CS544 Module 6 Assignment

Part1) Strings (60 points)

Use the ***stringr*** functions for the following:

Initialize the vector of words from Lincoln's speech with the following code:

```
file <- "http://people.bu.edu/kalathur/datasets/lincoln.txt"
words <- scan(file, what=character())
```

a) Detect and show all the words that have a punctuation symbol.
b) Replace all the punctuations in the corresponding words with an empty string. Remove any empty words. Convert all words to lower case. Make this the ***new_words*** data.
c) What are the top 5 frequent words in the ***new_words*** data?
d) Show the frequencies of the word lengths in the ***new_words*** data. Plot the distribution of these frequencies.
e) What are the words in the ***new_words*** data with the longest length?
f) Show all the words in the ***new_words*** data that start with the letter ***p***.
g) Show all the words in the ***new_words*** data that end with the letter **r.**
h) Show all the words in the ***new_words*** data that start with the letter ***p*** and end with the letter ***r***.

Bonus: 10 points

In c), you realize that the most spoken words are what are known as stopwords. In text mining, the stopwords are removed before analysis. Initialize the common English stopwords as follows:

```
stopfile <- "http://people.bu.edu/kalathur/datasets/stopwords.txt"
stopwords <- scan(stopfile, what=character())
```

Filter the stopwords from the ***new_words*** data. Use the %in% operator. Repeat c) and d) for this data without the stop words.

Part2) Data Wrangling (40 points)

Use the ***tidyverse*** library for the following:

Download the following csv file,
http://people.bu.edu/kalathur/usa_daily_avg_temps.csv
locally first and use read.csv to load the data into a data frame.

a) Convert the data frame into a tibble and assign it to the variable *usaDailyTemps*.
b) What are the maximum temperatures recorded for each year? Show the values and also the appropriate plot for the results.
c) What are the maximum temperatures recorded for each state? Show the values and also the appropriate plot for the results.
d) Filter the Boston data from *usaDailyTemps* and assign it to the variable *bostonDailyTemps*.
e) What are the average monthly temperatures for Boston? Show the values and also the appropriate plot for the results. Use the *bostonDailyTemps.*

Submission:

- You must work on your assignments individually. You are not allowed to copy the answers from the others.
- Each assignment has a strict deadline. However, you are still allowed to submit your assignment within 2 days after the deadline with a penalty. 15% of the credit will be deducted unless you made previous arrangements with your facilitator and professor. Assignments submitted 2 days after the deadline will not be

Create a folder, CS544_HW6_lastName and place the following files in this folder. Provide the R code, **HW6_lastName.R**, with each portion of the code clearly identified by the corresponding question. Prepare a corresponding Word/PDF document by pasting the output for each question including the plots and explanations.

Archive the folder (CS544_HW6_lastName.zip). Upload the zip file to the Assignments section of Blackboard.