Assignment #4

MSAN 593

***DUE**: Saturday, August 13th , 23:45*

Be sure to upload a single `*.R` script file named `asn4.R` to Canvas by the due date and time. No write-up is required. Late submissions will receive a grade of zero. Your `*.R` files will be run on local machines by graders. If your file does not run, you will automatically lose 30% of the grade. If you resubmit your corrected homework by the last class of the module, it will be graded out of the remaining 70%. Failure to resubmit will result in a grade of zero. Assume that the data file(s) being read into your `*.R` files are in the current local directory, e.g., `read.csv('myFile.csv')` will work. Do **not** hard code a specific directory structure.

Question 1

Using `tweets.csv`, answer the following:

1. Identify all tweets with the word 'flight' in them
2. How many tweets end in a question mark?
3. How many tweets have airport codes in them (assume any three subsequent capital letters are airport codes)
4. Identify all tweets with URLs in them
5. Replace all instances of repeated exclamation points with a single exclamation point
6. Replace consecutive exclamation points, question marks, and periods with a single period, split the tweet on periods, and create a list where each element is a vector of the split strings from each tweet
7. Return a list where each element is a vector of hashtags for each tweet

Question 2

The file `dates.csv` contains two variables `startDate` and `endDate`. Validate the data by ensuring that `startDate` is chronologically earlier than `endDate`. Remove all observations that do not conform. Do this using two techniques:

1. Use a `for()` loop. You may want to leave your computer running over night for this depending on how old your computer is.

2. Use a functional from the `apply()` family of functionals. This should be reasonably quick (a minute or two).

n.b. You may want to use the `lubridate` package to help process the date data.

Question 3

You now have the fundamental `R` tools to complete this exercise, but you will be required to explore new techniques and packages.

You will work with the full text of the State of the Union speeches from 1790 until 2012. The speeches are all in the file `stateoftheunion1790-2012.txt`. Read the text into `R` and manipulate it in order to create a data frame with

1. the president's name
2. year
3. month
4. day of month
5. day of week
6. length of the speech (lines)
7. number of sentences
8. number of words