Lab 10: Scrape the Rich

36-350, Fall 2013

15 November 2013

We will practice scraping data from web pages by extracting information about the 400 richest people in America from Forbes's list of them. Forbes displays this as four pages with 100 people each; to get around some issues with the way those pages are dynamically generated (apparently for ad placement), use the cached copies at `http://www.stat.cmu.edu/~cshalizi/statcomp/13/labs/10/rich-1.html` through `http://www.stat.cmu.edu/~cshalizi/statcomp/13/labs/10/rich-4.html`.

1. (10) Read in all four pages using `readLines`. Merge lines together with the newline character,
   `n`. Merge the files together into one string, called `richhtml`. Verify that `richhtml` is character vector of length 1. How many characters does it contain?

2. (10) Open up `rich-1.html` in a text editor (*not* a web browser), and find the entry for Bill Gates. Write a regular expression, with a parenthesized capture group, which will catch his name in the capture group. Write code, using `regexpr` and `regmatches`, which will return his name when applied to `richhtml`.

3. (15) Modify your code so that, when applied to `richhtml`, it gives you a vector of *all* the names, in order. Verify that you have a character vector of length 400. Check the first and last six entries against the web pages. Check that all 400 entries are in fact names (but do *not* turn in the vector of 400 names).

4. (10) Write a regular expression which should capture a person's net worth, in billions of dollars. Check that it is working by using it with `regexpr` and `regmatches` of Bill Gates. Then get a vector of all the net worths. What are the mean and median net worths?

5. (10) Write a regular expression to capture each person's age. After checking that it works on Bill Gates, get a vector of 400 ages. What are the minimum, median, and maximum ages?

6. (15) Write regular expressions to get the place of residence and the source of wealth. (These are two separate things.) Check that they work properly

on Bill Gates, giving a character vector of length two. Modify your code appropriately to work on all 400 individuals, returning a 400×2 array of characters.

7. (10) Create a data frame which gives, for each person (row), their name, their net worth, their age, their place of residence, and their source of wealth. *Hint:* `data.frame`.

8. (5) Make a histogram of the 400 net worths. Upload the image, as well as giving the command you used to make it.

9. (5) Make a scatter-plot of net worth against age. Upload the image, as well as giving the command you used to make it.

10. (5) What fraction of the total net worth of people in the top 400 comes from Walmart? (You should be able to do this in one line of R.)

11. (5) What is the most common *state* of residence?