Homework 9: Get (the 400) Rich(est list) Quick

36-350, Fall 2011

Due at 11:59 pm on Wednesday, 23 November 2011

You know the rules by now for turning in work.

1. `forbes.com` maintains an annual list of the 400 richest people in America at the `http://www.forbes.com/forbes-400`. This year (2011) the list spans 4 pages. The web page is actually generated dynamically by a technology, called AJAX, that allows your web browser and the web server to communicate even after the initial page load. This makes scraping a little more difficult, but some snooping in the HTML and Javascript reveals that the data for the Forbes 400 list is actually served from the URL `http://www.forbes.com/forbes-400/ajax/loadList`. Use the following command to load this *raw data* into R:

   ```
   raw400 <- readLines('http://www.forbes.com/forbes-400/ajax/loadList')
   ```

   For this homework you may work with either this raw data stream (at `http://www.forbes.com/forbes-400/ajax/loadList`), or the dynamically generated HTML (at `http://www.forbes.com/forbes-400`). To make things easier, the 4 HTML pages have been extracted for you (at `http://www.stat.cmu.edu/~cshalizi/statcomp/hw/09/pageX.html`, where X is 1,2,3, or 4). They can be loaded into R using the following commands:

   ```
   baseurl <- 'http://www.stat.cmu.edu/~cshalizi/statcomp/hw/09/'
   urls <- paste(baseurl, 'page', 1:4, '.html', sep = '')
   page <- lapply(urls, readLines, warn = FALSE)
   ```

   (a) (60) Write an R program that reads the Forbes 400 list into a data frame with the following columns:

   - `rank` — rank of the person (numeric)
   - `name` — name of the person (character vector)
   - `net.worth` — net worth in billions (numeric)
   - `age` — (numeric) Note that some billionaires' ages are not known!
   - `residence` — (character vector of the form: city, state)
   - `source` — (character vector of the form: source1, source2, ..., lastsource)

There should be 400 rows with each row corresponding to a person on the list.

(b) (10) Make a histogram of the 400 net worths.

(c) (10) Make a scatter plot of net worth and age. Who is the youngest person on the list? Oldest?

(d) (10) What were the ten most common sources of wealth?

(e) (10) What are the ten most common states of residence?