**Project Proposal**

For my project, I intend to scrape data from [www.setlist.fm](http://www.setlist.fm), a user-created repository of concert setlists, in order to create various summary statistics about concerts that have been played since 1/1/2000. Analyses may include the number of concerts per artist per year, common touring combinations of artists, favored artists by the setlist.fm community, and other interesting statistics as they may occur to me in the process of scraping the data from the website. Ultimately, I would like to register my own domain name (I’ve been eyeing [www.jenniferbrussow.com](http://www.jenniferbrussow.com) for a couple years now) and set up a website that would include the results of this project as well as other information typical to a researcher’s website (other projects, CV, etc.).

I have selected this project in order to acquire new skills that I have not had the opportunity to develop in my other coursework. Namely, I have had many opportunities to conduct simulation studies in other classes I have taken through the REMS program (Bayesian, IRT II, Fortran, etc.), but my degree path does not provide much in the way of practical programming and data analysis skills. I have been interested in data mining from websites for several years now, and this project seems like the ideal opportunity to develop those skills for myself. I am sure I will also pick up some new R tricks along the way.

Another avenue I have been interested in exploring is website development. It seems that many successful researchers have their own personal websites where they are able to showcase their research results and provide a point of contact for potential employers. As a person who would like to be successful in life, a personal website seems like a logical step for me. I suspect that the process of web hosting may have changed since I was last admin of a website (circa 2005), so I would like to set up this new page while able to benefit from the expertise of my instructor and fellow classmates.

Now, on to the specifics of the project. The website I plan to scrape data from, [www.setlist.fm](http://www.setlist.fm), is a website I have had an account on for several years. Users can add concerts, add and update setlists for concerts, mark the concerts they have attended to generate a profile of personalized statistics. The site also provides some high-level summary statistics such as the top artists, festivals, and venues in terms of concertgoers, but it doesn’t provide information about how statistics may have changed through time. Additionally, statistics may be skewed due to the user base of the website not being representative of the overall population of concertgoers – an idea that would be interesting to investigate if I am able to find some sort of data source to cross-validate. I plan to use the R packages “rvest” and “httr,” both of which are new to me, to scrape data from the site.

To set up a website, I will likely use DreamHost to register the domain name and secure hosting. I will likely use WordPress to create the basic website through DreamHost, then customize it to my purposes. I hope to use R Markdown and/or Shiny to create nice-looking, interactive pages to display the results of my analyses. This process will require me to learn about using FTP software to transfer files to the web host, learn how to better use R Markdown, and brush up on my html skills.