Homework 3
STAT 4410/8416 Section 001
Fall 2014
Due: October 12, 2014 by midnight

1. **Regular expression:** Write a regular expression to match patterns in the following strings. Demonstrate that your regular expression indeed matched that pattern by including codes and results. Carefully review how the first problem is solved for you.

   (a) We have a vector `vText` as follows. Write a regular expression that matches `g, og, go or ogo` in `vText` and replace the matches with dots.

   ```
   vText <- c('google','logo','dig', 'blog', 'boogie' )
   ```

   Answer:

   ```
   pattern <- 'o?go?'
   gsub(pattern, '.', vText)

   ## [1] "..le"  "l."    "di."   "bl."   "bo.ie"
   ```

   (b) Replace only the 2 or 3 digit numbers with the word `found` in the following vector. Please make sure that 4, 5 or 1 digit numbers do not get changed.

   ```
   vPhone <- c('35','6783','345', '20', '46349', '8' )
   ```

   (c) Replace all the characters that are not among the 26 English characters or a space. Please replace with an empty spring.

   ```
   myText <- "#y%o$u @g!o*t t9h(e) so#lu!tio$n c%or_r+e%ct"
   ```

   (d) In the following text, replace all the words that are exactly 3 characters long with triple dots '...'

   ```
   myText <- "All the three character words should be gone now"
   ```

   (e) Extract all the three numbers embedded in the following text.

   ```
   bigText <- 'There were three 20@14 numbers hidden in 500 texts'
   ```

   (f) Extract all the letters of the following sentence and show that it contains all the 26 letters.

   ```
   myText <- 'the quick brown fox jumps over the lazay dog'
   ```

2. Download the sample of a big data from blackboard. Note that the data is in csv format and compressed for easy handling. Now answer the following questions.

   (a) Read the data and select only the columns that contains the word 'human'. Store the data in an object `dat`. Report first few rows of your data.

   (b) The data frame `dat` should have 5 columns. Rename the column names keeping only the last character of the column names. So each column name will have only one character. Report first few rows of your data now.

   (c) Compute and report the means of each columns group by column b in a nice table.

   (d) Change the data into long form using id='b' and store the data in `mdat`. Report first few rows of data.

   (e) The data frame `mdat` is now ready for plotting. Generate density plots of value, color and fill by variable and facet by b.

(f) The data set `bigDataSample.csv` is a sample of much bigger data set. Here we read the data set and then selected the desired column. Do you think it would be wise do the same thing with the actual larger data set? Explain how you will solve this problem of selecting few columns (as we did in question 2a) without reading the whole data set first. Demonstrate that showing your codes.

Answer: We can use data table and read zero rows so that we only read the column names. After that we will pick the column names that contains the word 'human'. Finally we read only the few columns of the data.

3. **Extracting data from web:** Our plan is to extract data from web sources. This includes email addresses, phone numbers or other useful data. The function `readLines()` is very useful for this purpose.

   (a) Please read all the text in http://www.unomaha.edu/mahbubulmajumder/index.html and store your texts in `myText`. Show first few rows of `myText` and examine the structure of the data.

   (b) Now write a regular expression that would extract all the emails from `myText`. Include your codes and display the results that show only the email addresses and nothing else.

   (c) Now we want to extract all the phone/fax numbers in `myText`. Write a regular expression that would do this. Demonstrate your codes showing the results.

   (d) The link of ggplot2 documentation is http://docs.ggplot2.org/current/ and we would like to get the list of ggplot2 geoms from there. Write a regular expression that would extract all the geoms names (geom_bar is one of them) from this link and display the unique geoms. How many unique geoms does it have?

4. Download `lincoln-last-speech.txt` from the blackboard which contains the Lincoln's last public address. Now answer the following questions and include your codes.

   (a) Read the text and store the text in `lAddress`. Show the first 70 characters from the first element of the text.

   (b) Now we are interested in the words used in his speech. Extract all the words from `lAddress`, convert all of them to lower case and store the result in `vWord`. Display first few words.

   (c) The words like 'am', 'is', 'my' or 'through' are not much of our interest and these types of words are called stop-words. The package 'tm' has a function called `stopwords()`. Get all the English stop words and store them in `sWord`. Display few stop words in your report.

   (d) Remove all the `sWord` from `vWord` and store the result in `cleanWord`. Display first few clean words.

   (e) `cleanWord` contains all the cleaned words used in Lincoln's address. We would like to see which words are more frequently used. Find 15 most frequently used clean words and store the result in `fWord`. Display first 5 words from `fWord` along with their frequencies.

   (f) Construct a bar chart showing the count of each words for the 15 most frequently used words. Add a layer '`+coord_flip()`' with your plot.

   (g) What is the reason for adding a layer '`+coord_flip()`' with the plot in question (4f). Explain what would happen if we would not have done that.

   (h) The plot in question (4f) uses bar plot to display the data. Can you think of another plot that delivers the same information but looks much simpler? Demonstrate your answer by generating such a plot.