APIs and Web Scraping

There are now five rules that apply to all projects so far:

a) Follow instructions *precisely*. If I do not tell you what to write on a particular line, leave it blank.
b) Do not use any functions or approaches to problems that we have not yet learned in this course.
c) All code must be *scalable by sample size* unless specifically noted otherwise.
d) Any code using *magrittr* should contain a max of one verb per line.
e) Always use the standard set of headings up to this point unless otherwise specified: **R Studio API Code, Libraries, Data Import and Cleaning, Analysis, Visualization**

Part 1 – #RStats Community Study

1. **week11-twitter.Rmd**
    a. In an R Notebook, download per-post like data from the most recent 1000 posts from #rstats
    b. Create a tibble called *tweets_tbl* containing four columns: the screen name of the person tweeting, the content of their tweet, how many times they have been liked/favorited, and how many times they have been retweeted. Do not include retweets.
    c. Is there a relationship between length of tweet and retweet popularity? What about between length of tweet and likes/favorites? Calculate a statistical test and create a visualization of each relationship.
    d. Ensure you have sufficient headings and annotation so that a person looking at your notebook would be able to clearly understand what you did. When in doubt, over-explain. Export your notebook as a PDF to your *output* folder.

Part 2 – Google Scholar Web Scrape

2. **week11-scholar.Rmd**
    a. There is no API available to scrape Google Scholar, so you'll need to do this by hand. Choose someone's Google Scholar page – it does not matter who you choose, as long as they have at least 10 papers listed. Using R code alone, read this page and create a single tibble called *profile_tbl* containing columns representing the following information: name of article, author list, year, and citation count. Thus, a person with 20 papers listed should have 80 pieces of data: 20 rows x 4 columns. Don't worry about anything past 20 citations.
    b. Calculate a correlation between year and citation count.
    c. Create a scatterplot showing this relationship, superimposing a regression line.
    d. Once again, export a well-annotated notebook into a PDF in *output*.

Part 3 – Qualtrics API Access

3. **week11-qualtrics.R**
    a. The *qualtRics* package can be used to access your Qualtrics account from R. If you don't have a Qualtrics account yet, you can open it here: https://umn.qualtrics.com
    b. Using this package and ggplot(), display a bar chart showing number of surveys you've created over time (time on x, count on y). If you don't have any surveys in Qualtrics already, you might want to create one for testing purposes. Replace your API key with the text *apikeygoeshere* before submitting anything to GitHub or to Canvas.