USC Viterbi School of Engineering

INF 553: Foundations and Applications of Data Mining

Units: 3
Term—Day—Time:
Fall 2015 – TT – 9:30-10:50am (section 32423D)
Fall 2015 – TT – 5:00-6:20pm (section 32444D)

Location: KAP 163

Instructor: Yao-Yi Chiang
Office: AFH B55C
Office Hours: Tuesday after class
Contact Info: yaoyic@usc.edu, 213-740-7618

Instructor: Wensheng Wu
Office: GER 204
Office Hours: TBD
Contact Info: wuwens@gmail.com

Course Producer: Pooja Anand
Office: TBD
Office Hours: TBD
Contact Info: poojaana@usc.edu

Grader: Siddharth Mahendra Dasani
Office: TBD
Office Hours: TBD
Contact Info: sdasani@usc.edu

Catalogue Course Description

Data mining and machine learning algorithms for analyzing very large data sets. Emphasis on Map Reduce. Case studies.

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.

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.

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

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.

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 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.

Grading Breakdown

Quizzes: There will be weekly quizzes based on the material from the week before. There is no mid-term for this class.
Homework: There will be 5 homeworks based on the topics of the class each week.
Final Exam: There is a final exam at the end of the semester covering all of the material covered in the class.
Class Participation: Students are expected to come to class and participate in the class discussions and discussion board.

Grading Schema:	
Quizzes	30%
Homework	40%
Final:	25%
Class Participation	5%
Total	100%

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

94 - 100 = A 74 - 76 = C

90 – 93 = A- 70 - 73 = C-
87 – 89 = B+ 67 - 69 = D+
84 – 86 = B 64 - 66 = D
80 – 83 = B- 60 - 63 = D-
77 – 79 =C+ Below 60 is an F

Assignment Submission Policy

Homework assignments are due at 11:59pm on the due date and should be submitted in Blackboard. You can submit homework up to one week late, but you will loose 20% of the possible points for the assignment. After one week, the assignment cannot be submitted.

Course Schedule: A Weekly Breakdown

Week	Topic	Readings	Homework	Instructor
1 (8/24)	Introduction to Data Mining, MapReduce	Ch1: Data Mining and Ch2: Large-Scale File Systems and Map-Reduce		Wu
2 (8/31)	MapReduce (cont.)	Ch2: Large-Scale File Systems and Map-Reduce		Wu
3 (9/7)	Frequent itemsets and Association rules	Ch6: Frequent itemsets, Ch3: Finding Similar Items (section 3.5: Distance Measures)	Homework 1 assigned	Chiang
4 (9/14)	Frequent itemsets and Association rules	Ch6: Frequent itemsets		Chiang
5 (9/21)	Shingling, Minhashing, Locality Sensitive Hashing	Ch3: Finding Similar Items	Homework 1 due, Homework 2 assigned	Wu
6 (9/28)	Shingling, Minhashing, Locality Sensitive Hashing	Ch3: Finding Similar Items		Wu
7 (10/5)	Recommendation Systems: Content-based and Collaborative Filtering	Ch9: Recommendation systems, additional readings		Chiang
8 (10/12)	Recommendation Systems: Content-based and Collaborative Filtering	Ch9: Recommendation systems	Homework 2 due, Homework 3 assigned	Chiang
9 (10/19)	Clustering	Ch7: Clustering		Wu
10 (10/26)	Link Analysis: PageRank, Web spam and TrustRank, Random Walks with Restarts	Ch5: Link Analysis	Homework 3 due, Homework 4 assigned	Wu
11 (11/2)	Analysis of Massive Graphs (Social Networks)	Ch10: Analysis of Social Networks		Chiang
12 (11/9)	Analysis of Massive Graphs (Social Networks)	Ch10: Analysis of Social Networks	Homework 4 due, Homework 5 assigned	Chiang
13 (11/16)	Web Advertising	Ch8: Advertising on the Web		Wu
14 (11/23)	Mining data streams	Ch4: Mining data streams	Homework 5 due	Wu
15 (11/30)	Mining data streams Course Summary			Chiang/ Wu
Final (TBD 12/9-12/11)	Final Exam			

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.