USC Viterbi
School of Engineering

DSCI-553: Foundations and Applications of Data Mining

Syllabus

Units: 4

Spring 2022

Tuesday – 7:00-10:20 PM, Room THH-101

Instructor: Professor Wei-Min Shen, PhD
Office Hours: After class (by appointment)

Contact Info: wmshen@usc.edu (include "INF 533" in subject), viterbi-web.usc.edu/~wmshen

Teaching Assistant: TBA
Office: TBA
Office Hours: TBA
Contact Info: tba@usc.edu

Catalogue Course Description

Algorithms and techniques of Data Mining and Machine Learning for analyzing massive datasets. Emphasis on Map Reduce and others. Case studies and applications.

Expanded Course Description

Data mining is a fundamental skill for massive data analysis. 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

DSCI-550, DSCI-551 and DSCI-552 (or equivalents) are required prerequisites. Knowledge of probability, linear algebra, algorithm design, skilled programming, and some machine learning, are essential or highly desirable.

A basic understanding engineering principle is required, including skilled programming with the Python language is desirable. Most assignments are in Spark and require sufficient programming experience and algorithm design. The 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.

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 6 homework assignments and a final project. The assignments and final project 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. Students will be required to finish their homework with pySpark,

and a 10% bonus would be rewarded if they also implement the homework in Scala and their pySpark version is correct.

Grading Breakdown

Quizzes: There will be weekly quizzes based on the material from the week before. There is no mid-term for this class. There will be no makeup for quizzes and students can drop their 2 lowest quizzes .

Homework Assignments: There will be 5-6 homework assignments based on the topics of the class each week. 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.

Data Mining Competition Project: There will be a final project based on the topics introduced in class. The final project is to build an advanced recommendation system and compete with other students and TA for achieving the lowest recommendation errors. There will be some bonus (e.g., extra credits or recommendation letters) for the top few students for the competition.

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

Grading Schema:

Quizzes	30%
Homework	42%
Comprehensive Exam	20%
Data Mining Competition Project	8%
Total	100%

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

94 – 100 = A 90 – 94 = A-

87 – 90 = B+ 83 – 87 = B 80 – 83 = B-

75 – 80 =C+ 70 – 75 = C

67 –70 = C- 65 – 67 = D+ 63 – 65 = D 60 – 63 = D-

Below 60 is an F

Assignment Submission Policy

Homework assignments are due at 3:30pm (LA time) on the due date and should be submitted in Blackboard and/or Vocareum. 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. 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. There is no extension for the final project.

Schedule	**Topic**	**Readings and Assignments**	**Deliverables/Due Dates**
Week 1 1/11	Introduction to Data Mining, MapReduce	Ch1: Data Mining Ch2: Large-Scale File Systems and Map-Reduce	
Week 2 1/18	MapReduce (cont.)	Ch2: Large-Scale File Systems and Map-Reduce	Learn/Install Spark
Week 3 1/25	Frequent itemsets and Association rules	Ch6: Frequent itemsets.	Weekly Quiz starts Homework 1 assigned
Week 4 2/1	Similar Itemset: Shingling, Minhashing, Locality Sensitive Hashing	Ch3: Finding Similar Items	
Week 5 2/8	Recommendation Systems: Content-based and Collaborative Filtering	Ch9: Recommendation systems	Homework 1 due Homework 2 assigned
Week 6 2/15	Recommendation Systems: Content-based and Collaborative Filtering	Additional Readings see lecture notes	
Week 7 2/22	Analysis of Massive Graphs (Social Networks)	Ch10: Analysis of Social Networks	Homework 2 due Homework 3 assigned
Week 8 3/1	Analysis of Massive Graphs (Social Networks)	Ch10: Analysis of Social Networks	
Week 9 3/8	Mining data streams	Ch4: Mining data streams	
Week 10 3/15	Spring Break Week		

Week 11 3/22	Clustering massive data Link Analysis	Ch7: Clustering	Homework 3 due Homework 4 assigned Competition project assigned
Week 12 3/29	Link Analysis I	Ch5: Link Analysis	
Week 13 4/5	Link Analysis II	Ch5: Link Analysis	Homework 4 due Homework 5 assigned
Week 14 4/12	Web Advertising	Ch8: Advertising on the Web	Homework 5 due Homework 6 assigned
Week 15 4/19			Homework 6 due
Week 16 4/26	Comprehensive Exam		
Week 17 5/4			Competition project due

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://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 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/.