INLS 490: Programming for Data Analysis
Spring 2020

Basic Information

Date and time: Tuesdays and Thursdays, 11:00 a.m. to 12:15 p.m.
Location: Manning 14

Instructor Information

Instructor: Sayamindu Dasgupta
E-mail: sayamindu@unc.edu
Office: Manning 22
Office hours: Tuesdays from 3 p.m. to 5 p.m. (or by appointment)

Course overview

In a world that is increasingly driven by software and data, developing fluency with the basics of programming and data analysis is a crucial skill. This course will introduce basic programming and data science tools to give students the skills to use data to answer questions about local and online communities.

In particular, the class will cover the basics of the Python programming language, an introduction to web APIs including APIs from Wikipedia and Twitter, and will teach basic tools and techniques for data analysis and visualization. As part of the class, participants will learn to write software in Python to collect data from public datasets and web APIs and process that data to produce numbers, hypothesis tests, tables, and graphical visualizations that answer real questions.

The class will be built around student-designed independent projects. Every student will pick a question or issue they are interested in pursuing and will work with the instructor to build from that question toward an analysis of data that the student has collected using software they have written.

Please note that this course is designed for students with little or no prior programming experience. If you already consider yourself to be knowledgeable about programming, this is probably not the course for you. Furthermore, this introduction to programming is intentionally quick and dirty and is focused on what you need to get things done. If you want to become a professional programmers, this is also probably not the right class. If you want to learn about programming so that you can more effectively answer questions with data by writing your own software and by managing and communicating more effectively with programmers, you are in the right place.

Learning objectives

At the end of this course, you will be able to:

- Write or modify a Python program to collect and read a data-set
- Read web API documentation and write Python code to parse and understand a new and unfamiliar JSON-based web API
- Use both Python-based tools like MatPlotLib as well as tools like LibreOffice, Google Docs, or Microsoft Excel to effectively graph and analyze data
- Use data to effective answer a substantively interesting question and to present this data effectively in the context of both a formal presentation and a written report

Grading

You will be assessed based on the following elements:

- Participation: 30 points
- Project idea: 5 points
- Final project proposal: 10 points
- Final project presentation: 15 points
- Final paper: 40 points

There is a total of 100 points.

Final grades will be assigned according to the following schedule:

A	95 to 100
A-	90 to 94
B+	87 to 89
B	84 to 86
B-	80 to 83
C+	77 to 79
C	74 to 76
C-	70 to 73
D+	67 to 69
D	60 to 66
F	<60

Assessment Details

Project idea
Maximum length: 500 words
Due date: March 3, 11:00 AM
Submission: Sakai

In this assignment, you should concisely identify an community or context that you are interested in a source of data and/or and a list of at least 3-4 questions you might be interested in answering in the context of your final project. I am hoping that each of you will pick an area or domain that you are intellectually committed to and invested in (e.g., your town, or an online community that you participate in). You will be successful if you describe the scope of the problem and describe why you are interested in using the techniques you are learning in this class to tackle this problem.

If you are unsure, asking a question about Wikipedia is probably among the safer paths.

I will give you feedback on these write-ups and will let you each know if I think you have identified a questions that might be too ambitious, too trivial, too broad, too narrow, etc.

Project proposal
Maximum length: 1000 words
Due date: March 17, 11:00 AM
Submission: Sakai

Building on your project idea assignment, you should describe the specific types of data you will collect, the steps you will take to collect the dataset, the limits and strength of these data for answering the question you have selected, and a description of the kinds of report and visualization you will make.

An important step here is going to be framing your analysis. Why is this is an important question? Why do you care? What do we need to know (e.g., about the question, about underlying theories, about your business, about the topic, about the community) to understand this analysis? This will all need to be part of your final project.

I will give you feedback on these proposals and suggest changes or modifications that are more likely to make them successful or compelling and to work with you to make sure that you have the resources and support necessary to carry out your project successfully.

Final project
Presentation date: Week of April 21
Paper due date: April 27, 12:00 PM

For your final project, I expect you to build on the first two assignments to describe what you have done and what you have found. I'll expect every student to give both:

- **A short presentation to the class (10 minutes)**
- **A final report written as a Jupyter notebook that is not more than 2000 words**[1]

I expect that your reports will include text from the first two assignments and reflect comprehensive documentation of your project. Each project should include: (a) the description of the question and community you have identified and information necessary to frame your question, (b) a description of the how you collected your data, (c) the results.

A successful project will tell a compelling story and will engage with, and improve upon, the course material to teach an audience that includes me, your classmates, how to take advantage of programming with data more effectively. The very best papers will give us all a new understanding of some aspect of course material and change the way I teach some portion of this course in the future.

Paper and Code:
Your final project should include detailed information on:

- The problem or area you have identified and enough background to understand the rest of your work and its importance or relevance.
- Your research question(s) and/or hypotheses.
- The methods, data, and approach that you used to collect the data plus information on why you think this was appropriate way to approach your question(s).
- The results and findings including numbers, tables, graphics, and figures.
- A discussion of limitations for your work and how you might improve them.

If you want inspiration for how people use data science to communicate this kinds of findings broadly and effectively, take a look at great sources of data journalism including Five Thirty Eight or The Upshot at the New York Times. Both of these publish an large amount of excellent examples of data analysis aimed at broader non-technical audiences like the ones you'll be communicating with and quite a bit of their work is actually done using Python. A simple Five Thirty Eight story will include a clear question, a brief overview of the data sources and method, a figure or two plus several paragraphs walking through the results, followed by a nice conclusion. I'm asking you to try to produce something roughly like this.

[1] Python code and/or data does not count toward the word limit.

Keep in mind that most stories on Five Thirty Eight are under 1000 words and I'm giving up to 2000 words to show me what you've learned. As a result, you should do more than FiveThirtyEight does in a single story. You can ask and answer more questions, you can provide more background, context, and justification, you can provide more details on your methods and data sources, you can show us more graphs, you can discuss the implications of your findings more. You to use the space I've given you to show off what you've done and what you've learned!

As you will submit a Jupyter notebook as your final paper, I will automatically get to see your code. Make sure that you also submit your data (if you use a copy) with your submission. However, I will not be emphasizing the quality or quantity of your code but rather the degree to which you have been successful at answering the substantive questions you have identified.

Presentation
Your presentation should do everything that your paper does and should provide me with a very clear idea of what to expect in your final paper. I'm going to give you all at least a paragraph of feedback after your talk. This will be an opportunity for me to see a preview of your paper and give you a sense for what I think you can improve. It's too your advantage to both give a compelling talk and to give me a sense for your project.

- *Timing:* All presentations will need to be 'a maximum of 7 minutes long with additional 2-3 minutes for questions and answers. Timing is going to be tight and I'm going to set an alarm and stop presentations that go too long.
- *Presentation order:* Since presentations will happen over two days, I will be creating a random order for you to present in. You will get your time-slot toward the beginning of April.
- *Slides:* You are encouraged to use slides for your talk but I will need your slides ahead of class. If you want to submit slides, you must upload slides in PDF format in advance of the presentation date (specific deadlines will be shared in early April). I'm going to get everything in order on my laptop before class so we can make quick transitions. Because time will be very tight, if you do not submit slides, or if you submit them late, you will not be able to use slides for your talk. There will not be time in class for me to able to load your slides onto the computer.

Weekly coding activities
Every Thursday from January 21st onward will be dedicated to a set of coding activities that will involve changing or adding to code related to the topic of the week. **These coding activities will not be turned in and will not be graded.**

In many cases, you will find yourself continuing to work beyond the class on these activities. I will share my solutions answers to each of the coding activities by the subsequent Monday in a Sakai forum. As you will see over the course of the semester, there are many possible solutions to many programming problems and my own approaches will often be different than yours. That's completely fine! Coding is a creative act!

Please do not share answers to activities before midnight on Sunday so that everybody has a chance to work through answers on their own. After midnight on Sunday, you are all welcome to share your solutions and/or to discuss different approaches. We will discuss the coding activities for a short period of time at the beginning of each class.

Participation
The course relies heavily on participation. The material we're going to be covering is difficult and we're going to be covering it quickly. It is going to be extremely difficult to make up any missed classes. Attendance will be the most important part of participation and missing more than 1 sessuin is going to make it extremely difficult to excel in the class. Participation will be graded according to these criteria:

- *Attendance*
  It is important for you to attend class. Please be seated and ready when class begins. If personal difficulties (serious illness, etc.) make attendance problematic, please consult with me so that we can make an appropriate plan.
- *Deportment*
  You should be attentive in class and respectful of your classmates and the instructor. Turn off cell phones and other devices that might disrupt class. Use laptops and other devices to support current course activities only.
- *Engagement*
  Engagement includes: participating in class activities; responding to discussion questions or other questions that I might ask during a lecture; actively listening and taking notes. I value all informed opinions and encourage you to share them.

Engagement will be weighted more heavily than attendance and deportment.

Course technology

Sakai

Sakai will be used for assignments, forum discussions, and resources. A copy of this syllabus, as well as the textbook will be made available in the resources section of Sakai.

Jupyter notebooks

Although we will be using Python, you will not need to download and install Python on your own laptops. We will be using Jupyter notebooks to write programs. In order to use Jupyter notebooks, you will have to use a web-browser such as Mozilla Firefox or Google Chrome. I will share the link where you can sign in during class.

Semester Calendar

Note: This is a tentative schedule and is subject to change. Any changes will be announced in class and by email.

Apart from weekly coding activities, there will be readings for some of the days. I will announce those in advance (at least a week before) and share the material through Sakai's resource section.

Starting the week of January 21st, every Thursday will be dedicated to working hands-on on a set of coding activities that will involve changing or adding to code that is related to the topic of the week. These coding activities will not be turned in and will not be graded. In many cases, you will have to continue to work on these activities beyond class. I will share my solutions answers to each of the coding activities by the subsequent Monday morning in a Sakai forum. Please do not share answers to activities before midnight on Sunday so that everybody has a chance to work through answers on their own. After midnight on Sunday, you are all welcome to share your solutions and/or to discuss different approaches. We will discuss the coding activities briefly at the beginning of each Tuesday's class.

Date	**Topic**
Thursday, January 09	Introductions
Tuesday, January 14	What is programming?
Thursday, January 16	Getting started with Jupyter notebooks

Date	Topic
Tuesday, January 21	Getting started with Python (part 1)
Thursday, January 23	
Tuesday, January 28	Getting started with Python (part 2)
Thursday, January 30	
Tuesday, February 4	First data set - baby names
Thursday, February 06	
Tuesday, February 11	Playing with words
Thursday, February 13	
Tuesday, February 18	Data from Chapel Hill (part 1)
Thursday, February 20	
Tuesday, February 25	Data from Chapel Hill (part 2)
Thursday, February 27	
Tuesday, March 3	Data from the web: Wikipedia (part 1)
Thursday, March 5	
Tuesday, March 10	*Spring break (no class)*
Thursday, March 12	
Tuesday, March 17	Data from the web: Wikipedia (part 2)
Thursday, March 19	
Tuesday, March 24	Visualizing data
Thursday, March 26	
Tuesday, March 31	Data from the web: Twitter (part 1)
Thursday, April 2	
Tuesday, April 7	Data from the web: Twitter (part 2)
Thursday, April 9	
Tuesday, April 14	Review and final project prep
Thursday, April 16	
Tuesday, April 21	Final presentations
Thursday, April 23	

Policies

Instructor communication

For specific, concrete questions, e-mail is the most reliable means of contact for me. You should receive a response within a day or so, but sometimes it may take 2-3 days. If you do not receive a response after a few days, please follow up. Please keep this in mind when you are scheduling your own activities, especially those related to activities with due dates. If you wait until the day before a due date to ask me a clarification question, there is a good chance that you will not receive a response in time.

It is always helpful if your e-mail includes a targeted subject line that begins with “INLS 490.” Please use complete sentences and professional language in your e-mail.

For more complicated questions or help, come to office hours (no appointment necessary!) or make an appointment to talk with me at a different time. I cannot discuss grades over e-mail; if you have a question about grading, you must talk with me in person.

You are welcome to call me by my first name (“Sayamindu” -- pronounced “Shayomindoo”). However, you may also use “Dr. Dasgupta” or "Professor Dasgupta" if that is more comfortable for you. Any one of those is fine.

Academic integrity

The UNC Honor Code states that:

It shall be the responsibility of every student enrolled at the University of North Carolina to support the principles of academic integrity and to refrain from all forms of academic dishonesty…

This includes prohibitions against the following:

- Plagiarism.
- Falsification, fabrication, or misrepresentation of data or citations.
- Unauthorized assistance or collaboration.
- Cheating.

All scholarship builds on previous work, and all scholarship is a form of collaboration, even when working independently. Incorporating the work of others, and collaborating with colleagues, is welcomed in academic work. However, the honor code clarifies that you must always acknowledge when you make use of the ideas, words, or assistance of others in your work. This is typically accomplished through practices of reference, quotation, and citation.

If you are not certain what constitutes proper procedures for acknowledging the work of others, please ask the instructor for assistance. It is your responsibility to ensure that the honor code is appropriately followed. (The UNC Office of Student Conduct provides a variety of honor code resources.)

The UNC Libraries has online tutorials on citation practices and plagiarism that you might find helpful.

Students with disabilities

Students with disabilities should request accommodations from the UNC office of Accessibility Resources and Service (https://accessibility.unc.edu/).

Acknowledgements and thanks

This syllabus builds on the Community Data Science Course taught by Benjamin Mako Hill and Tommy Guy at the University of Washington. You can find their courses and material at https://wiki.communitydata.science/Workshops_and_Classes