Stat Method for Big Data: Homework 1

Instructor: Xi Luo

Friday 3$^{\text{rd}}$ February, 2017
Time: 14:52

Contents

Instructions

- Deadline: **11:59 pm, Feb 18, 2017** on Canvas.
- Please start working on this early, even though the deadline is in two weeks. It is hard to predict what kind of bugs will come up!
- You are allowed (also encouraged) to work in teams. Each team can have any number of members.
- Each team should submit only one solution, with all the team members clearly listed.
- The members of each team may self sign-up the project groups on Canvas. Please let me know if it does not work.
- You are encouraged to seek help from the instructor.
- Please submit your cdes only (R or other languages). Please do not submit the data unless required by the problem.
- Please submit your solution to Canvas online. Note that the online submission will be closed automatically after the deadline. Before the deadline, you may submit replacements.

---

1 Optional Materials on Linux and R

There are a lot of materials online on Linux and R. If you are unfamiliar with these, you are encouraged to fresh up using these additional resources, including:

- Intro to R: https://www.youtube.com/playlist?list=PLOU2XLYxmsIK9qQfztXeybpHvru-TrqAP
- Linux basics: https://www.youtube.com/results?search_query=linux+basics+
- Free Lynda.com courses: https://it.brown.edu/announcements/read/browns-new-lyndacom-subscription
- Manuals on regular expressions and xpath are mentioned in lecture slides

You don't need to write a solution for this question.

2 S & P 100 Stock Prices

We will use this case study to extend the case study on Yahoo's stock data introduced in class. The goal is to extract the historical prices of the stocks listed in S & P 100 index. Please carry out the following steps:

1. Wikipedia has a html page on the S & P 100 list at http://en.wikipedia.org/wiki/S%26P_100. The table under the section `Components` is the list that we want. Please extract the symbols and company names from the webpage.
    - Hint: you may consider using the R function readHTMLTable to simplify the task.
2. Because Yahoo Finance uses a slightly different system of symbols, you will have to change the symbol `BRK.B` to `BRK-B`.
3. Please use the symbol names extracted in step 2 to download the historical stock prices in csv format. You may use the downloading function that we developed in class.
4. Please add a column of the corresponding symbol for each csv file downloaded in step 3. Note that the new column should be the first column and comma separated from other columns. Please save the results as csv files.
5. Please concatenate all the csv files in step 4 together as a large csv file that contains the prices for all the stocks.

3 Funding and Publications

In this case study, we want to find out the relationship between funding and publications. Later on, we will introduce more advanced analysis of the data, but the first step is to collect the data!

The HW01 folder on Canvas provides a csv file that contains recent NIH awards of faculty and students from Harvard University. This list is provided by NIH. However, it did not provide the publications of each awardee. We will use PubMed as a complementary source of data. For simplicity, we will focus on the numbers of publications of each awardee only.

Please carry out the following steps:

1. We will focus on research awards only. Please remove the awards (column: Activity) starting with letter T or F, and then extract the unique PI names from the column: Contact PI / Project Leader.
2. The extracted names may contain middle names and/or initials. Please remove the middle names/initials from the results in Step 1. It could be possible that two names differ only by the middle name. We will ignore this rare case for now, or please let me know if you find any.
3. PubMed http://www.ncbi.nlm.nih.gov/pubmed/ is a online catalog of publications like Google Scholar, but it accepts more refined search criteria. We will use author and affiliation to restrict the matched publications. For example, to search for Professor Xihong Lin's publications, you can enter "LIN, XIHONG[Author] AND Harvard[Affiliation]" into the search box. The number of publications from the returned page, which may show "Results: 1 to 20 of 68" for Professor Lin. Please extract the number of publications for the names extracted in step 2.
    - Note that some search pages may be empty, and please build a mechanism to set the number of publications to zero in this case.
4. In addition to your program, please submit your results as a csv table that contains the names and number of publications.