STAT 33B Homework 7

Yuanrui Zhu (3034615728)

Dec 10, 2020

This homework is due **Dec 10, 2020** by 11:59pm PT.

Homeworks are graded for correctness.

As you work, write your answers in this notebook. Answer questions with complete sentences, and put code in code chunks. You can make as many new code chunks as you like.

Please do not delete the exercises already in this notebook, because it may interfere with our grading tools.

You need to submit your work in two places:

- Submit this Rmd file with your edits on bCourses.
- Knit and submit the generated PDF file on Gradescope.

If you have any last-minute trouble knitting, **DON'T PANIC**. Submit your Rmd file on time and follow up in office hours or on Piazza to sort out the PDF.

SpamAssassin Email Data

The SpamAssassin Email Data set is a collection of email messages used to train the SpamAssassin software to detect spam. The email messages are divided into legitimate "ham" emails and illegitimate "spam" emails. Each email is in a separate plain text file.

In this assignment, you'll only use a collection of "ham" emails. You can find the emails in the file `emails.zip` on the bCourse. You will need to unzip the file before proceeding with Exercise 1.

This data set is originally from the Apache SpamAssassin project.

Exercise 1

The `readLines` function reads lines of text from a file and returns them in a character vector with one element for each line. The first argument is the path to the file. By default, the function will read all of the lines in the file.

Write a function `read_email` that reads all of the text in a single email file. Your function should have a parameter `file` to set the path to the file. Your function should collapse all of the lines in the file into a single string with lines separated by the newline character `\n`.

Show that your function works for 3 of the email files.

Hint: The `paste` function is relevant here.

YOUR ANSWER GOES HERE:

```r
read_email = function(file) {
  email_lines = readLines(file, encoding="latin1")
  paste(file = email_lines, collapse ="\n")
}
cat(read_email("easy_ham/0003.acfc5ad94bbd27118a0d8685d18c89dd"))
```

```
## From timc@2ubh.com  Thu Aug 22 13:52:59 2002
## Return-Path: <timc@2ubh.com>
## Delivered-To: zzzz@localhost.netnoteinc.com
## Received: from localhost (localhost [127.0.0.1])
##  by phobos.labs.netnoteinc.com (Postfix) with ESMTP id 0314547C66
##  for <zzzz@localhost>; Thu, 22 Aug 2002 08:52:58 -0400 (EDT)
## Received: from phobos [127.0.0.1]
##  by localhost with IMAP (fetchmail-5.9.0)
##  for zzzz@localhost (single-drop); Thu, 22 Aug 2002 13:52:59 +0100 (IST)
## Received: from n16.grp.scd.yahoo.com (n16.grp.scd.yahoo.com
##     [66.218.66.71]) by dogma.slashnull.org (8.11.6/8.11.6) with SMTP id
##     g7MCrdZ07070 for <zzzz@example.com>; Thu, 22 Aug 2002 13:53:39 +0100
## X-Egroups-Return: sentto-2242572-52733-1030020820-zzzz=example.com@returns.groups.yahoo.com
## Received: from [66.218.67.198] by n16.grp.scd.yahoo.com with NNFMP;
##     22 Aug 2002 12:53:40 -0000
## X-Sender: timc@2ubh.com
## X-Apparently-To: zzzzteana@yahoogroups.com
## Received: (EGP: mail-8_1_0_1); 22 Aug 2002 12:53:39 -0000
## Received: (qmail 76099 invoked from network); 22 Aug 2002 12:53:39 -0000
## Received: from unknown (66.218.66.218) by m5.grp.scd.yahoo.com with QMQP;
##     22 Aug 2002 12:53:39 -0000
## Received: from unknown (HELO rhenium.btinternet.com) (194.73.73.93) by
##     mta3.grp.scd.yahoo.com with SMTP; 22 Aug 2002 12:53:39 -0000
## Received: from host217-36-23-185.in-addr.btopenworld.com ([217.36.23.185])
##     by rhenium.btinternet.com with esmtp (Exim 3.22 #8) id 17hrT0-0004gj-00
##     for forteana@yahoogroups.com; Thu, 22 Aug 2002 13:53:38 +0100
## X-Mailer: Microsoft Outlook Express Macintosh Edition - 4.5 (0410)
## To: zzzzteana <zzzzteana@yahoogroups.com>
## X-Priority: 3
## Message-Id: <E17hrT0-0004gj-00@rhenium.btinternet.com>
## From: "Tim Chapman" <timc@2ubh.com>
## X-Yahoo-Profile: tim2ubh
## MIME-Version: 1.0
## Mailing-List: list zzzzteana@yahoogroups.com; contact
##     forteana-owner@yahoogroups.com
## Delivered-To: mailing list zzzzteana@yahoogroups.com
## Precedence: bulk
## List-Unsubscribe: <mailto:zzzzteana-unsubscribe@yahoogroups.com>
## Date: Thu, 22 Aug 2002 13:52:38 +0100
## Subject: [zzzzteana] Moscow bomber
## Reply-To: zzzzteana@yahoogroups.com
## Content-Type: text/plain; charset=US-ASCII
## Content-Transfer-Encoding: 7bit
##
## Man Threatens Explosion In Moscow
##
## Thursday August 22, 2002 1:40 PM
## MOSCOW (AP) - Security officers on Thursday seized an unidentified man who
## said he was armed with explosives and threatened to blow up his truck in
## front of Russia's Federal Security Services headquarters in Moscow, NTV
## television reported.
## The officers seized an automatic rifle the man was carrying, then the man
## got out of the truck and was taken into custody, NTV said. No other details
## were immediately available.
## The man had demanded talks with high government officials, the Interfax and
## ITAR-Tass news agencies said. Ekho Moskvy radio reported that he wanted to
## talk with Russian President Vladimir Putin.
## Police and security forces rushed to the Security Service building, within
## blocks of the Kremlin, Red Square and the Bolshoi Ballet, and surrounded the
## man, who claimed to have one and a half tons of explosives, the news
## agencies said. Negotiations continued for about one and a half hours outside
## the building, ITAR-Tass and Interfax reported, citing witnesses.
## The man later drove away from the building, under police escort, and drove
## to a street near Moscow's Olympic Penta Hotel, where authorities held
## further negotiations with him, the Moscow police press service said. The
## move appeared to be an attempt by security services to get him to a more
## secure location.
##
## ------------------------ Yahoo! Groups Sponsor ---------------------~-->
## 4 DVDs Free +s&p Join Now
## http://us.click.yahoo.com/pt6YBB/NXiEAA/mG3HAA/7gSolB/TM
## ---------------------------------------------------------------------~->
##
## To unsubscribe from this group, send an email to:
## forteana-unsubscribe@egroups.com
##
##
##
## Your use of Yahoo! Groups is subject to http://docs.yahoo.com/info/terms/
```

```
cat(read_email("easy_ham/0011.07b11073b53634cff892a7988289a72e"))
```

```
## From exmh-workers-admin@redhat.com  Thu Aug 22 15:15:12 2002
## Return-Path: <exmh-workers-admin@example.com>
## Delivered-To: zzzz@localhost.netnoteinc.com
## Received: from localhost (localhost [127.0.0.1])
##  by phobos.labs.netnoteinc.com (Postfix) with ESMTP id 1AD7B43F99
##  for <zzzz@localhost>; Thu, 22 Aug 2002 10:15:11 -0400 (EDT)
## Received: from phobos [127.0.0.1]
##  by localhost with IMAP (fetchmail-5.9.0)
##  for zzzz@localhost (single-drop); Thu, 22 Aug 2002 15:15:12 +0100 (IST)
## Received: from listman.example.com (listman.example.com [66.187.233.211]) by
##     dogma.slashnull.org (8.11.6/8.11.6) with ESMTP id g7MECrZ09674 for
##     <zzzz-exmh@example.com>; Thu, 22 Aug 2002 15:12:54 +0100
## Received: from listman.example.com (localhost.localdomain [127.0.0.1]) by
##     listman.redhat.com (Postfix) with ESMTP id C57DA3ECC9; Thu, 22 Aug 2002
##     10:13:02 -0400 (EDT)
## Delivered-To: exmh-workers@listman.example.com
## Received: from int-mx1.corp.example.com (int-mx1.corp.example.com
##     [172.16.52.254]) by listman.redhat.com (Postfix) with ESMTP id 6854840C75
##     for <exmh-workers@listman.redhat.com>; Thu, 22 Aug 2002 10:12:27 -0400
##     (EDT)
## Received: (from mail@localhost) by int-mx1.corp.example.com (8.11.6/8.11.6)
##     id g7MECOK08343 for exmh-workers@listman.redhat.com; Thu, 22 Aug 2002
##     10:12:24 -0400
## Received: from mx1.example.com (mx1.example.com [172.16.48.31]) by
##     int-mx1.corp.redhat.com (8.11.6/8.11.6) with SMTP id g7MECOY08339 for
##     <exmh-workers@redhat.com>; Thu, 22 Aug 2002 10:12:24 -0400
## Received: from austin-jump.vircio.com
##     (IDENT:m7qlhYJ9XjzGHDfWy27ABasIREnlFU84@jump-austin.vircio.com
##     [192.12.3.99]) by mx1.redhat.com (8.11.6/8.11.6) with SMTP id g7MDvul20394
##     for <exmh-workers@redhat.com>; Thu, 22 Aug 2002 09:57:56 -0400
## Received: (qmail 31227 invoked by uid 104); 22 Aug 2002 14:12:23 -0000
## Received: from cwg-exmh@DeepEddy.Com by localhost.localdomain with
##     qmail-scanner-0.90 (uvscan: v4.1.60/v4218. . Clean. Processed in 0.340551
##     secs); 22/08/2002 09:12:23
## Received: from deepeddy.vircio.com (@[10.1.2.1]) (envelope-sender
##     <cwg-exmh@DeepEddy.Com>) by austin-jump.vircio.com (qmail-ldap-1.03) with
##     SMTP for <exmh-workers@redhat.com>; 22 Aug 2002 14:12:23 -0000
## Received: (qmail 25511 invoked from network); 22 Aug 2002 14:12:18 -0000
## Received: from localhost (HELO deepeddy.vircio.com)
##     (?LVKop2xGBpLZYqQBUZF/+jebI90KasL/?@[127.0.0.1]) (envelope-sender
##     <cwg-exmh@DeepEddy.Com>) by localhost (qmail-ldap-1.03) with SMTP for
##     <exmh-workers@redhat.com>; 22 Aug 2002 14:12:18 -0000
## X-Mailer: exmh version 2.5 07/13/2001 with nmh-1.0.4
## To: "J. W. Ballantine" <jwb@homer.att.com>
## Cc: exmh-workers@example.com
## Subject: Re: New Sequences Window
## In-Reply-To: <200208211351.JAA15807@hera.homer.att.com>
## References: <200208211351.JAA15807@hera.homer.att.com>
## X-Url: http://www.DeepEddy.Com/~cwg
## X-Image-Url: http://www.DeepEddy.Com/~cwg/chris.gif
## MIME-Version: 1.0
## Content-Type: multipart/signed;
##     boundary="==_Exmh_1547759024P";
##     micalg=pgp-sha1;
##     protocol="application/pgp-signature"
## Content-Transfer-Encoding: 7bit
## Message-Id: <1030025538.25487.TMDA@deepeddy.vircio.com>
## From: Chris Garrigues <cwg-exmh@DeepEddy.Com>
## X-Delivery-Agent: TMDA/0.57
## Reply-To: Chris Garrigues <cwg-dated-1030457538.09ad9f@DeepEddy.Com>
## X-Loop: exmh-workers@example.com
## Sender: exmh-workers-admin@example.com
## Errors-To: exmh-workers-admin@example.com
## X-Beenthere: exmh-workers@example.com
## X-Mailman-Version: 2.0.1
## Precedence: bulk
## List-Help: <mailto:exmh-workers-request@example.com?subject=help>
## List-Post: <mailto:exmh-workers@example.com>
## List-Subscribe: <https://listman.example.com/mailman/listinfo/exmh-workers>,
##     <mailto:exmh-workers-request@redhat.com?subject=subscribe>
## List-Id: Discussion list for EXMH developers <exmh-workers.example.com>
## List-Unsubscribe: <https://listman.example.com/mailman/listinfo/exmh-workers>,
##     <mailto:exmh-workers-request@redhat.com?subject=unsubscribe>
## List-Archive: <https://listman.example.com/mailman/private/exmh-workers/>
## Date: Thu, 22 Aug 2002 09:12:16 -0500
##
## --==_Exmh_1547759024P
## Content-Type: text/plain; charset=us-ascii
##
## > From:  "J. W. Ballantine" <jwb@homer.att.com>
## > Date:  Wed, 21 Aug 2002 09:51:31 -0400
## >
## > I CVS'ed the unseen/Sequences changes and installed them, and have only one
## > real issue.
## >
## > I use the unseen window rather than the exmh icon, and with the new code
## > I can't seem to be able to.  How many unseen when when I have the main window open
## > is not really necessary.
##
## hmmm, I stole the code from unseenwin, but I never tested it since I don't use
## that functionality.  Consider it on my list of things to check.
##
## Chris
##
## --
## Chris Garrigues                 http://www.DeepEddy.Com/~cwg/
## virCIO                          http://www.virCIO.Com
## 716 Congress, Suite 200
## Austin, TX  78701        +1 512 374 0500
##
##   World War III:  The Wrong-Doers Vs. the Evil-Doers.
##
##
##
##
## --==_Exmh_1547759024P
## Content-Type: application/pgp-signature
##
## -----BEGIN PGP SIGNATURE-----
## Version: GnuPG v1.0.6 (GNU/Linux)
## Comment: Exmh version 2.2_20000822 06/23/2000
##
## iD8DBQE9ZPFAK9b4h5R0IUIRAkjyAJ4jjjhAVRx5FiwuCMa+QBWsbbE2jQCaAj4x
## NhIgYqnx9/1wvdSgesQhMIU=
## =vA3k
## -----END PGP SIGNATURE-----
##
## --==_Exmh_1547759024P--
##
##
##
## _______________________________________________
## Exmh-workers mailing list
## Exmh-workers@redhat.com
## https://listman.redhat.com/mailman/listinfo/exmh-workers
```

```r
cat(read_email("easy_ham/0018.ba70ecbeea6f427b951067f34e23bae6"))
```

```
## From exmh-workers-admin@redhat.com  Thu Aug 22 16:37:36 2002
## Return-Path: <exmh-workers-admin@example.com>
## Delivered-To: zzzz@localhost.netnoteinc.com
## Received: from localhost (localhost [127.0.0.1])
##  by phobos.labs.netnoteinc.com (Postfix) with ESMTP id 50AF343F9B
##  for <zzzz@localhost>; Thu, 22 Aug 2002 11:37:35 -0400 (EDT)
## Received: from phobos [127.0.0.1]
##  by localhost with IMAP (fetchmail-5.9.0)
##  for zzzz@localhost (single-drop); Thu, 22 Aug 2002 16:37:35 +0100 (IST)
## Received: from listman.example.com (listman.example.com [66.187.233.211]) by
##     dogma.slashnull.org (8.11.6/8.11.6) with ESMTP id g7MFZLZ12577 for
##     <zzzz-exmh@example.com>; Thu, 22 Aug 2002 16:35:22 +0100
## Received: from listman.example.com (localhost.localdomain [127.0.0.1]) by
##     listman.redhat.com (Postfix) with ESMTP id 51F5140D92; Thu, 22 Aug 2002
##     11:35:26 -0400 (EDT)
## Delivered-To: exmh-workers@listman.example.com
## Received: from int-mx1.corp.example.com (int-mx1.corp.example.com
##     [172.16.52.254]) by listman.redhat.com (Postfix) with ESMTP id 6EED940E00
##     for <exmh-workers@listman.redhat.com>; Thu, 22 Aug 2002 11:26:01 -0400
##     (EDT)
## Received: (from mail@localhost) by int-mx1.corp.example.com (8.11.6/8.11.6)
##     id g7MFPwB25414 for exmh-workers@listman.redhat.com; Thu, 22 Aug 2002
##     11:25:58 -0400
## Received: from mx1.example.com (mx1.example.com [172.16.48.31]) by
##     int-mx1.corp.redhat.com (8.11.6/8.11.6) with SMTP id g7MFPwY25410 for
##     <exmh-workers@redhat.com>; Thu, 22 Aug 2002 11:25:58 -0400
## Received: from austin-jump.vircio.com
##     (IDENT:+BhY7EqRf5fwVT64o4aVh7UUHF1egIL+@jump-austin.vircio.com
##     [192.12.3.99]) by mx1.redhat.com (8.11.6/8.11.6) with SMTP id g7MFBTl04916
##     for <exmh-workers@redhat.com>; Thu, 22 Aug 2002 11:11:29 -0400
## Received: (qmail 4141 invoked by uid 104); 22 Aug 2002 15:25:57 -0000
## Received: from cwg-exmh@DeepEddy.Com by localhost.localdomain with
##     qmail-scanner-0.90 (uvscan: v4.1.60/v4218. . Clean. Processed in 0.327895
##     secs); 22/08/2002 10:25:57
## Received: from deepeddy.vircio.com (@[10.1.2.1]) (envelope-sender
##     <cwg-exmh@DeepEddy.Com>) by austin-jump.vircio.com (qmail-ldap-1.03) with
##     SMTP for <exmh-workers@redhat.com>; 22 Aug 2002 15:25:56 -0000
## Received: (qmail 13189 invoked from network); 22 Aug 2002 15:25:53 -0000
## Received: from localhost (HELO deepeddy.vircio.com)
##     (?AL1b2OiTMIHwZbG9ZCdQYQG3nsIe5jbe?@[127.0.0.1]) (envelope-sender
##     <cwg-exmh@DeepEddy.Com>) by localhost (qmail-ldap-1.03) with SMTP for
##     <exmh-workers@redhat.com>; 22 Aug 2002 15:25:53 -0000
## X-Mailer: exmh version 2.5 07/13/2001 with nmh-1.0.4
## To: Robert Elz <kre@munnari.OZ.AU>, exmh-workers@example.com
## Subject: Re: New Sequences Window
## In-Reply-To: <1029944441.398.TMDA@deepeddy.vircio.com>
## References: <1029882468.3116.TMDA@deepeddy.vircio.com>
##     <9627.1029933001@munnari.OZ.AU>
##     <1029943066.26919.TMDA@deepeddy.vircio.com>
##     <1029944441.398.TMDA@deepeddy.vircio.com>
## X-Url: http://www.DeepEddy.Com/~cwg
## X-Image-Url: http://www.DeepEddy.Com/~cwg/chris.gif
## MIME-Version: 1.0
## Content-Type: multipart/signed;
##     boundary="==_Exmh_-1317289252P";
##     micalg=pgp-sha1;
##     protocol="application/pgp-signature"
## Content-Transfer-Encoding: 7bit
## Message-Id: <1030029953.13171.TMDA@deepeddy.vircio.com>
## From: Chris Garrigues <cwg-exmh@DeepEddy.Com>
## X-Delivery-Agent: TMDA/0.57
## Reply-To: Chris Garrigues <cwg-dated-1030461953.beb807@DeepEddy.Com>
## X-Loop: exmh-workers@example.com
## Sender: exmh-workers-admin@example.com
## Errors-To: exmh-workers-admin@example.com
## X-Beenthere: exmh-workers@example.com
## X-Mailman-Version: 2.0.1
## Precedence: bulk
## List-Help: <mailto:exmh-workers-request@example.com?subject=help>
## List-Post: <mailto:exmh-workers@example.com>
## List-Subscribe: <https://listman.example.com/mailman/listinfo/exmh-workers>,
##     <mailto:exmh-workers-request@redhat.com?subject=subscribe>
## List-Id: Discussion list for EXMH developers <exmh-workers.example.com>
## List-Unsubscribe: <https://listman.example.com/mailman/listinfo/exmh-workers>,
##     <mailto:exmh-workers-request@redhat.com?subject=unsubscribe>
## List-Archive: <https://listman.example.com/mailman/private/exmh-workers/>
## Date: Thu, 22 Aug 2002 10:25:52 -0500
##
## --==_Exmh_-1317289252P
## Content-Type: text/plain; charset=us-ascii
##
## > From:  Chris Garrigues <cwg-exmh@DeepEddy.Com>
## > Date:  Wed, 21 Aug 2002 10:40:39 -0500
## >
## > > From:  Chris Garrigues <cwg-exmh@DeepEddy.Com>
## > > Date:  Wed, 21 Aug 2002 10:17:45 -0500
## > >
## > > Ouch...I'll get right on it.
## > >
## > > > From:  Robert Elz <kre@munnari.OZ.AU>
## > > > Date:  Wed, 21 Aug 2002 19:30:01 +0700
## > > >
## > > > Any chance of having that lengthen instead?   I like all my exmh stuff
## > > > in nice columns (fits the display better).   That is, I use the detache
## > d
## > > > folder list, one column.   The main exmh window takes up full screen,
## > > > top to bottom, but less than half the width, etc...
## >
## > I thought about that.  The first order approximation would be to just add
## > using pack .... -side top instead of pack ... -side left, however, since their
## > each a different width, it would look funny.
##
## I've done this.  It's not as pretty as I think it should be, but it works.
## I'm going to leave the cosmetic issues to others.  When I update the
## documentation, I'll add this to the exmh.TODO file.
```

```
##
## I'm leaving for a 2 1/2 week vacation in a week, so this is the last new
## functionality I'm going to add for a while.  Also, I now have pretty much
## everything in there that I want for my own use, so I'm probably pretty much
## done.  I'll work on bug fixes and documentation before my vacation, and
## hopefully do nothing more afterwards.
##
## Chris
##
## --
## Chris Garrigues                 http://www.DeepEddy.Com/~cwg/
## virCIO                          http://www.virCIO.Com
## 716 Congress, Suite 200
## Austin, TX  78701        +1 512 374 0500
##
##   World War III:  The Wrong-Doers Vs. the Evil-Doers.
##
##
##
##
## --==_Exmh_-1317289252P
## Content-Type: application/pgp-signature
##
## -----BEGIN PGP SIGNATURE-----
## Version: GnuPG v1.0.6 (GNU/Linux)
## Comment: Exmh version 2.2_20000822 06/23/2000
##
## iD8DBQE9ZQJ/K9b4h5R0IUIRAiPuAJwL4mUus5whLNQZC8MsDlGpEdKNrACcDfZH
## PcGgN9frLIM+C5Z3vagi2wE=
## =qJoJ
## -----END PGP SIGNATURE-----
##
## --==_Exmh_-1317289252P--
##
##
##
## _______________________________________________
## Exmh-workers mailing list
## Exmh-workers@redhat.com
## https://listman.redhat.com/mailman/listinfo/exmh-workers
```

Exercise 2

Write a function `read_email_all` that reads all of the files in the email directory and returns a character vector with one element for each email. Your function should have a parameter `dir` to set the path to the directory. Your function should call the `read_email` function from Exercise 1.

Make sure not to put other files in the email directory!

After writing your function, use it to read all of the email files into a character vector called `emails`.

How many email files are there?

Hint: The `list.files` function is relevant here.

YOUR ANSWER GOES HERE:

```r
read_email_all = function(dir){
    email_file_names = list.files(dir)
    email_file_paths = paste(dir, email_file_names, sep = "/")
    sapply(email_file_paths, read_email)
}
emails = read_email_all("easy_ham")
length(list.files("easy_ham"))
```

```
## [1] 2551
```

Exercise 3

Use stringr and regular expressions to write a function `extract_email_addr` that extracts all email addresses from a character vector. Your function should have a parameter `x` for the character vector, and should return a character vector with one element for each email address (duplicates are okay).

For simplicity, you can assume the formatting rules for email addresses are that they:

1. Must contain exactly one at-symbol `@`
2. Can also contain any number of letters, numbers, or characters in `._-`

Test your function on some made up strings and also on one of the email messages.

Hint: Using character classes `[ ]` in the regex pattern is important here.

Note: It's not necessary for this exercise, but if you're curious about the actual rules for email addresses, see Section 3.4.1 of RFC 5322, or this Wikipedia article.

YOUR ANSWER GOES HERE:

```r
myText = read_email("easy_ham/0003.acfc5ad94bbd27118a0d8685d18c89dd")
extract_email_addr = function(x) {
  unlist(regmatches(x,gregexpr("([_+a-z0-9-]+(\\.[_+a-z0-9-]+)*@[a-z0-9-]+(\\.[a-z0-9-]+)*(\\.[a-z]{2,1
}
extract_email_addr("<karry@bp.net><jerry@zhu.berkeley.com>")
```

```
## [1] "karry@bp.net"           "jerry@zhu.berkeley.com"
```

```r
extract_email_addr(myText)
```

```
##  [1] "timc@2ubh.com"
##  [2] "timc@2ubh.com"
##  [3] "zzzz@localhost.netnoteinc.com"
##  [4] "zzzz@example.com"
##  [5] "example.com@returns.groups.yahoo.com"
##  [6] "timc@2ubh.com"
##  [7] "zzzzteana@yahoogroups.com"
##  [8] "forteana@yahoogroups.com"
##  [9] "zzzzteana@yahoogroups.com"
## [10] "0-0004gj-00@rhenium.btinternet.com"
## [11] "timc@2ubh.com"
## [12] "zzzzteana@yahoogroups.com"
## [13] "forteana-owner@yahoogroups.com"
## [14] "zzzzteana@yahoogroups.com"
## [15] "zzzzteana-unsubscribe@yahoogroups.com"
## [16] "zzzzteana@yahoogroups.com"
## [17] "forteana-unsubscribe@egroups.com"
```

Exercise 4

Using your `extract_email_addr` function and the email message data:

1. How many different email addresses appear in the emails?
2. Which 5 email addresses appear the most?

YOUR ANSWER GOES HERE:

```r
length(unique(extract_email_addr(emails)))
```

```
## [1] 2985
```

```r
df = as.data.frame(table(extract_email_addr(emails)))
head(df[order(df$Freq, decreasing = T), ], 5)
```

```
##                            Var1 Freq
## 2062        fork-admin@xent.com 2948
## 2064           fork@example.com 2774
## 2063      fork-request@xent.com 2107
## 2950 yyyy@localhost.example.com 1699
## 2292              jm@jmason.org 1471
```

There are 2985 unique email addresses in the emails. The five email addresses appear the most are shown in the second dataframe above.

Exercise 5

The part of an email address after the `@` is called the *domain*. The domain refers to a website, so it usually contains at least one dot. For example, Cal email addresses use the domain `berkeley.edu`, which is also the Cal website address.

Which 10 domains are the most common among the email addresses you extracted from the email data?

How many domains in the email addresses end in `.edu`?

YOUR ANSWER GOES HERE:

```r
test = extract_email_addr(emails)
domain <- gsub(".*@(.+)$", "\\1", test)
df2 = as.data.frame(table(domain))
df3 = df2[order(df2$Freq, decreasing = T), ]
head_10 = head(df3, 10)
df4 = as.data.frame(table(sub("([_a-z0-9-]+\\.)+", "", df3$domain)))
head_10
```

```
##                      domain Freq
## 234             example.com 6882
## 828                xent.com 5927
## 255           freshrpms.net 3643
## 368              jmason.org 2854
## 235 example.sourceforge.net 2068
## 417   localhost.example.com 1838
## 414   lists.sourceforge.net 1716
## 613              redhat.com 1331
## 405                linux.ie 1081
## 833         yahoogroups.com  687
```

```
df4[df4["Var1"] == "edu",]
```

```
##    Var1 Freq
## 16  edu   51
```

The ten domains which are most common are shown above, and there are 51 email addresses end in ".edu".