Stat 133, Fall 04
Homework 3: Text Manipulation: Creating Spam-related Variables
Due: Monday, 4 Oct

For this homework, you need to create three variables from the email messages. These messages are available in an R dataset in the rda file, located at http://www.stat.berkeley.edu/users/nolan/stat133/data/Emails.rda.

Descriptions of thrity variables appear below. They are split into three groups: A, B, C. You are to transform the email data into three of these variables, one from each group. The list found at the end of the homework determines which three you will write code to create. For at least one of the three your code must be in an R function.

Email the code you use to create these three variables, as plain text in the body of your email. In addition, turn in a graphical comparison of spam and ham for each of these three variables. Discuss whether or not you think this variable will be useful in predicting if an email message is ham or spam.

GROUP A:

1. The subject is ”Re: something or other.”
2. The number of lines in the body of the email.
3. The number of characters in the body of the email.
4. The Reply-To has an underline and numbers/letters.
5. The number of exclamation marks in the subject.
6. The number of question marks in the subject.
7. The number of attachments.
8. The X-priority or X-Msmail-Priority set to high.
9. The number of recipients.

GROUP B:

1. The average length of words in the body.
2. The Received time in the current time zone.

3. The From: ends in numbers, e.g.

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

4. The subject is all capitals (excluding punctuation and numbers)
5. The percent of lines in the body of the email that begin with >.
6. The subject contains one of the following words: viagra, pounds, free, weight, guarantee, millions, dollars, credit, risk, prescription, generic, drug, money back, credit card.
7. The Message-Id has no hostname.
8. The body of the email contains a line with the two words ”Original” and ”Message” and no other alpha characters.
9. The percentage of blanks in the subject.
10. The body of the email contains a the word “wrote:” or “schrieb:” or “ecrit:” or a similiar expression in another language.

GROUP C:

1. The email contains the recipient’s email address.
2. The recipient list is sorted by address
3. The subject has punctuation or digits surrounded by characters, e.g. V?agra and pay1ng, but not New!
4. The difference between the Date and the Received date (be careful with time differences.
5. The header states that the message is multipart, but it is mostly text or html (i.e. the number of attachments that are plain text or html).
6. The email contains images.
7. The number of dollar signs in the body of the email.
8. The body contains Dear something, such as DEAR SIR, or Dear Madam
9. The Message-Id has a hostname that does not match the senders hostname, but does match a host name at a relay point.
10. For an HTML attachment, the percentage of characters in the html tags as a percentage of the total number of characters in the message (excluding blanks). Note that html tags start < and end >.

11. The percentage of the characters in the body of the email that are upper case (excluding blanks, numbers, and punctuation).

The emails you will use for this assignment are in the list 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” is itself a list, the first element of which is named ”text” and contains the body of the email message. This element is a character vector, with one string per line in the email message. A second element, if it exists, is named “attachments.” This element is a list containing one element per attachment. The individual attachment element is a list of two elements – one containing information about the format of the attachment and the other containing and the contents of the attachment.
- The element named spam is a logical vector of length 1 that indicates whether the message is spam (TRUE) or ham (FALSE).

To determine which three variable you are to write the code to create, look up the last letter in your SCF login, i.e. if your login is s133bu then your assignment is #8 in group A, #1 in group B, and #4 in group C.

A	B	C	SCF login	A	B	C	SCF login
1	1	1	a	2	2	2	b
3	3	3	c	4	4	4	d
5	5	5	e	6	6	6	f
7	7	7	g	8	8	8	h
9	9	9	i	9	10	10	j
5	3	4	k	9	2	6	l
8	2	7	m	7	4	8	n
6	5	9	o	5	6	10	p
4	7	3	q	3	8	1	r
2	9	2	s	1	10	3	t
8	1	4	u	7	2	5	v
6	3	7	w	5	4	6	x
1	7	2	y	4	5	9	z