Stat 6910 Data Technologies Fall Semester 2018

Project (10/26/2018)

100 Points — Multiple Due Dates

In this Data Technologies (DT) course project, you have to use a DT method we have discussed in class. You can scrape data in html, XML, or pdf format from the web, or you can do some extensive data manipulation via regular expressions or using the functionality from *tidyverse*. The data should not be readily available in any formats that are directly supported by R or Microsoft Excel (such as rda, txt, csv, xls, xlsx, etc.). This project can be closely related to your MS or PhD research. However, this should be a side–project with respect to your overall MS or PhD research and should not be used as a chapter of your MS report or dissertation. **The project is an individual project and not a group project!**

In a few cases in past years, interested students continued to work on this project after the end of this course and eventually were able to transform this project into a conference presentation and an accompanying proceedings paper. Other students used the data from their project in one of my follow–up courses such as *Statistical Visualization I & II* or *Applied Spatial Statistics*. If this is of interest to you, let's further discuss this once the course has ended.

This project consists of multiple stages, outlined in more details below:

(i) (10 Points) **Preliminary discussion of project proposal:** If you have some ideas for a project that meets the overall ideas outlined above, I would like to hear your suggestions! If you don't have any ideas (or no MS or no PhD topic yet), I have some possible suggestions for you. You have to meet with me in the week of October 22 (Mo 10/22/18 – Fr 10/26/18) for this preliminary discussion. This can be during my regular office hours or at an individually scheduled time. **Deliverables:** Please e–mail by Sunday 10/21/18, 11:59pm, when you want to meet with me for this preliminary discussion and indicate whether you have a topic idea.

(ii) (10 Points) **Project Proposal:** Based on the preliminary discussion of your project proposal, prepare a full two–page project proposal. This is kind of a road map to a successful completion of your project. You must indicate which data

set(s) or web pages you are going to use and which existing R packages (and functions) you are going to use. Be specific how you are going to use/apply/modify/extend the existing functionality. It must become clear what already exists and what needs to be implemented by you. Be realistic and only suggest what can be done over a five–week period. Cite and list supporting references (at least any required R packages needed for your project). Small graphics, diagrams, or sketches are fine to support your proposal. R code is not needed at this time.
Deliverables: Please e–mail your project proposal by Sunday 11/4/18, 11:59pm. As soon as you get my approval, you should start working on your project. If you submit your proposal early, I will try to provide feedback as soon as possible so you will have a few extra days to work on your project.

(iii) (40 [final version] + 10 [preliminary version] Points) **Project Paper:** Summarize your project in a project paper that resembles a first short (six–page) draft of a proceedings paper for the Joint Statistical Meetings (JSM). There must be a title, abstract, and keywords. Main sections are the introduction, methods (describing the data and previously existing DT methods you use), a section describing your results in this project, and a discussion/conclusion/outlook (on future work) section, followed by the reference list. Also, include your R code in the appendix. The appendix does not count towards the six–page limit.
Deliverables: Please e–mail your final project paper by Friday 12/14/18, 11:59pm. A preliminary version of your project paper (just the pdf) is due via e–mail by Sunday 11/18/18, 11:59pm. This preliminary version needs to show that you are on track with your project and that you will be able to finish within the next two weeks. For example, update the LaTeX template, prepare a first draft of the introduction and the methods section, prepare first versions of your figures (these don't have to be the final ones and may still violate some of the rules for good graphics) and include the new R code you have developed so far in the appendix.

(iv) (15 [slides] + 15 [presentation] Points) **Project Presentation:** Prepare a 20min presentation of your work, following the main sections of your (preliminary) project paper. Each of you will present his/her presentation to the other students and other people from the department interested in the work conducted in this class. Imagine that this is a contributed section at the JSM, something many of you likely will experience soon (or already have experienced in the recent past). I plan to hold this presentation session on Thursday 11/29/18 in our last scheduled lecture this semester.

Deliverables: Please e–mail your project presentation slides by 11:59pm the day before the presentations — and hold your 20min presentation during our presentation session (using my laptop).

General Instructions

Generalize from the instructions from previous homeworks in this course! All programming must be done in R (possibly with a combination of other external programming languages and tools if needed for your project). Your project proposal and project report must be prepared with LaTeX, sweave, and/or knitr. Your presentation has to be prepared with the LaTeX–beamer package. Please submit all your source and resulting output files (Rnw, pdf, tex, figures, local data files, etc.) for your final version of the project paper and your presentation. For the project proposal and the preliminary version of the project paper, you only have to submit the pdf file.

Before you submit your files, make sure that everything works on a computer other than your own computer! Your documents must be fully reproducible on a different computer, i.e., I must be able to re–translate your files without encountering any errors. If you haven't done so at the start of the semester, update R and all of your R packages to the most recent version (use the *installr* R package to quickly update to a new version of R).

Note: I will provide several templates, such as those for the JSM proceedings and a JSM presentation (and also the underlying source files).

Whenever you have questions or need clarifications, talk to me in person, via e–mail, or via Skype during the Thanksgiving week. Don't make any wrong assumptions, but rather check with me what I want you to do in this project!

Good luck!