MSDA 607 - Project 4

Project 4 is due end of day on Sunday April 26th. You may work in a small team on the project (and you may define what you feel is a small team).

Be courteous: Any time you are hitting multiple pages on a web site, you should consider putting a pause in between page reads, so your R application is not treated as a denial of service attack. For example:

```
Sys.sleep(1)
```

●

The site r-bloggers is a team blog, with a lot of great how-to content on various R topics. The page http://www.r-bloggers.com/search/web%20scraping provides a list of topics related to web scraping, which is also the topic of this project!

Grading rubric:

- For each of the reference blog entries on the first page, you should pull out the title, date, and author, and store these in an R data frame. Your code should be in github, and published to rpubs.com. You'll receive a maximum of 90% for completing this base assignment.
- To earn the full 100 points, you must do some kind of further data extraction and/or analysis. Here are four sample ideas. You don't need to do more than one of these, and you are free to instead choose your own area for further analysis. Maximum additional points: 10%
  - Extend your scraper to include the base information for blog entries on all of the tagged pages. Your R data frame should include any necessary additional rows.

  - Select what you feel are the (3 to 7?) most important R packages related to screen scraping. For each of the blog entries, identify which of these packages is mentioned. Your R data frame should include any necessary additional column(s).
  - Does the code that you created run against another group of similarly tagged blog entries on this site, e.g. http://www.r-bloggers.com/search/twitter? What modifications did you need to make to the code.
  - Does http://www.r-bloggers.com provide an API to search its blog entries? If you were to create an API for them, what would it look like?

You are encouraged to start early, ask many questions, actively post on the provided Project 4 discussion forum, etc.