USC Viterbi
School of Engineering

# DSCI-553: Foundations and Applications of Data Mining

*Syllabus*

**Units:** 4

**Spring 2024**

**Friday – 2:00-5:20 PM, SGM123/DEN**

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

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

**Teaching Assistants:** See Piazza
**Course Producers:** See Piazza
**Office Hours:** See Piazza
**Contact Info:** See Piazza

**Catalog Course Description**

Algorithms and techniques of Data Mining and Machine Learning for analyzing massive datasets. Emphasis on system building with Spark. 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 them into usable products. 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 build systems and solve real world problems.

**Recommended Preparation**

DSCI-551 and DSCI-552 (or equivalents such as CSCI-585 and CSCI-567) are required and hard prerequisites. Knowledge of probability (e.g., EE364), linear algebra (e.g., EE141), skilled programming (DSCI-510), algorithm design, and machine learning, are essential. This course is one of the most advanced classes in Data Science at USC, so please ensure your background before registration, or else you may be at a severe disadvantage compared to other well-prepared students in the class.

A basic understanding engineering principle is required, including skilled programming with Python. Most assignments are in ***Spark*** and require sufficient programming experience and algorithm design. The assignments are designed for the Unix environment (see vocareum.com); basic Unix skills will make programming assignments much easier.

**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 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 programming languages such as Python. Students are also expected to have their own laptop computer where they can install and run software to do the weekly homework assignments and quizzes.

**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 quiz scores at the end of semester.

**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 on Vocarem.com 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 for achieving the lowest recommendation errors. There will be some bonus (e.g., extra credits or recommendation letters) for the top winner students for the competition.

**Comprehensive Exam:** There will be an exam towards the end of the semester covering all 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. You will have plenty of opportunities to achieve good grades:

92 – 100 = A  88 – 92 = A-

85 – 88 = B+  80 – 85 = B  78 – 80 = B-

75 – 78 =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 on the date/time specified on the homework description upon release, and solutions should be submitted on 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 (but please submit your request to use the free extension days before the deadline, and you can always adjust it after the deadline). 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 due to the limited time in the last week of the semester.

To ensure the individual learning for the assignments, our system and intelligent agents will automatically check the similarity between any pair of submissions. If the matching score for two submissions is above a certain threshold, then both submissions will receive at least 50% for penalty depending on the severity.

| **Schedule** | **Topic** | **Readings and Assignments** | **Deliverables/Due Dates** |
|---|---|---|---|
| **Week 1** | Introduction to Data Mining, MapReduce | Ch1: Data Mining<br>Ch2: Large-Scale File Systems and Map-Reduce | |
| **Week 2** | MapReduce (cont.)<br>Spark (introduction) | Ch2: Large-Scale File Systems and Map-Reduce | Learn/Install Spark<br>Practice quiz online |
| **Week 3** | Frequent itemsets and Association rules | Ch6: Frequent itemsets, | Weekly Quiz starts<br>Homework 1 assigned |
| **Week 4** | Similar Itemset: Shingling, Minhashing, Locality Sensitive Hashing | Ch3: Finding Similar Items | |
| **Week 5** | Recommendation Systems: Content-based and Collaborative Filtering | Ch9: Recommendation systems | Homework 1 due<br>Homework 2 assigned |
| **Week 6** | Recommendation Systems: Content-based and Collaborative Filtering | Additional Readings see lecture notes | |
| **Week 7** | Analysis of Massive Graphs (Social Networks) | Ch10: Analysis of Social Networks | Homework 2 due<br>Homework 3 assigned |

| | | | |
|---|---|---|---|
| **Week 8** | Analysis of Massive Graphs (Social Networks) | Ch10: Analysis of Social Networks | |
| **Week 9** | Mining data streams | Ch4: Mining data streams | |
| **Week 10** | Clustering massive data<br>Link Analysis | Ch7: Clustering | Homework 3 due<br>Homework 4 assigned<br>Competition assigned |
| **Week 11** | Link Analysis I | Ch5: Link Analysis | |
| **Week 12** | Link Analysis II | Ch5: Link Analysis | Homework 4 due<br>Homework 5 assigned |
| **Week 13** | Web Advertising | Ch8: Advertising on the Web | Homework 5 due<br>Homework 6 assigned |
| **Week 14** | | | Homework 6 due |
| **Week 15** | Comprehensive Exam | | |
| **Week 16** | | | 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/.