# Problem 1: Data acquisition

## Problem Solution:

### Collecting Data:

The data from twitter is collected using “twitter” R Package which requires us to setup OAuth authorization to twitter RestAPI and then the tweets are collected using ‘*searchTwitter’* function. The collected tweets may have duplicates, we can eliminate the duplicate tweets by striping them based on the field “**ReplytoID**”. The **strip\_retweets** function from **twitteR** takes care of removing duplicates. Now after removing duplicates we are saving the data to JSON file by converting it into **JSON** using “**jsonlite**” R package.

#### Formatting and Summarizing Data:

The first step in this process is to load the data collected by earlier script. After collecting the data, we convert the data to R Data frame for flexibility to handle the structure of tweet data. Now we also “created” column to remove the time and only use date. After formatting created field we summarize the data per day using **ddply** package. This summarized data can be used to analyze data later in problem4

Files submitted for Problem 1 Report:

1. Q1\_partA.R file performs the Collecting Data step.

2. Q1\_partB.R file performs the Formatting and Summarizing Data.

3. A sample tweets file (tweets.json) is also provided along with this submission