USC Viterbi School of Engineering

INF 553: Foundations and Applications of Data Mining
Units: 4
Term—Day—Time:
Fall 2018 – Tuesday – 3:30-6:50pm

Locations: WPH B27

Instructor: Wensheng Wu
Office: GER 204
Office Hours: 9-9:45am MW and by appointment
Contact Info: wenshenw@usc.edu

Course producer: TBD
Office: SAL computing lab
Office Hours: TBD
Contact Info: TBD

A. **Catalogue Course Description**
    Data mining and machine learning algorithms for analyzing very large data sets. Emphasis on Map Reduce. Case studies.

B. **Expanded Course Description**
    Data mining is a foundational piece of the data analytics skill set. At a high level, it allows the analyst to discover patterns in data, and transform it into a usable product. The course will teach data mining algorithms for analyzing very large data sets. It will have an applied focus, in that it is meant for preparing students to utilize topics in data mining to solve real world problems.

C. **Recommended Preparation**:
    INF 550, INF 551 and INF 552. Knowledge of probability, linear algebra, basic programming, and machine learning.

    A basic understanding engineering principles is required, including basic programming skills; familiarity with the Python language is desirable. Most assignments are designed for the Unix environment; basic Unix skills 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.

D. **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, home works will be posted online

E. **Technological Proficiency and Hardware/Software Required**

Students are expected to know how to program in a language such as Python. 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.

F. Required Readings and Supplementary Materials

- Rajaraman, J. Leskovec and J. D. Ullman, *Mining of Massive Datasets*
  - Cambridge University Press, 2014. (2nd edition)
  - 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.

G. Grading Structure

Homework Assignments: There will be 5 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.

Quizzes: There will be weekly quizzes based on the material from the week before. There is no mid-term for this class.

Final Exam: There will be a midterm and a final exam.

Grade breakdown:

Quizzes	30%
Homework	30%
Midterm	15%
Final	25%
_____	
Total	100%

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

[93 – 100] = A   [73 – 77) = C
[90 – 93) = A-   [70 – 73) = C-
[87 – 90) = B+   [67 – 70) = D+
[83 – 87) = B   [63 – 67) = D
[80 – 83) = B-   [60 – 63) = D-
[77 – 80) =C+   Below 60 is an F

Note that [90, 93) means that your score is greater than or equal to 90 but less than 93. Note that every point in your coursework counts. We will strictly follow the above cut-off and NO roundup will be performed. Note that grades are NOT negotiable!

H. Grading Policy

Homework assignments are due at 11:59pm on the due date and should be submitted in Blackboard. Late homework will be deducted 10% of its points for every 24 hours that it is late. No credit will be given after 72 hours of its due time.

Makeup for quizzes and exams are not permitted unless there are medical emergencies. Doctor notes are needed as proof. Typically no makeups will be given for situations such as interview, job fairs, etc. Students are responsible for scheduling to avoid conflicts with class meeting times and for any missing coursework due to these situations.

Homework and quizzes regrading requests must be made within a week after the solutions have been posted. Grades are final after the regrading period.

I. **Course Schedule: A Weekly Breakdown (may be revised when the course progresses)**

Week	Topic	Readings	Homework
1 (8/20)	Introduction to Data Mining, MapReduce	Ch1: Data Mining and Ch2: Large-Scale File Systems and Map-Reduce	
2 (8/27)	MapReduce & Spark	Ch2: Large-Scale File Systems and Map-Reduce • Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing, Matei Zaharia, et. al., NSDI, 2012. • Dean and S. Ghemawat, MapReduce: simplified data processing on large clusters," Communications of the ACM, vol. 51, pp. 107-113, 2008.	
3 (9/3)	Frequent itemsets and Association rules	Ch6: Frequent itemsets, Ch3: Finding Similar Items (section 3.5: Distance Measures)	Homework 1 assigned
4 (9/10)	Frequent itemsets and Association rules	Ch6: Frequent itemsets	
5 (9/17)	Shingling, Minhashing, Locality Sensitive Hashing	Ch3: Finding Similar Items	Homework 1 due, Homework 2 assigned
6 (9/24)	Shingling, Minhashing, Locality Sensitive Hashing	Ch3: Finding Similar Items	
7 (10/1)	Midterm (10/2), in-class		
8 (10/8)	Recommendation Systems: Content-based and Collaborative Filtering	Ch9: Recommendation systems, additional readings	

9 (10/15)	Recommendation Systems: Content-based and Collaborative Filtering	Ch9: Recommendation systems	Homework 2 due, Homework 3 assigned
10 (10/22)	Clustering	Ch7: Clustering	
11 (10/29)	Link Analysis: PageRank, Web spam and TrustRank, Random Walks with Restarts	Ch5: Link Analysis	Homework 3 due, Homework 4 assigned
12 (11/5)	Analysis of Massive Graphs (Social Networks)	Ch10: Analysis of Social Networks	
13 (11/12)	Analysis of Massive Graphs (Social Networks)	Ch10: Analysis of Social Networks	Homework 4 due, Homework 5 assigned
14 (11/19)	Web Advertising	Ch8: Advertising on the Web	
15 (11/26)	Mining data streams	Ch4: Mining data streams	Homework 5 due
Final exam	December 11, Tuesday, 2-4pm, same classroom, closed-notes and book		

J. **Statement on Academic Conduct and Support Systems**

Academic Conduct
Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in *SCampus* in Section 11, *Behavior Violating University Standards* https://scampus.usc.edu/1100-behavior-violating-university-standards-and-appropriate-sanctions. 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 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://capsnet.usc.edu/department/department-public-safety/online-forms/contact-us. 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 Center for Women and Men* http://www.usc.edu/student-affairs/cwm/ 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.