McCormick

Northwestern Engineering

Civil & Environmental Engineering

Data Science for Urban Systems

Civil and Environmental Engineering Department
McCormick School of Engineering

Instructor: Ying Chen
Office: TECH A224 or TC Room 214
In-person Office Hours: TBD
On-line Office Hours: Flexible but by appointment
Email: y-chen@northwestern.edu
Website: http://sites.northwestern.edu/ych168/

Textbook:

- Think Python – How to Think Like a Computer Scientist
- Data Visualization in Python
- Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning

Class Times: MW 10:00 ~10:50am, F 11:00 ~ 11:50am
Location: TBD

Software: Python

Course Description

Data Science in Urban Systems introduces most state-of-the-art data science concepts, techniques, and teaches students to select and apply the right algorithms to solve problems.

As an important component of big data processes, data management plays a critical role. Therefore, this course aims to close the gaps within students' data science knowledge toolbox. This course will help students to gain the fundamental knowledge and skills they need in data engineering such as its ecosystem, lifecycle, and tools to manage data for implementing data analytics/science and building machine learning models to solve urban systems challenges like optimizing transportation, urban design and energy portfolios.

Who should take this class?

This course is designed for students who want to learn to program in python for data science. This course guides students in work through basic Python programming language, from basic concepts, and data processing to final data analysis using the correct python and external packages.

In order to present this idea clearly and increase relevance for engineering students, we will use applications from the urban systems areas (including transportation, urban planning, and civil and environmental engineering, but not limited to those fields). At the end of this course, students will gain a competitive edge by tapping into the power of data science.

Prior Expectations:

- Basic familiarity with computer programming in other languages at the introductory level (i.e., R, Matlab, SAS, etc.) and no prior knowledge of *Python* is required;
- Basic college-level math knowledge (probability/statistics/matrices) (such as, CIV_ENV 306).
- Since this course includes some hands-on labs, students are expected to install Python and relevant packages (recommend from Anaconda) and bring their laptop to every class.

Learning Objectives

- To provide students a solid starting point for using Data Science in their work and research;
- Students will be able to understand and use standard sequential, conditional, and iterative control structure of automated data analysis through computers;
- To familiarize students with leading tools used in modern data science practice;
- To help students understand how to manipulate data (store, query, and summarize) using a database designed to analyze structured data;
- To help students understand and use computer programming to collect, analyze and visualize data related to various urban systems challenges;

ABET Program Outcomes

- An ability to identify, formulate, and solve complex engineering problems by applying principles of engineering, science, and mathematics.
- An ability to communicate effectively with a range of audiences.
- An ability to function effectively on a team whose members together provide leadership, create a collaborative and inclusive environment, establish goals, plan tasks, and meet objectives.
- An ability to develop and conduct appropriate experimentation, analyze and interpret data, and use engineering judgment to draw conclusions.
- Ability to acquire and apply new knowledge as needed, using appropriate learning strategies.

Tools

- Jupyter Notebook (https://jupyter.org/)
- Scientific computing package Numpy (https://numpy.org)
- Using Python and SQL to create and query relational databases, in particular MySQL
- Pandas(http://pandas.pydata.org/) for handling and visualizing structured data and time series
- Machine learning packages: scikit-learn(http://scikit-learn.org/stable/documentation.html)

Tentative Schedule

It is a tentative schedule of lectures for this course. We will try to keep approximately on this schedule. (Note that we may change the agenda during the quarter.)

Schedule	Topics
Week 1	Introduction to basic programming concepts Install and run Python programs
Week 2	Sequences: Strings, Lists, and Files
Week 3	Data structures (List, Dictionary, Tuple and Set) in python
Week 4	Modules, Functions, parameters, and scientific computing: NumPy
Week 5	Database operations and Preparation: SQLite and MySQL
Week 6&7	Data Processing: Pandas; Project Version Control: Git and GitHub
Week 8	Data Visualization: Matplotlib, Plotly, Seaborn, Bokeh, and GeoPandas etc.
Week 9	Machine Learning: scikit-learn, NLTK
Week 10	Machine Learning in Urban areas
Week 11	Project Presentation

Websites for Instruction

Canvas

We will use Canvas to distribute readings, assignments, and grades.

DataCamp

I applied for a datacamp classroom for our course. You will automatically have full access to the entire course curriculum on DataCamp. Please use the link below to join the classroom. This is a good source for self-learning.

Kaggle (or Github)

Sometimes, I will run the code using Kaggle notebook in class and you may consider using it for your presentation or your collaboration work for your final project.

OneDrive

I will share the large dataset with some of you depending on the project you plan to work on.

Assignments

We have five homework assignments. These assignments are mainly from the lectures. They will cover basic data querying, data preprocessing, data visualization etc. These assignments will help you understand concepts and ideas you've learned from lectures. You need to submit a report and your code at the same time.

Plagiarism Policy: For any programming course, students may attempt to submit homework that is not coded by themselves. Please keep in mind that it is not hard to detect copying of programs although a program is modified to try to hide its source. **Copying a program, or letting someone else copy your program, is a form of academic dishonesty and the penalties can be found here.**

Late Assignment Policy: the penalty is **10%** off the grade of your project or each assignment for every additional day after the deadline.

Project

We will have a group-based class project using real data. The size of each self-assigned group is three at maximum. Each group will be assigned a case with real data and problems in the real world. Each group also can use existing online datasets or download their own datasets from online resources, like Facebook, Twitter, Yelp, etc. We expect each group could generate a technical report to show some interesting findings by running existing big data analysis algorithms. We encourage each group/student to use the dataset relevant in their fields. You need to submit a detailed technical report along with the source code.

Grading

Your final grade will be composed of the following items:

Attendance:	1% * 10 = 10%
Assignments:	10% *5 = 50%
Midterm Exam:	15% * 1 =15%
Final Project:	25% * 1 = 25%

Letter grades are assigned as follows:

Points Letter Grade Percentage

A	100 – 90
A-	89 – 85
B+	84 – 80
B	79 – 75

B-	74 – 70
C+	69 – 65
C	64 – 60
F	Below 60

Office Hours, E-mail

Your visits, whether online or in-person, are not confined to my regular office hours. However, I prefer scheduling appointments outside of these hours via email. Even during regular office hours, please email to confirm your visit, as this helps me manage any potential conflicts. Email is my preferred communication method, as I typically respond to messages within one day of receiving them.