Stat 133, Spring 2011
Homework 2: Data in R and Regular Expressions
Due Friday, Mar. 4 at 11:59 PM

Reading Data from Web Pages Using R

In this assignment, you will write programs to gather data from various websites. As a courtesy to the maintainers of the web sites, I ask that you download a local copy of the web page to work with, but make sure that when your program is completed, it can successfully extract the data directly from the web page. You may find it convenient to use the R function `download.file` to make a local copy of the pages.

You should not edit the file you're trying to extract from, nor should you use cut and paste to extract the data. Your program should operate on the page in its entirety.

1. At http://news.ask.com/news, there are a number of headlines broken down into categories, like "Top Stories" and "Science & Tech" Extract all the headlines from the page into a vector of character strings in R, one headline per string, and print them.

2. The pdf file http://www.bea.gov/scb/pdf/2011/02 February/D Pages/0211dpg_c.pdf contains, among other things, the Gross Domestic Product of the United States for the last 50 years. You can convert a pdf file called, say `file.pdf` to a text file using the command:

   ```
   pdftotext -layout file.pdf
   ```

   which will automatically create a file called `file.txt`. Using this data, create a data frame with columns for year and GDP. The `pdftotext` command is available on any of the SCF computers – if you'd like to install `pdftotext` on your own computer, go to the xpdf home page.

   At http://www.treasurydirect.gov/govt/reports/pd/pd.htm, there are a set of links entitled "Historical Debt Outstanding". These links lead to five tables, showing dates and the amount of the national debt on those dates.

   Read the data for the debt from the tables that correspond to the GDP data, *i.e.* the last fifty years, combine them with the GDP data, and merge them together with the GDP data based on the year of the data. Finally, calculate the percentage of the GDP that the debt represents for each year of available data, and make a plot of this percentage over time.

3. The populations of the US states in 2004 through 2009 can be found at http://www.infoplease.com/ipa/A0004986.html. Create a data frame with columns for the state name, 2009 population, 2004 population, and the percent change from 1990-2000. Calculate the change in population between the 2004 and 2009 as a column in

the data frame, and compare that change with the one from 1990-2000. Which states had the most different population growths from 1990 to 2000 than they did from 2004 to 2009.

4. At http://www.forbes.com/lists/2009/54/rich-list-09_The-400-Richest-Americans_Rank.html is the first page of a list of the 400 richest Americans from 2009. The complete list of the top 400 is spread over 16 pages.

   Write a program that reads *all* of the pages, and create a data frame with the rank, name, net worth, age, residence, and source of wealth for the top 400 richest Americans. Who was the youngest person on the list? What were the ten most common sources of wealth among the 400 richest Americans?

Submit for grading all of your R code and the output requested by the assignment, along with any output you used to answer the questions posed by the assignments.

Your submission should be contained in a *single* pdf file.

Email this file to me (`s133@stat.berkeley.edu`) by 11:59PM on the due date. Make certain to save a copy of your email submission.