Statistics 5014: Homework 9

Due In Class November 15, 9am

2017-11-03

For each assignment, turn in by the due date/time. Late assignments must be arranged prior to submission. In every case, assignments are to be typed neatly using proper English in Markdown.

This week, we spoke about text mining and sentiment analysis. Most of the material came from http://tidytextmining.com/. While this is not the only way to mine textual data, it fits nicely into the tidy process we used in our search for Reproducible Analysis.

Problem 1 (GitHub 2 pts + 2 Style pts)

This week we will change gears a little on GitHub. Today we start making homework something we can use to highlight our professionalism and abilities. You will continue to retrieve the lectures and homeworks from my GitHub (anyway you like), but you will:

1. create a new GitHub Repository (online)
2. invite me as a collaborator (settings –> collaborators)
3. setup an Rproject pointing to this new repository
4. create a new RNotebook file within the project folder
5. save the notebook HW9_lastname_firstname

Note that this homework will be graded for professionalism as well as your ability to solve the problems (passing is ≥ 7. Remember, reproducibility requires weaving text (appropriate for a reader), code (commented) and necessary output as a compendium for what was done.

Problem 2: Text analysis (4 pts)

Duplicate one of the following analysis (NOT the simple.py one) in python in an RNotebook.

https://github.com/amueller/word_cloud

Note, you are DUPLICATING someone else's work, reference it appropriately. Explain the steps appropriately.

What is the point to this problem?
1. You are demonstrating you read some of the lecture notes.
2. You have python correctly installed. The simple.py ported to my lecture notes worked correctly in python3.
3. You are getting a feeling for how easy it is to incorporate differnt codes bases in a single reproducible document.

Note that I used Conda to do all my installation stuff.
https://conda.io/docs/user-guide/getting-started.html

Problem 3 (2 pts)

Push your homework to YOUR GitHub.

This is YOUR GitHub Repository. Save what you want, there is a limit to how big a repo can be, so be a little sparing on what you put up there. Please save the file you would like me to read as: HW#_lastname_firstname.html. I am not going to go through your Markdown to try to figure out what you were doing.

Optional preperation for next class:

Read up on regular expressions.