McCormick

Northwestern Engineering

Civil & Environmental Engineering 474
Data Analytics for Urban Systems

Civil and Environmental Engineering Department
McCormick School of Engineering

Instructor: Ying Chen
Office: Tech A224 /TC Room214
In-person Office Hours: Friday, 10 AM~12:00 PM (by appointment)
On-line Office Hours: Flexible but by appointment
Email: y-chen@northwestern.edu
Website: http://sites.northwestern.edu/ych168/

Textbook: 1) Mining of Massive Datasets (Edition 3)
Hardcopy: https://www.amazon.com/Mining-Massive-Datasets-Jure-Leskovec/dp/1108476341
E-version: Free available http://www.mmds.org/
2) Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems 2nd Edition
3) Data Visualization in Python
4) Python Machine Learning

Class Times and Locations:
Thursdays, 9:00AM ~ 12:00PM
Location: TECH LG52

Software: Python

Course Description

We live in a world occupied by various information. Big data is everywhere. With the rapidly evolving of the web technology and mobile use, people are becoming more and more enthusiastic about interacting, communicating, and sharing with each other through different social platforms and media. In recent years, this collective intelligence has spread to many different domains, with a particular focus on e-commerce, healthcare, and social network, causing the volume of user-generated data to expand exponentially. The extraction of knowledge from such a large amount of unstructured dynamically changed is a challenging task. Those typical data include social comments from Facebook, online customer reviews, Twitter and other popular social platforms, shopping transaction records, mobile messages, financial news and climate data, etc. In the transportation field, mobile devices like GPS or apps in the smartphone make it possible to track vehicle traces, and some traffic surveillance data including speed, link counts, etc. also generate big data in large volumes.

However, the methods, models and algorithms that are used in the transportation field to mine and explore data from estimation, prediction, validation of traffic to transportation theories and models may not perform well under the new situation. The same issue also exists in other fields.

Data Analytics is a graduate-level class, which introduces most state-of-the-art data analytical concepts, techniques, and right algorithms to solve problems.

In this course, we will cover the basic concepts of big data framework presented by Hadoop and Spark and how to build a data pipeline. We also will include some algorithms in data mining, machine learning, and social network analysis. We will summarize recent research in big data applications that could help establish fundamental knowledge, concepts, and technologies related to the specific data analytics task. To present this idea clear, we will take the application in transportation and traffic engineering as an example. More importantly, we will cover how to solve large-scale data problems using right algorithms (such as Deep learning, SVM, XGBoost). The ultimate goal of this course is to master the basic data science techniques and analytical tools for solving problems through hands-on experiences and projects.

This course has some prerequisites: data mining and information retrieval techniques (optional); basic computer programming skills; basic college-level math knowledge (probability/statistics/matrices). Since the big data have been evolved quickly and is a newly emerging topic in transportation, we do not have a specific and fixed curriculum. The primary format of this course will be teaching, class discussion, hands-on case study, and projects.

Objectives

1. To provide students a *starting point* for Data Analytics in their work and research.
2. To present students to the data pipeline and different data science platforms.
3. To introduce students to the popular algorithms and methods in Data Analytics and Data Science.
4. To expose students to recent study in Data Analytics/Data Science.
5. At the end of this course, each student should successfully generate a Data Analytics report.

Tentative Schedule

It is a tentative schedule of lectures and readings for this course. We will try to keep approximately on this schedule.

(Note that we may change the agenda during the semester. Chapters are in the book: Mining of Massive Datasets; HML Chapters are in the book: Hands-on Machine Learning; DV Chapters are in the book: Data Visualization in Python); PM: Python Machine Learning.

Weeks	Topics	Readings	Handouts	Hand-ins
Week2 (April 4)	Introduction to Big Data	Chapter 1. Data Mining	Syllabus, Project Topic List	
Week3 (April 11)	Data Exploration	PM Chapter 4	HW1	
Week4 (April 18)	Data Visualization	DV		
Week5 (April 25)	Machine Learning In Transportation	Other Materials	HW2	HW1
Week6 (May 2)	Classification *Supported Vector Machine	PM Chapter 3 HML: Classification		
Week7 (May 9)	Tree-based Methods	Other Materials	HW3	HW2
Week8 (May 16)	Clustering, Reinforcement Learning	Chapter 7: Clustering HML: Unsupervised Learning Techniques		
Week9 (May 23)	Neural Network Deep Learning	PM Chapter 13 HML: Training Deep Neural Networks		HW3
Week10 (May 30)	Text Mining, Topic Modelling Network Analysis	PM Chapter 8 Chapter 10: Analysis of Social Networks Community Detection in graphs HML: NLP with RNS and Attention		
Week11 (June 6)	Project Presentation		Presentation Schedule	Report and Code

---

Websites for Instruction

Canvas

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

DataCamp

I applied a datacamp classroom for our course. You will automatically have full access to the entire course curriculum on DataCamp from March 31 to October 6. 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 to use it for your presentation or your collaboration work for your final project.

Piazza

We are going to use Piazza to create a virtual community to exchange ideas and interact with your classmates.

OneDrive

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

Assignments

We have three homework assignments. These assignments are mainly from the lectures. They will cover basic data visualization, decision tree, k-Means, text mining or social network analysis, 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 a programming course, a few people inevitably submit the 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 class project for each group. The size of each group is two at maximum. Each group will be assigned a case with the real data and problems in the real world. Each group also can use existing online datasets or download your 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 in their fields. You need to submit a detailed technical report along with the source code.

Grading

Your final grade will be composed from the following items:

Attendance: 1% * 10 = 10%

You is expected to show in person each class, but if you have any conflicts, please send me an email in advance.

Class participation: 5% * 2 = 10%

Sometimes I will bring some open questions for the next lecture, and you will get something to read or think about it in advance. Please be prepared for a three or five-minute presentation.

Assignments: 10% *3 = 30%

Final project: 50% *1 = 50%

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 online/in-person office visits are certainly not limited to my regular office hours, but appointments by email preferred for non-regular office hour time. Even my regular office hours, if you could send me an email to confirm that will be great in case I have any other conflicts. Email is a good way to communicate with me since I usually answer messages within one day of receiving them.