BIOF085: Introduction to Data Science using Python

Introduction

This is a 3-day workshop on using Python for Data Science. Over the three days, we will introduce the Python scripting language and also the major packages used in Python for data science. This is quite a condensed introduction, so we will move at a pretty fast clip. After this workshop, you should know the following skills:

1. Use Python through an integrated development environment (we will introduce two: Spyder and JupyterLab) and/or a shell terminal
2. Gain an understanding of how to use freely available Python packages in your environment
3. Load data into Python
4. Clean, manipulate and munge raw data to make it amenable for analysis
5. Visualize data through statistical graphs and some interactive graphs
6. Run basic statistics and regression models on data
7. Run basic machine learning models on data
8. Have a high-level idea of how to apply Python to various data science problems, including some bioinformatics problems

Instruction Information

Dr. Abhijit Dasgupta
Contact: via Slack (see below) or Canvas

Format

This is an online workshop, which seems an oxymoron. We will use hybrid (in-person and online) materials during this workshop, to help instructor-student engagement as well as provide a modicum of independent, self-paced learning. We expect that you will participate fully in this workshop over the 3 days. In particular, all in-person sessions (see schedule) will assume that you have completed the asynchronous material and progress checks assigned for the previous time period. Our engagement during this workshop will take several forms:

- **Class materials:** All materials, including screencasts, slides, videos, textual handouts and assignments/progress checks, will be available on the FAES Canvas Site under this class's site.
- **In-person engagement:** We will have several sessions of in-person engagement through Zoom. The links will be posted in Canvas, and Zoom can also be accessed through Canvas

- **Communications:** We will have a dedicated **Slack channel, biof085.slack.com**, for this class. You will receive an invitation to join this channel at the e-mail you used during registration, through Canvas. I will be monitoring this Slack channel from 1 week before the workshop starts to 2 days after the workshop ends. You will receive an invitation to join this channel at the e-mail you used during registration. Please promptly join this channel. There will be separate tracks within the channel for installation issues, general Python issues and general data science issues. You can also communicate via Canvas.

Software

We will be using Python 3 for this workshop, since Python 2 is no longer being maintained. In particular, we will be using the Python 3.7 distribution provided by Anaconda. This distribution comes "batteries included" for all the data science work we'll be doing, including all the requisite packages. This distribution is available for Windows, MacOS and Linux.

The Anaconda Python distribution is available at https://www.anaconda.com/products/individual. Scroll down to the bottom of the page and download the Python 3.7 Graphical Installer appropriate for your operating system. You should install the **64-Bit Graphical Installer** unless you have a really old computer.

If you have installation issues, please ping me on Slack **no later than 9pm the day before class starts** so I can help you with any issues. The Anaconda distribution is robust and should install effortlessly on all operating systems, in my experience, but you never know.

Schedule

Day	Time	Format	Topics
Day 1	9am - noon	In-person on Zoom	Why data science in Python? A Python primer for Data Science
	1pm - 2pm	In-person on Zoom	Python tools for data science Data wrangling, cleaning, summarizing and munging
	2pm-4pm	Asynchronous material	Data munging
	4pm-4:30pm	In-person on Zoom	Q & A
Day 2	9am-10 am	In person using Zoom	Data visualization
	10am-noon	Asynchronous material	Data visualization
	1pm-2pm	In person using Zoom	Statistical Analysis using Python
	2pm-4pm	Asynchronous material	Statstistical Analysis using Python
	4pm-4:30pm	In-person using	Q & A

		Zoom	
Day 3	9am-10am	In person using Zoom	Data analytics; Machine Learning
	10am-noon	Asynchronous material	Data analytics; Machine Learning
	1pm-2pm	In person using Zoom	String manipulation; Introduction to bioinformatics
	2pm-3:30pm	Asynchronous material	Introduction to bioinformatics; working with other languages; additional resources
	3:30pm-4:30pm	In-person using Zoom	Q & A, Miscellaneous topics

Asynchronous material

Asynchronous material will be a combination of screencasts, videos, slides and textual material, as well as progress checks. The idea will be to learn for 15 minutes, then complete a progress check, and then move to the next module. I will also have quick progress checks during the in-person sessions as well.

Office hours

I won't have any office hours per se, but I will be available at the end of each day on Zoom, and throughout the 3 days of the workshop on Slack. I will try to respond promptly to all queries, and will either reply individually or, if the question is of general interest and would fill a gap in the materials, on the general slack channel.

FAQs

Q: Do I need to know programming or Python to do this workshop?

A: You will not need to know programming or Python to do this workshop. However, some familiarity with general programming concepts or some experience with a scripted language like R, SAS, Stata or Java would be very useful to pick up concepts, due to the pace of the workshop. I'll try my best to explain the basic concepts thoroughly.

Q: Do I need to know biology or genomics or the like to do this workshop?

A: The material in this workshop is mainly domain-agnostic, except for the section on bioinformatics. I will cover bioinformatics at a high level and show examples, so you do not need a bioinformatics background to do this workshop.

Q: Do you expect me to finish the asynchronous sections before coming to the evening Q & A session?

A: **Yes, absolutely!** If a question is answered in the material, I'll just refer back to it, unless further explanation is needed. I do expect engagement with the material and finishing the progress checks before coming to Q & A. This is in the interests of respecting everyone's time. I will have a hard stop each evening at 5pm for Q & A so that both you and I can rest and recharge for the following day, and you have some opportunity to digest material.

Q: How long are you available each day?

A: I will be available from 8 am to 6 pm each day for questions/comments/explanations on Slack. If there are questions/comments sent after 7pm I'll answer them based on my time availability that evening, but will definitely address them before the next day's session. For the two days after the workshop, I will answer all queries, but maybe not as immediately; you can expect responses around lunch time and at the end of the day.