CSE 587 – Data Intensive Computing

Problem 1: Data acquisition

Author: Bhavika Jain
bhavikap@buffalo.edu

Data Collection in R:

In this problem we were required to understand the process of data collection from an external source like Twitter, Facebook etc. in R. For this purpose, I used Twitter as my external source and collected few samples of tweets for a period of 1 week (28-02-2016 – 05-03-2016) to understand the process of data collection in R.

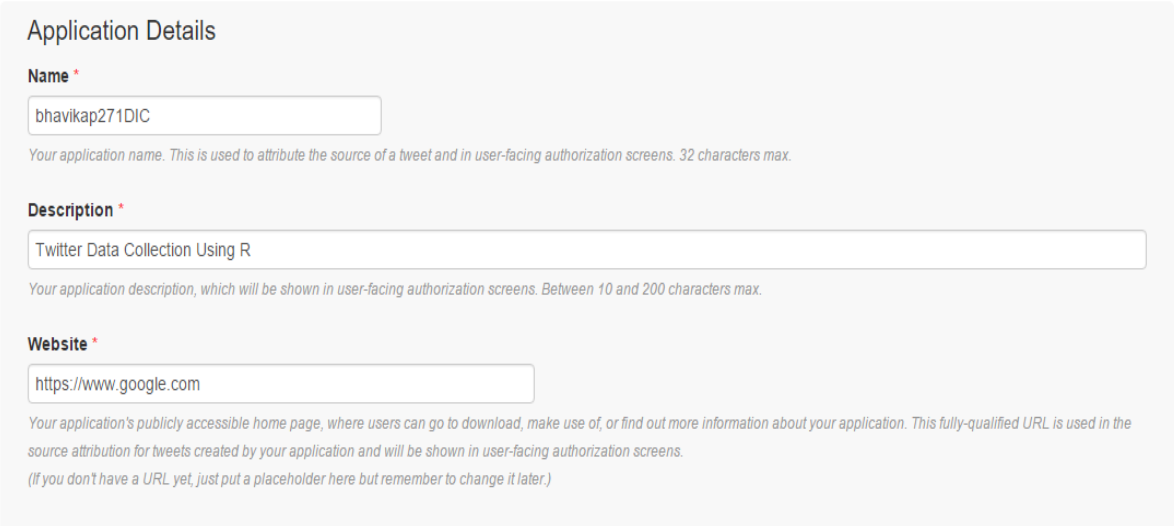
How to get started with Twitter Search API and R?

1) Create a new application on Twitter:

To able to query the Twitter Search API and import data into R, I created an account on Twitter.

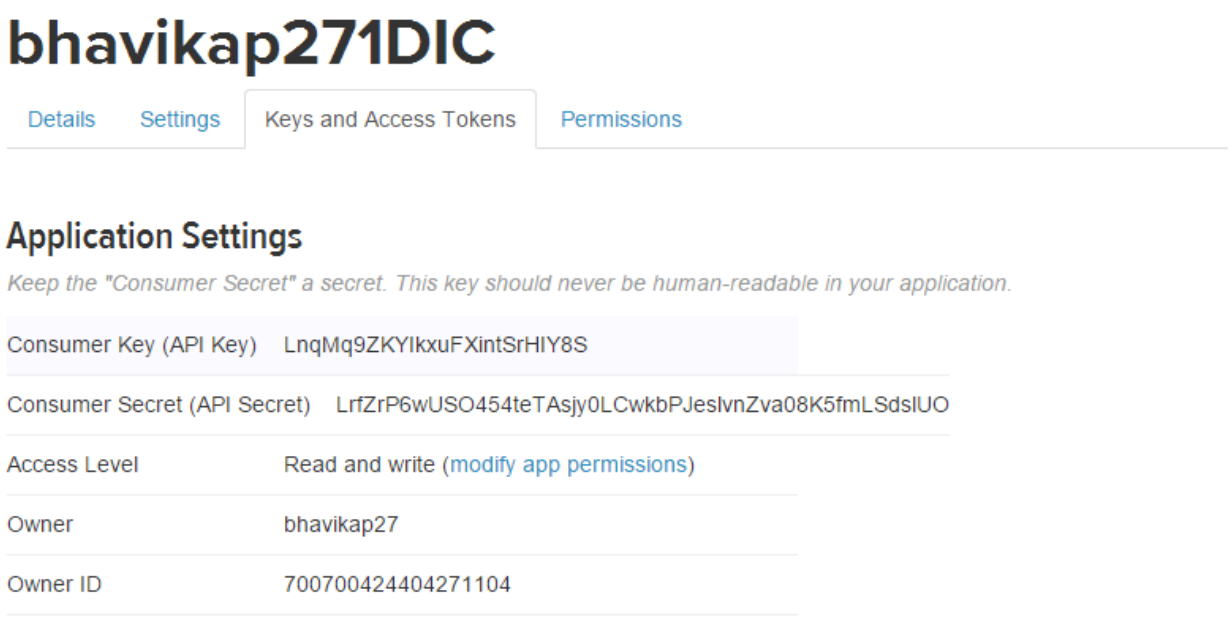
Below is the process of creating an application on Twitter:

i) Login to Twitter Developers site.
ii) Create a new application.
iii) Enter the application details as shown below.



i) Once the application is created, following API credentials (keys and access tokens) as shown below are given to set up a connection to the Twitter Search API.
iv) For connecting to Twitter Search API, credentials for **Consumer Key, Consumer Secret Key, Access Token, and Access Key** are needed.

Author: Bhavika Jain
bhavikap@buffalo.edu



2) Install Twitter Packages in R-Studio :

The following packages were required to be installed.

a) TwitteR: Provides us with an API of several methods which can be used to collect tweets from twitter.
b) Jsonlite: Provides us with an API of several methods helpful for exporting/importing tweets to/from json file.
c) ROauth: Provides an interface to the OAuth 1.0 specification allowing users to authenticate via OAuth to the server of their choice.

Author: Bhavika Jain
bhavikap@buffalo.edu

3) R-Script for Collecting Tweets :

Once the packages are installed and the credentials ready, we can move to the next step of writing the R script. Below I have mentioned what each of the method in the script does:

Methods used from *twitteR* package:

setup_twitter_oauth: Using this method we set up a connection to Twitter API. The methods needs the OAuth credentials (**Consumer Key, Consumer Secret Key, Access Token, and Access Key**) we received after we created a new application on Twitter Developer Site. Using these credentials it will make a direct HTTP connection to the Twitter API.

searchTwitter: This method provides the interface for collecting tweets from Twitter. The method takes several parameters, few of them mentioned below which I had used in my script also for collecting required tweets. It will return a list of tweets which satisfy the conditions given.

a) searchString = Using this value, the method returns all the tweets which contain this keyword or hashtag. We can even search multiple keywords using '+' operator between individual keywords.
b) N = Indicates the number of tweets to be collected.
c) Lang = Using this we can collect tweets in different languages like English, Spanish, French etc. For this assignment I have restricted the language to English.

Below is the sample structure of a tweet returned by searchTwitter method for ***keyword = "election":*** The tweet has several fields which provides us details like creation date/time, source, tweet text etc.

{

"text": "my favorite thing about this election cycle is that US Americans have learned two new words: bloviate, and demagogue",

"favorited": false,

"favoriteCount": 0,

"created": "2016-03-05 15:49:18",

"truncated": false,

"id": "706144563128365056",

"statusSource": "Hootsuite",

"screenName": "egoistetx",

"retweetCount": 0,

"isRetweet": false,

"retweeted": false

}

strip_retweets: This method removes the Retweets returned by the searchTwitter method.

Author: Bhavika Jain
bhavikap@buffalo.edu

twListToDF: This method converts the list of tweets returned by search Twitter method to an R data Frame. Data Frame is later used for exporting the tweets to a json file.

Methods used from *jsonlite* package:

toJSON: This method converts an R data Frame returned by twListToDF method to a Json object. Also by setting pretty = TRUE, it formats the Json object for display.

write: This method writes the json object returned by toJSON method in json file. It needs the json object to be written in a file and a filename as a parameter to create a json file.

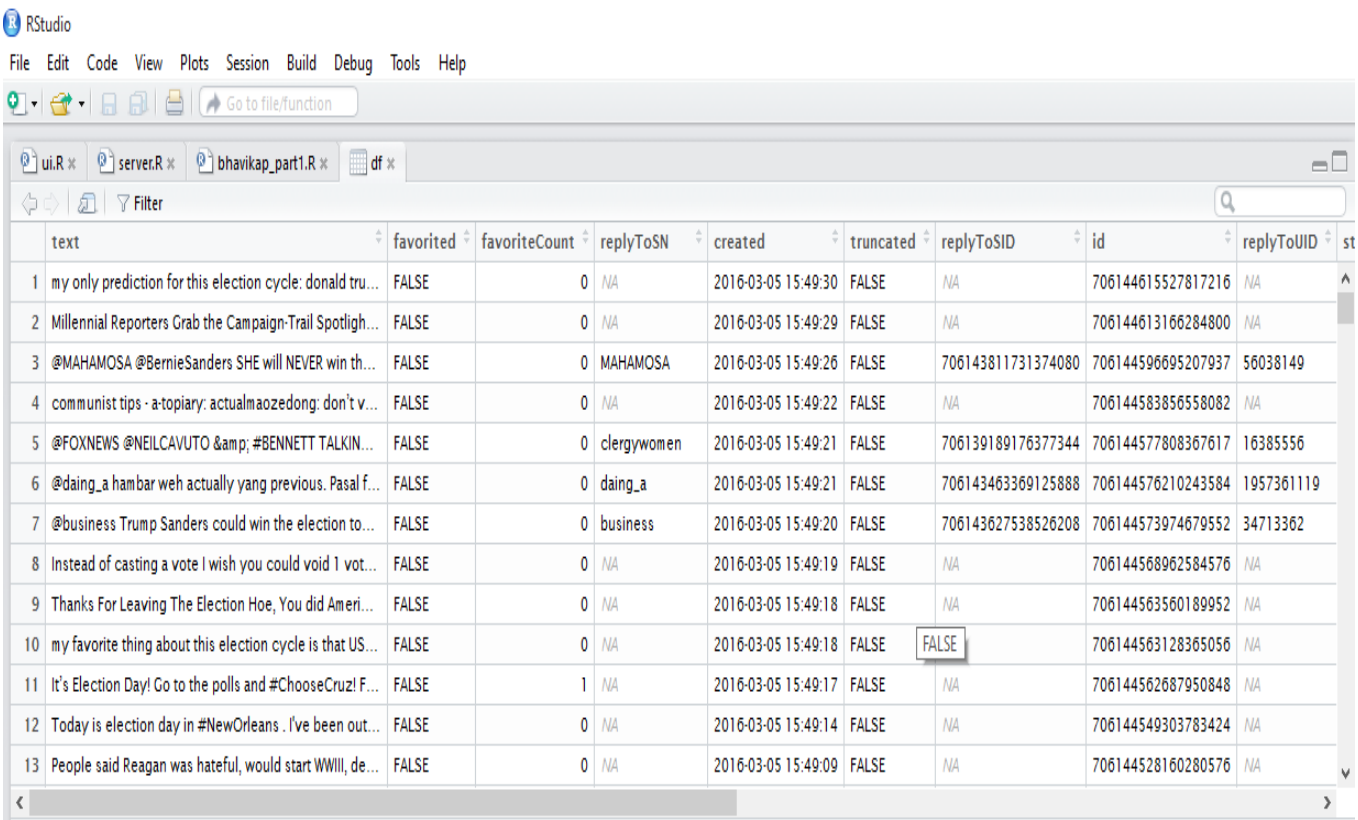
fromJson: This method converts a Json object into an R data Frame. The method takes as input the filename and return an R object.

In the first R script, I have used all of the above methods for doing the task of exporting tweets in a json file.

In the second R script, I had to import the json file back, created by executing the first script into an R object i.e. data Frame and display it. For doing this, I used the ***fromJson*** method from the Jsonlite package. The method takes the input parameter as the filename and converts it into an R data Frame. This data Frame can then later be used for doing any kind of analysis on the tweets.

R Data Structure:

The data structure that I used in this problem is Data Frame. It is used for storing data tables consisting of length of equal vectors. Below screenshot shows how tweets are stored inside data frame.



	text	favorited	favoriteCount	replyToSN	created	truncated	replyToSID	id	replyToUID	st
1	my only prediction for this election cycle: donald tru...	FALSE	0	NA	2016-03-05 15:49:30	FALSE	NA	706144615527817216	NA	
2	Millennial Reporters Grab the Campaign-Trail Spotligh...	FALSE	0	NA	2016-03-05 15:49:29	FALSE	NA	706144613166284800	NA	
3	@MAHAMOSA @BernieSanders SHE will NEVER win th...	FALSE	0	MAHAMOSA	2016-03-05 15:49:26	FALSE	706143811731374080	706144596695207937	56038149	
4	communist tips - a-topiary: actualmaozedong: don't v...	FALSE	0	NA	2016-03-05 15:49:22	FALSE	NA	706144583856558082	NA	
5	@FOXNEWS @NEILCAVUTO & #BENNETT TALKIN...	FALSE	0	clergywomen	2016-03-05 15:49:21	FALSE	706139189176377344	706144577808367617	16385556	
6	@daing_a hambar weh actually yang previous. Pasal f...	FALSE	0	daing_a	2016-03-05 15:49:21	FALSE	706143463369125888	706144576210243584	1957361119	
7	@business Trump Sanders could win the election to...	FALSE	0	business	2016-03-05 15:49:20	FALSE	706143627538526208	706144573974679552	34713362	
8	Instead of casting a vote I wish you could void 1 vot...	FALSE	0	NA	2016-03-05 15:49:19	FALSE	NA	706144568962584576	NA	
9	Thanks For Leaving The Election Hoe, You did Ameri...	FALSE	0	NA	2016-03-05 15:49:18	FALSE	NA	706144563560189952	NA	
10	my favorite thing about this election cycle is that US...	FALSE	0	NA	2016-03-05 15:49:18	FALSE	FALSE	706144563128365056	NA	
11	It's Election Day! Go to the polls and #ChooseCruz! F...	FALSE	1	NA	2016-03-05 15:49:17	FALSE	NA	706144562687950848	NA	
12	Today is election day in #NewOrleans . I've been out...	FALSE	0	NA	2016-03-05 15:49:14	FALSE	NA	706144549303783424	NA	
13	People said Reagan was hateful, would start WWIII, de...	FALSE	0	NA	2016-03-05 15:49:09	FALSE	NA	706144528160280576	NA	

Author: Bhavika Jain
bhavikap@buffalo.edu

4) **Conclusion :**

By working on this assignment, I learnt several new things, most significant of them mentioned below:

- R programming language.
- Working on R Studio.
- Twitter Search API for collecting tweets.
- Importing/Exporting data in R.
- Twitter API packages in R.
- Creating application in Twitter.
- Twitter OAuth setup.
-

I have attached the json file created by executing first script containing sample of tweets collected by me.

P.S: Please set the working directory in the script for OAuth setup and importing/exporting json data.

Author: Bhavika Jain
bhavikap@buffalo.edu