Stat 133, Spr 15
Homework 5: Regular Expressions & Creating Variables for Spam Prediction
Due: Friday Apr 4

For this homework, you need to create a data frame by deriving variables from a set of email messages. These messages are available in an R dataset in the rda file, located at
http://www.stat.berkeley.edu/users/nolan/data/emailsStat133.rda

Descriptions of variables appear below. In addition, the file called spamDerivedVars.R provides the signature for each function.

1. The subject is "Re: something or other."
2. The percentage of characters in the subject that are capitals (exclude punctuation and numbers from the denominator).
3. The From: address ends in numbers, e.g.

   ```
   david gezi <davidgezi12@hotmail.com>
   ```

4. The number of exclamation marks in the body.

The emails you will use for this assignment are in the list called emails, where each element of the list contains one email message. Each email is itself a list consisting of three elements:

- The element named "header" is a named character vector, where each name corresponds to a key in the email header and the value of the element corresponds to the text following the ':' in the key:value of the header.
- The element named "body" which contains the body of the email message. This element is a character vector, with one string per line in the email message.
- The element named isSpam is a logical vector of length 1 that indicates whether the message is spam (TRUE) or ham (FALSE).

Turn in (1) the spamDeriveVars.R file containing the completed functions that you use to create these variables, (2) a completed spamReport.Rmd file that contains a graphical comparison of spam and ham for two of these variables that you think might be promising for predicting spam, and (3) the knitted version of your report.