DSCI 553: Foundations and Applications of Data Mining
Syllabus

USC Viterbi School of Engineering

Units: 4

Term — Day — Time: Fall 2020
Location: Online

Instructor: Fred Morstatter, Andrés Abeliuk
Contact Info: fredmors@isi.edu; aabeliuk@isi.edu

Catalogue Course Description

Data mining and machine learning algorithms for analyzing very large data. Emphasis on case studies. We base a large part of our course on **Anna Farzindar**'s own INF 553 course. Hence, while this course material will look very similar to hers, we are doing our own twist.

Expanded Course Description

Data analysis is foundational to everything from social media to high energy physics (see, e.g., https://tinyurl.com/y2a57cc9). It allows for pattern-discovery, such as predictable human browsing behavior, that we can use to improve websites or make policy changes. This course will teach algorithms for mining "Big Data", i.e., data that can be GBs to TBs. By the end of this course we should prepare you to use data mining to solve real-world problems.

Recommended Preparation

INF 550, INF 551 and INF 552. Knowledge of probability, linear algebra, basic programming, and some machine learning.

A basic understanding of engineering principles is required, including basic programming skills; familiarity with the Python language (and Scala) is desirable. Most assignments are designed for the Unix environment; basic Unix skills, including using Bash, will make programming assignments much easier. Students will need sufficient mathematical background, including probability, statistics, and linear algebra. Some knowledge of machine learning is helpful, but not required.

Course Notes

The course will be run as a lecture class with student participation strongly encouraged. There are weekly readings and students are encouraged to do the readings prior to the discussion in class. All of the course materials, including the readings, lecture slides, and homework will be posted online.

Technological Proficiency and Hardware/Software Required

Students are expected to know how to program in Python (Python 3.X strongly preferred). Students are also expected to have their own laptop or desktop computer where they can install and run software to do the weekly homework assignments.

Required Readings and Supplementary Materials

- Rajaraman, J. Leskovec and J. D. Ullman, *Mining of Massive Datasets*
  - Cambridge University Press, 2012.
  - Available free at: http://infolab.stanford.edu/~ullman/mmds.html

In addition to the textbook, students may be given additional reading materials such as research papers. Students are responsible for all assigned reading assignments.

Description and Assessment of Assignments

Homework Assignments: There will be 4 homework assignments. The assignments must be done individually. Each assignment is graded on a scale of 0-100 and the specific rubric for each assignment is given in the assignment. *Each submission will be checked for plagiarism.*

Comprehensive Exam: There will be an exam at the end of the semester covering all of the material covered in the class.

Data Mining Project: An integral part of this course is the course project, which builds on the topics and techniques covered in the class, focusing on extending and evaluating methods to solve problems. Students will write a written proposal for the project, conduct the project, and then write a paper about the project, and present the project in class. Students are encouraged to identify a new problem, apply or extend the methods they learned in class to propose an approach to solve the problem. Emphasis is placed on quantitative evaluation of the approach. Working as a group is permitted if the project is large enough to justify this. A team can consist of no more than 3 persons.

Project Timeline:

- Aug 24 – Sep 21: Identifying team members and project topics
- Sep 21: Proposal due (team member, topics and milestone)
- Oct 3: Mid-term report due (data description, preliminary results)
- Nov 16 (MW section), or 17 (TR section): Project presentations
- Nov 30: Final report due (task and model description, major discovery, lessons learned)

Grading breakdown of the course project:

- Proposal: Not Graded
- Mid-term report: 5%
- Final report: 15% Reports are 5 pages long, describing the goal, existing solutions to the problem and challenges, proposed approach, its evaluation and limitations.
- Presentation: 5% Presentations are 15-20 minutes long, depending on the number of projects.
- Total: 25% of course grade

Grading Schema:

Class Participation	5%
Homework	50%
Comprehensive Exam	20%
Data Mining Project	25%
Total	100%

Grades will range from A through F. The following is the breakdown for grading:

94 – 100 = A	74 – 77 = C
90 – 94 = A-	70 –74 = C-
87 – 90 = B+	67 – 70 = D+
84 – 87 = B	64 – 67 = D
80 – 84 = B-	60 – 64 = D-
77 – 80 =C+	Below 60 is an F

Assignment Submission Policy

Homework assignments are due at 11:59 pm on the due date and should be submitted in Blackboard. Every student has ***FIVE free late days*** for the homework assignments. You can use these five days for any reason separately or together to avoid the late penalty. There will be no other extensions for any reason. You cannot use the free late days after the last day of the class. You can submit homework up to one week late, but you will ***lose 20% of the possible points*** for the assignment. After one week, the assignment cannot be submitted.

Schedule

	Topic	Readings and Assignments	Deliverables/Due Dates
Week 1 *Week of…* 8/24	Introduction to Data Mining, MapReduce	Ch1: Data Mining and Ch2: Large-Scale File Systems and Map-Reduce	
Week 2 8/31	Frequent itemsets and Association rules	Ch6: Frequent itemsets,	Homework 1 assigned
Week 3 9/07	Shingling, Minhashing, Locality Sensitive Hashing	Ch6: Frequent itemsets, Ch3: Finding Similar Items (section 3.5: Distance Measures)	
Week 4 9/14	Shingling, Minhashing, Locality Sensitive Hashing	Ch3: Finding Similar Items	Homework 1 due, Homework 2 assigned
Week 5 9/21	Recommendation Systems: Content-based and Collaborative Filtering	Ch9: Recommendation systems	
Week 6 9/28	Recommendation Systems: Content-based and Collaborative Filtering	Ch9: Recommendation systems	Homework 2 due, Homework 3 assigned
Week 7 10/5	Analysis of Massive Graphs (Social Networks)	Ch10: Analysis of Social Networks	
Week 8 10/12	Analysis of Massive Graphs (Social Networks)	Ch10: Analysis of Social Networks	Homework 3 due, Homework 4 assigned,
Week 9 10/19	Mining data streams	Ch4: Mining data streams	
Week 10 10/26	Clustering	Ch7: Clustering	
Week 11 11/02	Link Analysis	Ch5: Link Analysis	Homework 4 due.

Week 12 11/09	Link Analysis/ Web Advertising	Ch8: Advertising on the Web	
Week 13 11/16* *Last week of class	Advanced Topics	TBA	
Week 14 11/23 (MW section) 11/24 (TR section)	Comprehensive Exam		

Statement on Academic Conduct and Support Systems

Academic Conduct

- **Plagiarism** – Plagiarism is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in *Campus* in Section 11, *Behavior Violating University Standards* https://policy.usc.edu/student/scampus/part-b/. Other forms of academic dishonesty are equally unacceptable. See additional information in *SCampus* and university policies on scientific misconduct, http://policy.usc.edu/scientific-misconduct.
- **Discrimination, sexual assault, and harassment are not tolerated by the instructors nor the university.** You are encouraged to report any incidents to the *Office of Equity and Diversity* http://equity.usc.edu or to the *Department of Public Safety* http://adminopsnet.usc.edu/department/department-public-safety. This is important for the safety of the whole USC community. Another member of the university community – such as a friend, classmate, advisor, or faculty member – can help initiate the report, or can initiate the report on behalf of another person. *The Relationship and Sexual Violence Prevention Services* http://engemannshc.usc.edu/rsvp/ provides 24/7 confidential support, and the sexual assault resource center webpage http://sarc.usc.edu describes reporting options and other resources.

Support Systems

A number of USC's schools provide support for students who need help with scholarly writing. Check with your advisor or program staff to find out more. Students whose primary language is not English should check with the *American Language Institute* http://dornsife.usc.edu/ali, which sponsors courses and workshops specifically for international graduate students. *The Office of Disability Services and Programs* http://sait.usc.edu/academicsupport/centerprograms/dsp/home_index.html provides certification for students with disabilities and helps arrange the relevant accommodations. If an officially declared emergency makes travel to campus infeasible, *USC Emergency Information* http://emergency.usc.edu will provide safety and other updates, including ways in which instruction will be continued by means of blackboard, teleconferencing, and other technology.

Resources for Online Students

The Course Blackboard page has many resources available for students enrolled in our graduate programs. In addition, all registered students can access electronic library resources through the link https://libraries.usc.edu/.