Guidelines for Preparing for Kutztown University's Graduate Certificate Program in Data Analytics in the Department of Computer Science & Information Technology

1. Overview

This four-course, twelve-credit program, detailed on this web page, is a stand-alone graduate certificate program whose credits can be applied to the Master of Science in Computer Science tracks and the Master of Business Administration. The following pages provide additional information on these programs:

- MS in Computer Science/Software Development
- MS in Computer Science/Information Technology
- MS in Computer Science/Interdisciplinary
- Master of Business Administration

Prerequisites for successful entry into the program include entry-level experience programming in a modern scripting language, preferably Python version 3, and completion of an undergraduate course in applied statistics or equivalent. A subset of the assignments in CSC458 Data Mining and Predictive Analytics I and CSC459 Introduction to Big Data use Python to retrieve, verify, clean, and format data for subsequent analysis. All courses depend on a general understanding of statistical concepts as a basis for analyzing data relationships.

A semester or summer of self-directed study using the following free, on-line tutorials is sufficient to prepare students for taking courses in the Graduate Data Analytics Certificate program.

2. Python preparation

All resources linked in this section are free.

How to Think Like a Computer Scientist is a good tutorial for Python version 3 newcomers. Work through chapters 1 through 14, skipping chapter 10 on events.

If you install the most recent stable python 3.x on your own machine, just running **python** will get you the simpler-to-use interpreter. You can download Python 3.x from here. The interactive ipython interpreter includes some useful features such as maintaining a history of user commands. The Python website is at http://www.python.org/.

Documentation including tutorials for the 3.x library is here.

3. Statistics preparation

Successful completion of an undergraduate applied statistics course is enough. The goal is to start this program feeling comfortable with concepts such as statistical distributions, basic measures such as mean, median, mode, and variance, and probabilities. The courses in the program pick up from there.

Perhaps the best way to lay this groundwork is to get hold of an inexpensive, used textbook on applied probability and statistics and work through exercises on the following topics. CSC458 reviews these topics.

- Averages (mean, median, mode), variance, population & sample standard deviation, centiles.
- Statistical distributions such as uniform, Gaussian ("bell-shaped" or "normal" curve), bimodal, and exponential distributions. Related concepts in probabilities.
- The *Pearson r* correlation coefficient.
- Graphing & visualizing numeric relationships.

An apparently good on-line alternative is to use the free Open Learning Initiative course sponsored by Carnegie Mellon University, covering the same topics.

4. Other suggestions

Start projects soon after they are handed out, work at least 50 minutes per work session, and work no more than 15 to 20% of your project time late at night. These suggestions come from analyzing successful software student work habits.

Attend instructor office hours and graduate assistant tutorial hours if you run into problems. Do not wait to address such problems.

Take physical notes. Studies have shown that the act of writing notes by hand helps reinforce and retain facts and concepts.

Cultivate enjoyment of this work. It is especially exciting when applied to an application data domain that interests you.

5. After you start the courses.

This section outlines tools used in the courses. You do not need to use these resources before starting the courses. This information is here for browsing purposes.

CSC458 & CSC 558 Data Mining & Predictive Analytics I & II

After one Python project for data cleaning we use the free Weka toolkit, written in Java, with interaction via a graphical user interface (GUI). There is an on-line tutorial here and video tutorials here. The courses cover all of this material. Some students start with CSC558, depending on course offering schedules.

CSC459 Introduction to Big Data

This course includes map-reduce tools for analyzing large datasets and use of AWS (Amazon Web Services) tools. This is also a good place to start the program when offered before CSC458.

Capstone Courses – you must take one of these three to complete the twelve-credit program.

CSC 570 Independent Study is a one semester, one-on-one research project with an advisor using a data domain of your choice.

OR

CSC523 Scripting for Data Science is a Python-intensive semester of scripting for using Python in data acquisition, cleaning, formatting, visualizing, and analysis. We mostly use the scikit-learn machine learning library. There is also an option for using other libraries or using Python to drive the command-line Weka interface.

OR

CSC590 Internship, approved by your certificate advisor that uses data science knowledge and techniques.

Please contact the Department of Computer Science and Information Technology for more information.