BSDS 100: Intro to Data Science with R
Case Study 2: Text Mining

by James D. Wilson (University of San Francisco)

Directions: For all questions in this assignment, write complete sentences and fully answer any question that is asked, and use R to answer each question. Provide all R code and solutions by *knitting* your final RStudio file into a single file. You can work in groups of up to 5, and only one person needs to upload the final .pdf to Canvas. Be sure to put all of your group members' names at the top of the document

1. Using the `tweets.csv` data that is available on the GitHub site, provide code to do the following
   (a) Identify all tweets with the word 'flight' in them
   (b) How many tweets end in a question mark?
   (c) How many tweets have airport codes in them (assume any three subsequent capital letters are airport codes)
   (d) Identify all tweets with URLs in them
   (e) Replace all instances of repeated exclamation points with a single exclamation point
   (f) 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
2. You now have the fundamental R tools to complete this exercise, but you will may still have 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` on the GitHub site. Read the text into R and manipulate it in order to create a data frame with the following summaries for each speech:
   (a) the President's name who gave the speech
   (b) the year the speech was given
   (c) the month the speech was given
   (d) day of the week the speech was given
   (e) the number of sentences in the speech
   (f) the number of words in the speech