STAT 430, FALL 2008
HOMEWORK 6: REGULAR EXPRESSIONS
DUE MONDAY

Warm-up

1. Write a regular expression that matches lines 70 or more characters long. No need to use R…
2. Which of the following matches the regular expression `a(ab)*a`
    a. abababa
    b. aaba
    c. aabbaa
    d. aba
    e. aabababa
3. Which of the following matches the regular expression `[a-z]+[\\.\\?!]`
    a. battle!
    b. Hot
    c. Green
    d. swamping.
    e. jump up.
    f. undulate?
    g. is.?
4. Which of the following matches the regular expression `[a-zA-Z]*[^,]=`
    a. Butt=
    b. BotHEr,=
    c. Ample
    d. FIdDlE7h=
    e. Brittle =
    f. Other.=

Problem – Catching Dr. Rein's Cheaters

As you know Dr. Rein suspects that some students may have cheated during his exam. Weblogs for the studio servers for 10/28 and 10/29 are available at http://statweb.calpoly.edu/aschaffn/430/Data/ex081028.txt and http://statweb.calpoly.edu/aschaffn/430/Data/ex081029.txt.

Helpful Info (from Dr. Rein's handout):

- As mentioned by Dr. Rein, the times in these files are in GMT, so to convert to PDT, you'll have to subtract 7 hours.
- The IP addresses of all machines on campus are in the range 129.65.0.1 to 129.65.254.254
- The IP addresses of all studio machines are 129.65.21.208 to 129.65.21.232
- The exams took place from 1:10 – 2:00 and 2:10 – 3:00 on each of the two days.
- The only allowed files from Dr. Rein are midterm2.htm and MT2.htm

1. You are to write a function that has the following inputs:
    - `log.name` = the location of the log file (ex: 'ex081029.txt'. Assume the file has the header info that needs to be cleaned.)
    - `perm.files` = a character vector with the names or regular expressions of the files to search for (ex: 'MT2\\.html?')
    - `instr` = a character string of the name of the instructor (ex: 'srein')
    - `time.window` = a charater vector of length 2 giving the beginning and end times for exam to be converted to POSIX format(ex: c('2008-10-29 00:00:05', '2008-10-29 12:34:22'))
    - `time.include` = a logical to search within the time window or outside the time window (ex: T)

    And the following outputs as a data frame:
    - `ID` = the index position of the record in the log file.
    - `date` = a POSIXct value for the date/time the file was requested
    - `IP` = the IP address of the machine being requested
    - `filename` = the name of the file being requested

2. Use your function to find any cases where students accessed Dr. Rein's files with answers to sample exam questions during exam hours. How many times were files accessed? Which files? Which IPs?
3. Use your function to find any cases where students attempted to access exam materials after the first exam ended (day 1), but before the 2nd exam began (day 2). How many attempts were there? What were the IP addresses? How many were on campus? Off campus?

4. Write a function that has the following inputs:
    - `log.name` = the location of the log file (ex: 'ex081029.txt'. Assume the file has the header info that needs to be cleaned.)
    - `instr` = a character string of the name of the instructor (ex: 'srein')
    - `topN` = a numeric indicating the maximum number of files to list

    And the following output:
    - a named numeric vector (i.e., table output) listing the name and frequency of the top `topN` files requested from the particular instructor. The names should only include the last part of the URL (i.e., /aschaffn/430/Data/081029.txt should only appear as 081029.txt)