EPPS 7v81 Advanced Data Programming

Summer 2020
Wednesday: 4:00 pm - 6:45 pm On-line

Instructor: Dr. Karl Ho kyho@utdallas.edu
Online Office Hours: On Microsoft Teams (Team code: **1vxfka4**[1])
Monday through Friday 10:30 am to 12 noon, by appointment
Teaching Assistant: Yalin Yang Yalin.Yang@UTDallas.edu
Online Office Hours: On Microsoft Teams
Monday and Thursday from 2 pm to 3:30 pm, by appointment

Overview:

This is an applied course introducing advanced data programming training for data science projects. Data programming is a practice that works around and evolves with data. This course covers the general principles for data programming or coding involving data and techniques in building data models and applications. Students will practice data programming on collecting, managing, visualizing and modeling data from a variety of sources including social media, websites and API's. While this is a survey course covering major languages and data methods, students are encouraged to develop own specialization in areas including web scraping, data management, text and complex data analytics, data visualization and specialized learning methods. No prerequisites required but some programming experience will be helpful.

Learning Objectives:

At completion of this course, students will be familiar with:

1. General programming principles
2. Programming languages for data science projects
3. Applications on social science and policy research in particular new methods such as machine learning, web scraping and text analytics.
4. Application development process including design and debugging of programs

Tools:

Git and GitHub (https://github.com)
PyCharm (https://www.jetbrains.com/shop/eform/students) *
RStudio (https://rstudio.com/products/rstudio/)
R (https://cran.r-project.org)
Python 3.x (https://www.python.org/downloads/)
Anaconda (https://www.anaconda.com)

(*) Free Community/Education version available

GitHub:

https://github.com/datageneration/advanceddataprogramming

Required readings:
Brooker, Phillip D., 2019. *Programming with Python for Social Scientists*. SAGE Publications Limited. **PB**
Chacon, Scott and Straub, Ben. 2014. *Pro git*. Apress. (https://git-scm.com/book/en/v2) **SC&BS**
Gentzkow, Matthew and Shapiro, Jesse M., 2014. Code and data for the social sciences: A practitioner's guide. *Chicago, IL: University of Chicago*. **MG&JS**
Garrett Grolemund and Hadley Wickham, 2017, *R for Data Science*, O'Reilly. (https://r4ds.had.co.nz/) **GG&HW**

[1] Click **Join or create a team** below your teams list and look for the **Join a team with a code** card

Wickham, H., 2019. *Advanced R*. CRC press. (https://adv-r.hadley.nz/) **HW**
Math Primer:
Fernández, Maribel, 2014. Mathematical Background. In *Programming Languages and Operational Semantics* (pp. 21-41). Springer, London. **MF‡**
Laaksonen, Antti, 2017. *Guide to Competitive Programming*. Springer International Publishing. Appendix A. (https://link-springer-com.libproxy.utdallas.edu/book/10.1007/978-3-319-72547-5) **AL**

(‡) Available electronically at McDermott Library

Camera-readiness: This online course requires students to attend with camera on. Camera-readiness is critical for future career in academia and industry alike. Data scientists must excel in communicating data projects and effective face-to-face presentation is very important part of data science training. Learn how to best learn and work remotely from this free Udemy course https://www.udemy.com/course/quick-guide-to-working-remotely-from-vp-of-learning/.

BYOD: Students are expected to have own functional, up-to-date computer for this course. Personal computer (not mobile device) running MacOS, Linux or Windows operating systems is recommended.

Participation: Full attendance of all classes is required and imperative. Attendance however is not enough. All class members have to actively participate in class preparation and discussion. Participation entails full preparation for class including research of class materials, completion of assignments and full involvement in class discussion. Only medical emergency (with documentation) is considered excused absence. If you have a special medical condition, please contact me in first week and special arrangements can be made. Participation is responsible for 10% of the final grade.

Assignments

Exercises will be assigned. Students are expected to finish all assignments and apply techniques to own project. At beginning of each class, one student will be randomly selected to present and discuss assignment solutions.

Grading and Requirements:

Project:

Each student must design and implement a data project using the systems and techniques covered in class. Identify one original topic to investigate. Originality is the first quality the instructor demands. Original means the topic or data or methods must be originally thought out and designed and has never been published. Replication with new data and/or method will count. The final report on design and implementation of data project must be between 15-20 pages in length, not including codes and data appendices. Please read carefully the University's policy about cheating and plagiarism (see below for the link to the document). Due dates of the proposal presentations will be announced in class. No late submission will be accepted. Proposal of the project **must be consulted with instructor in advance**. It constitutes 30% of the final grade. The final presentation, which constitute a highly important part of the class, accounts for 30% of the final grade. In total, the final project is responsible for 60 percent of the entire grade. In summary, the grade structure is as follows:

Participation	10%
Project Proposal	30%
Final project presentation:	30%
Final project	30%

Document guideline:
All documents in this class must adhere to the following general guidelines:

- Must be typed or word-processed on letter size papers, stapled on upper left-hand corner and one inch on all margins
- No binders or plastic covers
- For final project, use a cover sheet with topic and name

Topics:

	Topic	Readings
1	Introduction - Why data programming? - Survey of programming languages	MG&JS PB: 1-3 Math primers
2	Programming languages - Python - R - Other	PB: 9 GG&HW: 4-8
3	Git and GitHub - Version control - Collaboration and Replicability	SC&BS
4	Python I - Syntax and concepts - IDE	PB: 4-5
5	Python II - Objects and Classes - Functions	PB: 6-8
6	Advanced R - Functional programming - Object-oriented programming	GG&HW: 8-21 HW: 1-16, 22-24
7	Data collection - API's - Web scraping - Social media: Twitter data	PB: 11, 13
8	Data management - Data formats - Relational data and NoSQL	PB: 12
9	Data visualization - Dashboard - Markdown and notebooks	PB: 14 GG&HW: 2, 26-28
10	Data modeling - Statistical learning models - Model selection	TBA

Comet Creed

This creed was voted on by the UT Dallas student body in 2014. It is a standard that Comets choose to live by and encourage others to do the same:

"As a Comet, I pledge honesty, integrity, and service in all that I do."

UT Dallas Syllabus Policies and Procedures

The information contained in the following link constitutes the University's policies and procedures segment of the course syllabus.
Please go to http://go.utdallas.edu/syllabus-policies for these policies.

The descriptions and timelines contained in this syllabus are subject to change at the discretion of the Professor.