CSC 523 – Advanced Scripting for Data Manipulation, Analysis, and Machine Learning
Fall 2022 First Day Handout

Mon 6-8:50 PM, Old Main 158, Zoom classes & recordings, https://faculty.kutztown.edu/parson

My courses are multimodal this semester, meaning you can attend in-person or remotely via Zoom. I insist on maintaining 6 feet of distance from myself in order to reduce the odds of carrying the virus home to my wife & two-year-old granddaughter. I plan to wear a mask. I strongly encourage the unvaccinated to get vaccinated unless a medical condition precludes that. The unvaccinated will not only infect each other, they will also provide an environment in which the virus can mutate to more dangerous strains. I will post Zoom videos for all classes.

The class time interactive Zoom link appears on **D2L** Course CSC523 -> Content -> Overview.

Office Hours Monday 2-4, Wednesday 4-6 (Zoom only), Thursday 10-11 or by appt. All available via Zoom.

This course covers advanced study and practice in using a modern scripting language to integrate off-the-shelf code libraries for the retrieval of unstructured and partially structured data, and for the cleaning, integration, formatting, storage, analysis, and visualization of large data sets. Modern scripting languages include powerful built-in features for storing, retrieving, mapping, and integrating data; code libraries extend such features greatly. Libraries include those for regular-expression based extraction of textual data, data integration, statistical analysis and correlation, machine learning, natural language processing, machine vision and listening visualization, and storage in files and database systems. Emphasis is on using a scripting language to glue together off-the-shelf library modules without writing the complex, underlying library code.

Optional Textbooks: *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, Aurélien Géron, 978-1492032649; (super optional: *Python for Data Analysis*, Wes McKinney, 978-1491957660)

Grading (A = 92:100, A- = 90:91, B+ = 87:89, B = 82:86, B- = 80:81, C+ = 77:79, C = 70:76, D = 60:69, F = 0:59) There is no "D" grade for student grading in 400- and 500-level courses at KU.

Projects 100% divided equally among the project assignments.

Final one or two projects are student presentations and materials for an individual student project.

Programming project assignment grading criteria

Grading rubrics will be part of each assignment handout. Late penalty is 10% per each day late, up until I go over the solution. Any assignment turned in after that is worth 0%.

We will use the CS&IT documentation requirements:

http://faculty.kutztown.edu/parson/CSCDocumentationStandards.pdf

The academic integrity policy:

http://faculty.kutztown.edu/parson/AcademicIntegrityPolicy.pdf

Please read the above policy statement.

You may openly discuss ideas, algorithms, pitfalls, and the use of programming tools.

You may not share code, test drivers or test data except within groups for group projects.

Group projects, when assigned, have documented partitioning of student responsibilities.

There will be a 10% per day late penalty for projects that come in after the due date.

Class attendance is not graded, but I will be teaching using data sources and concepts both inside and outside the scope of the textbook. You are responsible for all material covered in class, including technical information, coding standards and conventions, verbal specification of assignments, and your questions about topics that are not clear to you. Please, there should be no classroom conversations, cell phones, text messaging, eating, sleeping, obscenities, listening to music or other disruptions of the class.

If you have already disclosed a disability to the Disability Services Office (215 Stratton Administration Building) and are seeking accommodations, please feel free to speak with me privately so that I may assist you. If you have an injury sustained during military service including PTSD or TBI, you are also eligible for accommodations under the ADA and should contact the Disability Services Office.

Please let me know if I pronounce your name incorrectly, or use an incorrect gender pronoun, or if you prefer a nickname or a name different from that in the MyKU roster. Feel free to let me know in private.

Week	Lecture Topics[1]
1	Intro to the course, Zoom. Interactive Python 3[2]. Makefile-driven testing.
2	Regular expressions[3] for extracting, cleaning, and formatting data.
3	Assignment 1 using Python regular expressions.
4	Scikit[4] classification, decision trees, Kappa statistic, measures of error.
5	Assignment 2 for classification of discrete target attributes (dependent variables).
6	Regression, model trees. Correlation coefficient & other measures of error.
7	Assignment 3 using regression to predict numeric target attribute values.
8	Bayesian techniques in scikit-learn.
9	Scripted invocation of Weka using bash and Python scripts.
10	Assignment 4 invoking Weka in command-line mode from scripts.
11	Instance-based learning & ensemble learning in sci-kit.
12	Clustering in sci-kit. Start of individual Assignment 5 project.
13	Overview of neural nets for signal-based learning. Consolidation.
14	Student presentations of individual projects.
15	Student presentations of individual projects.

We will be using Zoom for remote attendance during class time. Recorded archives of class sessions will be available within a day. We will go over Zoom & recording permissions in the first class.

[1] This is a schedule-in-progress that may need revision as we go along. We are proving in the 2nd offering of this course.
[2] https://ipython.org/, https://www.python.org/ We use Python 3.x.
[3] https://docs.python.org/3/library/re.html, https://pythex.org/
[4] https://scikit-learn.org/stable/