INFO 6101: DATA SCIENCE ENGINEERING WITH PYTHON

Course Information

Course Title: Data Science Engineering methods and tools
Course Number: INFO 6101, CRN 19612
Term and Year: Fall 2020
Credit Hour: 4 credits
Location: Online

About this course

This course will be taught via livestream, synchronous teaching, every week.

Is your secret wish to become a good computer programmer? Or maybe you are an artist or humanities or an accounting person who wants to use computer programming as a tool to broaden and deepen your impact. Or maybe you already know a good deal of programming and are looking to dive deep into numerical computation and data science engineering? If you answered yes to any of these questions, then INFO 6101 is the course for you.

INFO 6101 is unique because it helps people who are totally new to programming get a friendly and fun introduction to programming. For the more experienced, INFO 6101 also provides advanced numerical computation material to prepare them for data science engineering.

We use the Python programming language in this course for two reasons. One, Python is a high- level language that feels almost like English. Two, you can get a lot done in Python by writing only very few lines to software code!

What will you learn in this course?

Basic material for beginners

1. Start programming in Python
2. How to input and output data in Python.
3. How to use Python to “scrape” data from webpages.
4. How to use Python to read data from excel worksheets and sql tables.
5. Store data in Python lists, tuples and dictionaries.
6. How to visualize data in Python in less than 5 lines of code.
7. What are functions and how to write them?
8. Regular expression and why you should be learning about them.
9. Getting Python to do the repetitive and conditional stuff (if, elif and for loops in Python)

10. Write Python in style: the PEP 8 style guide and why you really need to have style while coding!

Advanced material (separate track)

11. Learn tips and tricks for reading in large datasets into pandas.
12. Develop a deep understanding of built-in and abstract datatypes in Python and how one should leverage them for specific tasks.
13. Computational Linear algebra in Python.
14. Learn about numpy, scipy and pyTorch

Students interested in only the basic part of the course will be graded separately from students who are also taking the advanced part of the course !!!

Course Syllabus

In addition to covering the material below, you are expected to work on a project. Project selection will depend on the track chosen by students.

Basic track

Data input & output. Python built-in data structures — lists, tuples and dictionaries.
Hashing explained in a fun way.
Conditionals (if, elif, else) and loops (for) for flow of control.
Writing functions in Python: a great way to modularize python code.
Web scraping with Python: introduction to urllib library in python.
Using regular expressions to search for fuzzy words.
Easy data visualization with the seaborn library

Advanced track

Course introduction — structure, operations, components, and content overview. Why Python over other languages? Brief introduction to Python, and an overview of Python libraries commonly used by data scientists.

Exploring the power of pandas to read data from multiple formats. Basic and advanced pandas operations with specific emphasis on numerical computation.

Data structures in Python: built-in and abstract. Which data structures are suitable for large-scale numerical computation and why? The problem of search and how different data structures hack it.

Linear algebra with Python, numpy and scipy. The use of optimized Python libraries to enable fast numerical compute. Some examples of input-output, web scraping and regular expressions.

Python and machine learning: decision trees to Random Forests. Introduction to ensemble models and bagging. Using excel and Python to create a random forest classifier. More about Random Forests. Hyper-parameter tuning, and exploratory data analysis with random forests and scikit-learn.

Python frameworks for deep learning: Keras-tensorflow Vs Pytorch. Introduction to artificial neural networks and deep learning — why are they popular, examples, their relation to linear algebra. Different neural network architectures. Basics of linear algebra. Neural networks as successive transformations of the input vector.

Data visualization packages in Python and their efficiency: matplotlib, seaborn and Bokeh (interactive visualization). Tips and tricks to machine learn on large datasets. Memory efficient computations.

From prototyping to productizing in Python: considerations.

Meet the instructor

Ramkumar Hariharan is currently head of applied AI with Macro-Eyes, Seattle, where he drives diverse projects in the Artificial Intelligence & Healthcare space. Previously, he has led multiple high-impact data-driven projects at some of the leading institutes in Seattle. These include Fred Hutch, University of Washington (UW), and the Institute for Systems Biology. His areas of focus include data analyses, data visualization, and predictive analytics of both structured and unstructured data.

Ram has a 15-year history of developing and delivering more than 20 computational, biomedical, and data science courses at a variety of levels. His courses, lectures, online teaching, and motivational talks have been overwhelmingly well-received in Seattle, Japan and in India. He has also “edutained” on local and national Television and Radio in India. Ram serves as affiliate faculty at Northeastern University, affiliate of UW e-sciences institute, bootcamp leader at General Assembly, and mentor with Springboard. He has also led education and training programs for Fred Hutch. He specializes in using powerful, yet simple analogies to explain seemingly complex computational and data science concepts and math.

Ram’s teaching philosophy is grounded in one strong belief: there is no one size fits all approach to teach, or to learn a new concept.

His brand statement — telling humorous stories with at least one takeaway!

How is the course going to be delivered?

This course is organized as a series of modules, one per week. Each module will contain

(1) One to two hours of lively, livestreaming sessions.
(2) Slides from the videos
(3) Jupyter notebooks with python code used in the videos
(4) Data for running the examples, assignments and final project
(5) Links to great resources
(6) Textbook suggestions

How are you going to be graded?

Please note that students choosing only the basic track will have less stringent requirements than those who take the advanced track.

There will be three assignments in total, one going out every third week. Each assignment will have questions that either require you to write a response, select a response, or more likely, write Python code to solve a problem. Your performance on the assignments will contribute 70% towards your final grade.

You are required to complete one final project beginning at week 4. Project can span prototyping a data science model to app development. Scores on your project will contribute 30% towards your final grade for this course.

Grade Scale

96-100%	**A**	**87-90.9%**	**B+**	**77-79.9%**	**C+**	69.9% or below F
		84-86.9%	**B**	**74-76.9%**	**C**	
91-95.9%	**A-**	**80-83.9%**	**B-**	**70-73.9%**	**C-**	

Pre-requisites

Interest to learn or check out programming. Willingness to work hard.

Attendance Policy

Please let me know in advance of at least 1 day if you cannot make it to the class.

Late Work Policy

Students must submit assignments by the deadline in the time zone noted in the syllabus.
Students must communicate with the faculty prior to the deadline if they anticipate work will be submitted late.
Work submitted late without prior communication with faculty will not be graded.

Course reviews by previous INFO 6105 students

“Ram is the friendliest professor I have had…” **—** Spring 2020 student

“Ram can teach machine learning to my grandma and she will completely understand it” **—** Spring 2019 student

“I thoroughly enjoyed the course. would 100% recommend your course to anyone interested in starting out with Data Science” **—** Summer 2019 student

“I found the course to be very interesting as its design is very simple and understandable” **—** Summer 2019 student

“Used techniques from your course for my data science internship. Thank you” — Spring 2019 student

How to ask for help and other benefits

Ram and TA’s will be available by email throughout the duration of this course and will gladly help out students with INFO 6105.

Perks: for active data science job seekers, Ram will be happy to leverage his professional network to pass along CVs of students! This has resulted in some of his previous students landing jobs, or sometimes getting interviews from companies!

Text Books

- Beginning Programming with Python For Dummies: Edition 2, John Paul Mueller, 2018, Sold by John Wiley & Sons

- Numerical Python: Scientific Computing and Data Science Applications with Numpy, SciPy and Matplotlib. Robert Johanssen. 2018, Apress.

- Python for Data Analysis, 2nd edition. Wes McKinney. O'Reilly publishing.

- Introduction to Machine Learning with Python: A Guide for Data Scientists, Andreas C. Müller and Sarah Guido

- Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, 2nd Edition, Sebastian Raschka and Vahid Mirjalili

Academic Integrity

A commitment to the principles of academic integrity is essential to the mission of Northeastern University. The promotion of independent and original scholarship ensures that students derive the most from their educational experience and their pursuit of knowledge. Academic dishonesty violates the most fundamental values of an intellectual community and undermines the achievements of the entire University.

As members of the academic community, students must become familiar with their rights and responsibilities. In each course, they are responsible for knowing the requirements and restrictions regarding research and writing, examinations of whatever kind, collaborative work, the use of study aids, the appropriateness of assistance, and other issues. Students are responsible for learning the conventions of documentation and acknowledgment of sources in their fields. Northeastern University expects students to complete all examinations, tests, papers, creative projects, and assignments of any kind according to the highest ethical standards, as set forth either explicitly or implicitly in this Code or by the direction of instructors.

Go to http://www.northeastern.edu/osccr/academic-integrity-policy/ to access the full academic integrity policy.

Student Accommodations

Northeastern University and the Disability Resource Center (DRC) are committed to providing disability services that enable students who qualify under Section 504 of the Rehabilitation Act and the Americans with Disabilities Act Amendments Act (ADAAA) to participate fully in the activities of the university. To receive accommodations through the DRC, students must provide appropriate documentation that demonstrates a current substantially limiting disability.

For more information, visit http://www.northeastern.edu/drc/getting-started-with-the-drc/.

Library Services
The Northeastern University Library is at the hub of campus intellectual life. Resources include over 900,000 print volumes, 206,500 e-books, and 70,225 electronic journals.

For more information and for Education specific resources, visit http://subject-guides.lib.neu.edu/edresearch.

Diversity and Inclusion
Northeastern University is committed to equal opportunity, affirmative action, diversity and social justice while building a climate of inclusion on and beyond campus. In the classroom, member of the University community work to cultivate an inclusive environment that denounces discrimination through innovation, collaboration and an awareness of global perspectives on social justice.
Please visit http://www.northeastern.edu/oidi/ for complete information on Diversity and Inclusion

TITLE IX
Title IX of the Education Amendments of 1972 protects individuals from sex or gender-based discrimination, including discrimination based on gender-identity, in educational programs and activities that receive federal financial assistance.

Northeastern's Title IX Policy prohibits Prohibited Offenses, which are defined as sexual harassment, sexual assault, relationship or domestic violence, and stalking. The Title IX Policy applies to the entire community, including male, female, transgender students, faculty and staff.

In case of an emergency, please call 911.

Please visit www.northeastern.edu/titleix for a complete list of reporting options and resources both on- and off-campus.