COSC 4610/5610: Data Mining

Fall 2019

Instructor:	Shion Guha	**Office Hours:**	Tue-Thu 2-3:30 pm
Email:	shion.guha@marquette.edu	**Office:**	318/240 Cudahy Hall

Description: Techniques for extracting and evaluating patterns from large databases. Introduction to knowledge discovery process. Fundamental tasks including classification, prediction, clustering, association analysis, summarization and discrimination. Basic techniques including decision trees, neural networks, statistics, partitional clustering and hierarchical clustering.

Real Talk: I will be teaching this course as a rigorous, graduate level research seminar in computational social science (still accessible by upper level undergraduates) from the perspectives of modern machine learning approaches. We will focus on relevant social network theory, network science models (for modeling structure of networks) and topic modeling (for analyzing content of networks). This is not a "survey of methods" course. This course will focus on real world applications as most students (undergraduate or graduate) go into industry where you focus on "social/people" data. Finally, because most people analyze data extremely poorly, we will have a strong focus on algorithmic biases and ethics on social data as befits your Marquette education.

Prerequisites: cosc 1010 (intro programming) or cosc 2100(data structures) or equivalent. I will assume that you either know or will pick up the basics of git and latex.

Location: Tue-Thu, 5:00-6:15 pm, 108 Cudahy Hall

Github: `https://github.com/shionguha/cosc5610-datamining-fa19`

Piazza: `https://piazza.com/class/jzpqydafpt14ub`

Books: This course is a research seminar. There are plenty of articles and books for this course. All article readings are available as online pdfs and you don't have to "buy" all the books per se though many are quite cheap; I will provide some pdfs and many are available online. however, think about these books as investments in your data science careers. They will be extremely useful beyond the classroom. Regardless, following my usual academic philosophy, I will not make any of these books "mandatory" for purchase but will instead **strongly recommend** them. They are relatively cheap on Amazon and the Marquette Librarians are pretty amazing at getting books via the Interlibrary Loan system. Make judicious use of these resources!

- Social Network Analysis by John Scott
- Networks, Crowds and Markets by Easely and Kleinberg.
- Text Mining with R by Silge and Robinson

Objectives: This course aims to develop skills that can translate to building ethical, human centered, data products and services within organizations. At the end of the course, a successful student should be able to:

- design and develop data mining skills to understand structure and content of social networks
- understand modern network dynamic models based in machine learning perspectives,
- understand a modern understanding of topic modeling and other text mining approaches,

- evaluate complex, social networks with data-driven models.

Additional Graduate Student Objectives: In addition to the objectives outlined above, graduates students are expected to:

- critically read, analyze and respond to current state-of-the-art literature in data science.
- develop and lead a final project academic paper that is ”almost-ready” for submission to a reputable computer science conference.

Graduate Student Book Review Sign up: Editable Google Sheet link

Timeline:

- aug 27: introduction to the course; expectations and class policies; machine learning, data mining and computational social science.
- aug 29: introduction to social network theory (scott chapter 1-2)
- sep 3: social network metrics: density and community structures (scott chapter 1-2)
- sep 5: social network metrics: centrality (scott chapter 5)
- sep 10: social network structures: clusters, components, cliques etc. (scott chapter 6)
- sep 12: social network structures: strong and weak ties (easley chapter 3)
- sep 17: social network structures: homophily and social selection (easley chapter 4)
- sep 19: social network structures: relationships and structural balance (easley chapter 5)
- sep 24: social network dynamics: information cascades (easley chapter 16,19)
- sep 26: social network dynamics: small worlds and power laws (easley chapter 18, 20)
- oct 1: social network dynamics: epidemic models (easley chapter 21)
- oct 3: social network content: word counts and sentiment analysis
- oct 8: social network content: text clustering and classification
- oct 10: graduate student book review presentations
- oct 15: graduate student book review presentations
- oct 17: **no class! mid term break!**
- oct 22: evaluation: A/B testing fundamentals
- oct 24: social network content: introduction to topic modeling
- oct 29: social network content: pca/lda/nnmf models
- oct 31: social network content: text mining and hypothesis testing
- nov 5: social network content: text mining and hypothesis testing
- nov 7: **short in-progress group project presentations**
- nov 12: **no class. I am away to CSCW. Work on group projects!**

- nov 14: **no class. I am away to CSCW. Work on group projects!**
- nov 19: social network structure and content: answering research questions
- nov 21: social network structure and content: answering research questions
- nov 26: **no class. tuesday before thanksgiving! Work on projects!**
- nov 28: **no class. thanksgiving!**
- dec 3: final project presentations! (pizza party!)
- dec 5: final project presentations! (pizza party!)

Grading Policy: Both individual as well as group skills will be tested as part of this course. The following are the items that will be assessed as part of this course:

- individual weekly reading responses and discussion on piazza (15%)
- individual class activity performance and participation (15%)
- group project in-progress evaluations (20%)
- group final project presentations (20%)
- group final project report (30%)

Additional Grading Policy for Graduate Students: Commensurate with the standards and expectations for course outcomes, graduate students will have additional grading expectations as follows:

- individual weekly reading responses and discussion on piazza: graduate students will be graded more critically for weekly reading responses. Instead of summarization, graduate students are expected to critically, read, reflect and respond to state-of-the-art literature in more detail and depth than undergraduate students.
- book review and presentation: graduate students will be asked to read a book critically and individually throughout the semester and then present in-class.
- final project report: graduate students will be asked to lead their groups and transform their project report into a final ACM-style academic paper. This should be ”almost ready” for submission into a computer science conference and the instructor will identify several opportunities for students to do so. The instructor will be happy to mentor student-led submissions.

This course will **not** be graded on a curve. The final grades will depend on the following scores:
A: 96 - 100; A-: 91 - 95; B+: 86 - 90; B: 81 - 85; B-: 76 - 80; C+: 71 - 75; C: 66 - 70; C-: 61 - 65; D+: 56 - 60; D: 51 - 55; D-: 46 - 50; F: 0 - 45;
There are no regrade requests.

Course Policy:

- You are responsible for your own progress. Please check your marquette email, the course github repo and piazza about course announcements and news regularly.
- This is a course that utilizes active learning principles. As a result, there are no traditional lectures and you will be required to do readings for class everyday before arriving. These readings (for any week) will be posted the previous friday by 12:00 pm except the first week, when it will be posted on saturday. You will be required to post reading responses on piazza as part of class participation the day before (monday and wednesday) by **9 pm**.

- You are expected to ask, discuss and contribute to questions on piazza. Not only does this help you and enrich the course, but this will also count towards your grade. Please monitor piazza everyday as I will be posting readings, questions and polls there regularly.
- Please bring a computing device. I will demo sample code on how to put into practice some of the principles that we've been learning in class.
- If you don't already, each of you will create your own free github accounts to maintain your project and any code you write that you create for this course. This is part of a data science project's lifecycle and is expected to be shown to employers by data science job applicants.
- If you don't already, you will be expected pick up LaTeX, specifically overleaf. This is the lingua franca in computer science and data science and is often used to write academic or white papers in serious data science teams. I recommend Overleaf as it is the ”Google Docs” equivalent of LaTeX and allows you to write
- All submission of project reports will be online via d2l. Any submission after the deadline will be considered late. In addition, you will be required to share any code that is part of your project submission, you will be required to send a pull request to the class github repository.
- You must cite every reference in your papers and every library (if used) in your project in ACM style. Failure to do so will be regarded as a violation of academic integrity. Please refer to the section on academic integrity for more details.
- We will follow CSCW proceedings format for final project report. CSCW is regarded as the top computational social science conference. Proceedings Formats are here. Specifically, we will be using the ACM Small Overleaf template. Submissions in any other format shall not be accepted.
- Regular attendance is essential to an active learning process. A student who incurs an excessive number of absences may be withdrawn from the class at the instructor's discretion.
- Please make sure your cell phone is turned fully off, or silent. No texting, reading emails, playing games, or anything else. I will reserve the right to ask you to leave the class if you are being distracting or disruptive.

Ott Memorial Writing Center: The Ott Memorial Writing Center offers free one-on-one consultations for all writers, working on any project, at any stage of the writing process. Marquette's writing center is a place for all writers who care about their writing, because every writer can benefit from conversation with an interested, knowledgeable peer. Writing center tutors can help you brainstorm ideas, revise a rough draft, or fine-tune a final draft. You can schedule a 30- or 60-minute appointment in advance (288-5542 or www.marquette.edu/writing-center), but walk-ins (in 240 Raynor or our other satellite locations) are also welcome. The Ott Memorial Writing Center also offers free workshops and hosts writing retreats.

Academic Integrity: Marquette University takes academic integrity very seriously. This is a core part of who we are as reflected by our Jesuit values. All students are required to take the Academic Integrity Tutorial. If you haven't, please go take it right now. All students are required to adhere to the Honor Pledge and follow the Honor Code. Please familiar yourself with the Academic Integrity website. There is a lot of useful information there. I take a zero tolerance policy with violations of academic integrity. All papers and projects are run through a plagiarism detection software. If you are flagged, be assured that we will have a conversation. Let's not have that conversation shall we? If I determine that you have indeed violated academic integrity, you will receive a failing grade for that component of the course or for the entire course depending on the nature and severity of the violation. Please help me to make sure that there are no such incidents.

Accessibility Policy: If you have any accessibility needs, please contact Office of Disability Services (ODS) to register them as soon as possible. ODS works with students with documented disabilities to provide accommodations for their educational needs. The course has been designed for multiple different styles of learning. However, if you have any specific learning styles that you want me to know about which would not be addressed by AES, please reach out to me within the first week of class so I can try to accommodate.