CS544 Module 6 Assignment

Part1) Strings (60 points)

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

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

```
file <- "https://people.bu.edu/kalathur/datasets/mlk.txt"

words <- scan(file, what=character())
```

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

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

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

Part2) Data Wrangling (40 points)

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

Download the following csv file,
https://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:

When the term *lastName* is referenced, please replace it with your last name.

Provide all R code in a single file, **CS544_HW6_LastName.R**. Clearly mark each subpart of each question.

Provide the corresponding outputs from the R console in a single PDF document, **CS544_HW6_LastName**.pdf

Upload the two files to the Assignments section of Blackboard.

Note: Only ONE submission is allowed. Please be sure that what you are submitting is your final submission.