSE | NA | 704355918612295680 | NA | <a href="htt |

Fig: Sample Data

Instead of using a python program to retrieve data and then an R script to convert the data to json format, I have used **"twitteR"** package of R. This package is a convenient way of retrieving tweets and converting it to json format. The sample code used to retrieve tweets is shown in the figure below.



```r




library(twitteR)
library(RCurl)
library(RJSONIO)
library(stringr)
# Declare Twitter API Credentials
api_key <- "cPUm9nuc6OSfUa6PS5CeYw1pb"
api_secret <- "Ybb33RKYJXtJ4lYzu3DJfbu6MgO72se2O9Gw35LzDNYbazEv4U"
token <- "3599656937-WOeQk7ICa548kzz2niDepAVWNegurSB8O8NZKDy"
token_secret <- "KBrMzisW8hTSrBZe75WY88t1wSeR24QBe76GghTgBxr5n"

# Create Twitter Connection
setup_twitter_oauth(api_key, api_secret, token, token_secret)

# Run Twitter Search. Format is searchTwitter("Search Terms", n=100, lang="en", geocode="lat,lng", also accepts since and until).
tweets <- searchTwitter("NY rentals OR rental in NY OR from:craignewmark OR from:NewYorkHabitat", n=10000, lang="en", since="2016-02-08")
tweets <- searchTwitter("'NY rentals'", n=1000, lang="en",since="2016-02-08")
tweets <- searchTwitter("'rental in NY'", n=1000, lang="en", since="2016-02-08")

```

Fig: Shows the code for collecting data using “twitter” Package.

The command “library(twitteR)” is used to load the library and all the authorization keys acquired from the Twitter Api is entered. To run the twitter search, the following command is used.

**tweets <- searchTwitter("'NY rentals'", n=1000, lang="en",since="2016-02-08")**

In the above command, the retrieved tweets are stored in “tweets” variable, **“NY rentals**” is the search term, “**n**” is equated to 1000, to retrieve **1000 tweets** from the date **2016-02-08** in **English.**

**twt<-toJSON(tweets)**
**write(twt,file="tweet.json")**

The above command is used to write the tweets in to a json file called “tweet.json”. The “toJSON” function converts the data in a variable to Json and then it is written to a file.

The file RP1pgolluin.R is an R script which has retrieves the tweets as well as converts it into a JSON file.