New York University Tandon School of Engineering
Department of Computer Science & Engineering

CSUY 3943 D — Fall 2020
Mining Massive Datasets — Professor Sandoval

Contact
Email: gustavo.sandoval@nyu.edu
My background is here.

Student hours:

- Monday 2 – 3pm
- Zoom: nyu.zoom.us/my/Sandoval

Student Assistants:

See Piazza for their student hours.

Course Pre-requisites

- CS-UY 2134 (Data Structures and Algorithms)
- CS-UY 3224 Intro to Operating Systems
- Some undergraduate probability and linear algebra. No ML experience is expected for this class

Course Description

This course is an introduction to Data Mining. Data mining is the art of extracting useful patterns from large bodies of data. This involves both methods for discovering possible patterns, and methods for checking or validating candidate patterns. Data Mining is related to statistics and Machine Learning, but has its own aims and scope. We will use tools from statistics and Machine Learning, but we will use them as tools rather than as things to study in their own right.

Topics that we will discuss include: MapReduce/Spark, PageRank, Mining Social-network graphs, Mining Data Streams, Recommendation Systems, and Large-Scale Machine Learning among others. The course will primarily consist of technical readings, discussions and hands-on practice.

After taking this class, if you are faced with a new data mining problem you will be able to:

1. Select the appropriate method for pattern discovery and validation
2. Use the right libraries or custom code to implement those methods
3. Critically evaluate the results and present them to non-technical peers

Prerequisites:

- You should be comfortable with the basic principles of Data Structures and Algorithms at the level of (CS-UY 1134) and also able to write a non-trivial computer program.
- You should have a good knowledge of how the Operating Systems at the level of (CS-UY 3224)
- You should have good knowledge of Python since most assignments will require the use of Spark and Colab notebooks.
- Familiarity with basic probability theory
- Familiarity with basic linear algebra

Readings

Textbook:

- Leskovec, Rajaraman and Ullman. **Mining of massive datasets**. Available online for free and also for sale at the NYU bookstore.

Useful references:

- Tan, Steinbach, Karpatne and Kumar. **Introduction to Data Mining**. 2nd edition. Pearson. 2019

Course requirements

- Attendance will not be taken, but it is highly recommended and it will help with your participation.
- Assignments must be received by midnight on the day they are due. Late homework will not be accepted.

Cooperation Policy

You will work individually on every assignment. You may discuss solutions with your classmates but stop short or sharing your code with them.

Academic Honesty

All work submitted in this course must be your own. Cheating and plagiarism will not be tolerated. If you have any questions about a specific case, *please ask me.*

NYU Poly's Policy on Academic Misconduct:
http://engineering.nyu.edu/academics/code-of-conduct/academic-misconduct

Course schedule (**Tentative**)

The course schedule is tentative, it's likely to change as the weeks go on.

Theme	Topic	Reading
Parallel Processing	Course Information. What is Data Mining.	Chapters 1 and 2
	Distributed File Systems	
	Map Reduce	
	Spark	
High Dimensional Data	Locality Sensitive Hashing	Sections: 3.1 – 3.8
	Clustering	Sections: 7.1 – 7.4
	Dimensionality Reduction	Sections: 11.4
Infinite Data	Mining Data Streams	Sections: 4.1 – 4.7
	Computational Advertising	Chapter: 8
Graph Data	PageRank	Sections: 5.1 – 5.5
	Analysis of Social Networks	10.1, 10.2 and 10.6
Applications	Recommender Systems	Chapter: 9
	Association Rules	Chapter: 6
Machine Learning	Perceptron	Chapter: 12

	Support-Vector Machines	
	Nearest Neighbors	
	Decision Trees	

Grading

Grading will be based on the following weights.

30% Colab Notebooks (Will drop lowest)
30% Homework (Will drop lowest)
20% Exam (take home)
20% Final Project

Grading Schema:

A	95
A-	90
B+	87
B	83
B-	80
C+	77
C	73
C-	70
D+	67
D	60
F	0

Other Grading notes:

Please take the following into consideration during and after the semester and save yourself one or many emails.

1) **I must grade every student EXACTLY the same way.** To this end, I cannot give you special consideration as a result of your academic status (probation or otherwise), scholarships, work status, family situation, visa status, race, color, creed, religious beliefs, past alien abductions, current moon cycle, location of the sun in the sky or anything other than your academic performance. **Your grade must be based on your academic performance in my class**.

2) **I cannot change your grade simply because you ask me to.** Your grade is calculated based on your performance from the first day of class to moment you turn in the final exam.
3) **I will not give you additional work.** Please remember that I must treat all students the same, so if I give you additional work, I would have to give it to the entire class. This is unfair to the students who complete their work on time.
4) **Your grade is a measure of your performance in my class.** If you receive an "F" it is because you have demonstrated that you do not understand the material in the course; if you receive an "A" it is because you have demonstrated that you fully understand the material covered in the course. Other grades are assigned accordingly.

Moses Center Statement of Disability

If you are student with a disability who is requesting accommodations, please contact New York University's Moses Center for Students with Disabilities (CSD) at 212-998-4980 or mosescsd@nyu.edu. You must be registered with CSD to receive accommodations. Information about the Moses Center can be found at www.nyu.edu/csd. The Moses Center is located at 726 Broadway on the 3rd floor.

NYU School of Engineering Policies and Procedures on Academic Misconduct – complete Student Code of Conduct here

A. Introduction: The School of Engineering encourages academic excellence in an environment that promotes honesty, integrity, and fairness, and students at the School of Engineering are expected to exhibit those qualities in their academic work. It is through the process of submitting their own work and receiving honest feedback on that work that students may progress academically. Any act of academic dishonesty is seen as an attack upon the School and will not be tolerated. Furthermore, those who breach the School's rules on academic integrity will be sanctioned under this Policy. Students are responsible for familiarizing themselves with the School's Policy on Academic Misconduct.

B. Definition: Academic dishonesty may include misrepresentation, deception, dishonesty, or any act of falsification committed by a student to influence a grade or other academic evaluation. Academic dishonesty also includes intentionally damaging the academic work of others or assisting other

students in acts of dishonesty. Common examples of academically dishonest behavior include, but are not limited to, the following:

1. Cheating: intentionally using or attempting to use unauthorized notes, books, electronic media, or electronic communications in an exam; talking with fellow students or looking at another person's work during an exam; submitting work prepared in advance for an in-class examination; having someone take an exam for you or taking an exam for someone else; violating other rules governing the administration of examinations.
2. Fabrication: including but not limited to, falsifying experimental data and/or citations.
3. Plagiarism: intentionally or knowingly representing the words or ideas of another as one's own in any academic exercise; failure to attribute direct quotations, paraphrases, or borrowed facts or information.
4. Unauthorized collaboration: working together on work meant to be done individually.
5. Duplicating work: presenting for grading the same work for more than one project or in more than one class, unless express and prior permission has been received from the course instructor(s) or research adviser involved.
6. Forgery: altering any academic document, including, but not limited to, academic records, admissions materials, or medical excuses.

NYU School of Engineering Policies and Procedures on Excused Absences – complete policy here

A. Introduction: An absence can be excused if you have missed no more than **10 days of school**. If an illness or special circumstance has caused you to miss more than two weeks of school, please refer to the section labeled Medical Leave of Absence.

B. Students may request special accommodations for an absence to be excused in the following cases:

1. Medical reasons
2. Death in immediate family
3. Personal qualified emergencies (documentation must be provided)
4. Religious Expression or Practice

Deanna Rayment, deanna.rayment@nyu.edu, is the *Coordinator of Student Advocacy, Compliance and Student Affairs* and handles excused absences. She is located in 5 MTC, LC240C and can assist you should it become necessary.

NYU School of Engineering Academic Calendar – complete list here.

The last day of the final exam period is _December 21st ____. Final exam dates for undergraduate courses will not be determined until later in the semester.

Also, please pay attention to notable dates such as Add/Drop, Withdrawal, etc. For confirmation of dates or further information, please contact Susana: sgarcia@nyu.edu