Homework 5
STAT 4410/8416 Section 001
Fall 2014
Due: November 29, 2016 by midnight

1. **Scrapping HTML data:** We often obtain data from Wikipedia. This exercise will guide us to collect some data about the native speakers of some common languages. The information can be obtained from the following link (remember to copy it from pdf not Rnw file).

   http://en.Wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

   Now answer the questions below.

   (a) Read all the HTML tables available in the above link and store the result in an object called `tables`. Note that you may have ti use the function `getURL()` as we talked in class.

   (b) Now notice that the table in this link does not have any ID to specifically get the data from. But if you examine the source of the page, the first table is the data table. Thus we pick first table in the list of tables as our data. Store the data of first table in an object called `datRaw`.

   (c) We are particularly interested about columns 2 and 3 of `datRaw`. Subset columns 1 and 2 of `datRaw` and store the data in `dat`. Give the column names as language, nativeSpeaker. Display some data from `dat`.

   (d) Notice that the data is not clean. We have a comma(',') and refernces in parenthesis in the number of native speaker. We have to carefully review the data before we can use it. But first let us remove the ',' and additional numers with parenthesis from column 2 and make the column numeric. Also we may need to do the same operation with the first column. Clean the data and store it in `cleanDat`. Display some cleaned data.

   (e) Now plot the data to show language wise ranks and their relative position. For this we plan to select only top 20 languages based on number of speakers. Generate a bar chart showing the top 20 languages. Order the bars according to the number of speakers. Please **don't show** totals.

2. **Extracting twitter data:** In this problem we would like to extract data from twitter. For this refer to the documentation in the following link.

   `https://github.com/geoffjentry/twitteR/`

   (a) **Twitter API set up** Set up twitter API using any of the following methods. Make sure you installed all the packages as mentioned in the class.

   Method 1: Read Getting Started section of the above link and create a twitter application by going to the link `https://apps.twitter.com/`. Once you created your application connect twitter from `R` using the secrets and keys obtained from your twitter application.

```
library(twitteR)
api_key <- "your api key"
api_secret <- "your api secret"
access_token <- "your access tocken"
access_token_secret <- "your access tocken secret"

setup_twitter_oauth(api_key,api_secret,access_token,
                    access_token_secret)
```

   Method 2: If you don't like creating an account with twitter and going through all the trouble, you can use my keys (ssh, don't tell anyone). For this download the `hw5-twitter-auth` file from blackboard and load it as follows.

```
load("hw5-twitter-auth")

library(twitteR)
setup_twitter_oauth(api_key,api_secret,access_token,
                    access_token_secret)

## [1] "Using direct authentication"
```

(b) Now search twitter messages for "data science job". Display few job informations.

(c) Search 300 tweets using the hash tag "#chess" and save them in an object called `rTweets`. Show the top 7 sources of tweets (such as android or iphone) in a ordered bar plot.

(d) Notice that the object `rTweets` is a list. Convert it into a data frame using function `twListToDF` and store it in an object called `dTweets`. Display some data from `dTweets`.

(e) `dTweets` has a column showing the time the tweet was created. Generate a plot showing number of tweets on each of the hours.

(f) Arrange the dataframe `dTweets` based on the `retweetCount`. While doing this select only columns `text, screenName, retweetCount`. Store the data in a object called `mostTweets`. Display five texts that are most retweeted.

(g) Generate a bar chart showing top 10 screen names and count of retweets from `mostTweets`. Order the bars based on the retweet counts.

3. **Working with databases:** For this exercise we will use `MySQL` database available in the data science lab or the `datascienceVM`. Answer the following questions.

(a) Write down the connection string that would establish a connection to the `MySQL` database `trainingDB`.

(b) Write down a `SQL` command to select pclass, sex, survived and their average age from the titanic table. Store the selected data in data frame `avgAge` and display all the aggregated data.

(c) Now generate a line plot showing average age vs pclass colored by survived and faceted by sex.

(d) Use the package `dplyr` to obtain the same result as you did in question 3b. Display the results and the underlying `SQL` command used by `dplyr`.

(e) Find the name, age, sex and pclass of the 5 oldest and 5 youngest persons who died. Remove the people whose age information are not available for this computation.

4. **Exploring data:** Explore the crime data by downloading it from the blackboard. Provide nice tables and some plots that explain some important features revealed from the data. Discuss what you have found.