CSE587: DATA INTENSIVE COMPUTING

PROBLEM1: DATA ACQUISITION

For this problem, I collected Twitter Data using both stream API and Rest API.

STREAM API

The Streaming APIs give developers low latency access to Twitter's global stream of Tweet data. An example of Stream API, which can be used with R Language, is StreamR API.

streamR API

This package includes a series of functions that give R users access to Twitter's Streaming API, as well as a tool that parses the captured tweets and transforms them in R data frames, which can then be used in subsequent analyses. streamR supports authentication via OAuth and the ROAuth package, as well as basic authentication with screen name and password.

Therefor the following packages have to be installed for streamR API:

1. library(streamR)
2. library(ROAuth)

The packages RJSONIO and rjson are used for importing json input into R.

The script for accessing twitter via streamR API includes signing up as a developer on Twitter and then acquiring a access key, access token, consumer key and consumer token.

The access key and access token authenticate the app to access twitter while the consumer key and consumer token let the app access the user's timeline data.

When first creating the OAuth token, it is necessary to execute the OAuth handshake.

After this , the filterstream() function can be used to scrape Twitter data in real time.
The tweets attribute specifies the maximum number of tweets to be collected and track has the keywords to search for. We also have a timeout attribute which says when the stream connection will close.

There are two additional parameters follow, which can specify the subset of users for whom the tweets are to be collected and the location parameter which helps in scraping location-oriented tweets.

I have also illustrated the samplestream() function that allows the user to capture a small random sample of tweets that are being sent at the moment.

After the tweets are captured, the parsetweet() function is used to read the captured tweets from the file in which they were stored in disk and are converted into a dataframe in the R environment.

The gettimeline() function allows to capture tweets from the timeline of a given user.

Next we look at twitter API.

REST API

REST Api typically runs over the HTTP protocol and has a stateless existence. It underlies on the principle that the interactions between clients and servers can be enhanced by having a limited number of operations

twitteR API

twitteR is a REST API which can be used to scrape data from Twitter
The following packages are installed for twitteR

1. library(twitter)

The setup_twitter_oauth () function sets up the OAuth connection required to collect data from Twitter using the API key from the developer account.
The main function used to search tweets in twitter API is searchTwitter() function. For repeated search and import to R, we use a for loop involving the searchTwitter() function.

The limitations of REST API over the Streaming API is that the Streaming API has less overhead than the REST API and also the Stream API is used to capture realtime data which is published at that moment without any latency.

DATA SET INFORMATION:

Data Source: For this problem, the data source is Twitter.
Period of Time: 2 days
Data Format :JSON