McCormick

Northwestern Engineering

Civil & Environmental Engineering 495
Data Analytics for Transportation and Urban Infrastructure Applications

---

Civil and Environmental Engineering Department
McCormick School of Engineering

Instructor: Ying Chen
Office: TC Room 214
Office Hours: Friday, 3:00~5:00 PM
Phone: (847)-467-2948
Email: y-chen@northwestern.edu

Textbook: 1) Mining of Massive Datasets (and some other optional textbooks)
Hardcopy: www.amazon.com
E-version: Free available
http://infolab.stanford.edu/~ullman/mmds/book.pdf

Class Times and Locations: Th, 9:00AM~12:00PM (Tech M349)

Software: R or 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 includes 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 MapReduce. 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. In order 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. The ultimate goal of this course is to master the basic data analytical techniques and 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 introduce students to the popular algorithms and methods in Data Analytics;
3. To expose students to recent study in Data Analytics;
4. 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.)

Weeks	Topics	Readings	Handouts	Hand-ins
Week1 (April 5)	Introduction to Big Data	Chapter 1. Data Mining	Syllabus	
Week2 (April 12)	Data Exploration	Py 4		
Week3 (April 19)	Data Visualization	Data Visualization Tools and Case Study	HW1	
Week4 (April 26)	Classification Supported Vector Machine	Py 3	Project Topic List	
Week5 (May 3)	Tree-based Methods	Other Materials		Project Topic

Week6 (May 10)	Clustering, Finding Similar Items (optional)	Chapter 7: Clustering Py 11		HW1 TBD
Week7 (May 17)	Neural Network Deep Learning	Py 13	HW2	
Week8 (May 24)	Text Mining, Topic Modelling Sentiment Analysis (optional)	Py 8	TBD	TBD
Week9 (May 31)	Network Analysis	Chapter 10: Analysis of Social Networks Community Detection in graphs	Presentation Schedule	
Week10 (June 7)	Project Presentation or Makeup Lecture			HW2
Exam Week (June 14)	Project Presentation			Report and Code

Assignments

We have two 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 50% off the grade of your project or each assignment.

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: 2% * 10 = 20%

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 one or two-minute in-class presentation. Depending on the time, I may randomly ask some students to present their findings.

Assignments: 20% *2 = 40%

Final project: 40% *1 = 40%

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

I am on campus for most of the day, and you are welcome to come in anytime if you have any questions. Your 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.