DS-GA 3001 Advanced Python for Data Science
Syllabus

Term: Spring 2020
Instructor: Milan Bradonjić mb7535@nyu.edu
Section Leader and Grader: Manikanta Srikar Yellapragada msy290@nyu.edu

Course Description

In this course, we will examine a range of advanced techniques for improving the performance of Python programs, including the use of parallel computation and GPU acceleration. We will also investigate how Python can be used for big data analysis using frameworks such as Apache Hadoop and Apache Spark. Students will have the opportunity to employ these techniques and gain hands-on experience developing advanced Python applications.
The course will take a student-centered, active learning, approach to teaching this material. Class will typically consist of a short introduction to programming techniques, followed by hands on computing exercises.

Course Objectives

Upon successfully completing this course students will be able to:

- write relatively advanced, well structured, computer programs in Python;
- understand principles and techniques for optimizing the performance of Python numeric applications
- understand parallel computing and how parallel applications can be written in Python
- experiment with developing GPU accelerated Python applications
- develop Python applications that utilize big data services such as Hadoop and Spark

Course Requirements

The course comprises weekly 100 minutes lectures, 50-minutes of lab sessions, and they will be a mixture of direct instruction and interactive activities. The lab sessions will provide students with

the opportunity to review the lecture material and address any concerns they may have. Students will also be expected to spend time studying outside of class, mainly tackling the homework assignments and the projects.

Resources

There is no primary textbook for the course. The following texts provide very useful information:

- Introduction to Parallel Computing, Ananth Grama, Anshul Gupta, George Karypis, Vipin Kumar, Pearson; 2 edition (January 26, 2003), ISBN 978-0201648652
- Big Data: Principles and best practices of scalable realtime data systems, 1st Edition, Nathan Marz, James Warren, ISBN 978-1617290343

The syllabus and other relevant class information, resources and notebook will be posted at NYU classes.

Evaluation Plan

There will be regular (weekly based) assignments and a final project. These elements will be combined into a course average using the following weights:
Assignments 60%
Project Proposal 10%
Final Project 30%

1 Grading

All programming assignments and the final project will be graded based on the following criteria:

- Program correctness
- Completely addressing the requirements of the assignment
- Able to handle invalid input correctly
- Able to handle exceptions correctly
- The adequate and judicious use of comments
- Correctly following assignment instructions

In addition, the final project will be graded based on:

- Demonstrated understanding of advanced Python concepts
- Demonstrated understanding of the final project problem
- Quality of documentation
- Quality of the presentation.

2 Student's Responsibilities

Students are expected to read/view assigned material of the class for which they are scheduled, attend class, participate in class, complete assignments, complete projects, and ask for help early if they are having trouble.

3 Instructor's Responsibilities

I expect myself to read/view the material prior to the class for which they are scheduled, prepare and deliver high quality introductions to the material, prepare exercises and assignments that are relevant to research in data science, and provide comments on assignments and projects intended to help students develop their abilities to work with computers and data.

4 Academic Honesty

NYU students and faculty to maintain the highest standards of academic honesty. Students can find information on the core principles and standards in the university's policy on academic integrity, which is accessible at http://www.nyu.edu/about/policies-guidelines-compliance/policies-and-guidelines/academic-integrity-for-students-at-nyu.html.

5 Class and Classroom Manners

I do not take attendance and therefore I expect that if you are in class you are here to learn. So, please, turn off your cell phones, resist the urge to send email and text messages, etc. Basically, I'm just asking that you be respectful of your fellow students and myself. This class is a collaborative learning experience. If you have already finished with what we are working on then find another student to help.

6 Fairness

I will do my absolute best to make this a fair class. If you are having problems in the class, or just not doing as well as you would like, I strongly encourage you to approach me as soon as possible to get help during the semester. Please do not approach me at the end of the semester and ask me to change your grade, allow you to do extra credit, etc. Your grade will be the one you have earned and I am ethically required to report that grade. Of course, if I've made a mistake grading, I encourage you to let me know.

7 Schedule of Classes

This course will comprise 14 lectures and 14 labs over a period of 14 weeks, and attendance will be mandatory.

Week: Lecture Content

Week 01: Shell
Week 02: Python Performance Tips
Week 03: The itertools module
Week 04: Python Performance Tuning
Week 05: Cython
Week 06: Numba
Week 07: Project proposal presentations
Spring Recess - No Class
Week 08: Optimization in Python
Week 09: Python Concurrency
Week 10: Parallel Programming
Week 11: Parallel Programming
Week 12: Python for GPUs
Week 13: BigData with PySpark
Week 14: Final project presentations