San José State University
Computer Engineering Department
CMPE 239, Web and Data Mining, Section 1, Spring 2013

Instructor:	Magdalini Eirinaki
Office Location:	ENG 283F
Telephone:	(408) 924 3828
Email:	magdalini.eirinaki@sjsu.edu
Office Hours:	Tuesday 3:30 – 5:30 PM, Thursday 2 – 4 PM (or by appointment)
Class Days/Time:	Tuesday, 6 PM – 8:45 PM
Classroom:	ENG 301
Prerequisites:	CMPE 272 or instructor's consent

Course Web Page and Messaging

Copies of the course materials such as the syllabus, major assignment handouts, etc. may be found on the course shell available from the eLearning platform D2L (Desire2Learn) at: http://sjsu.desire2learn.com.
You are responsible for regularly (i.e. every couple of days) checking with the messaging system through D2L.

For some homework assignments we will use the Gradiance Learning System at: http://www.newgradiance.com/services/servlet/COTC . In order to enroll to the CMPE 239 class please use the token 0AFC3E06.

Course Description

Data mining and Web mining, data preprocessing, association rules and sequential patterns, classification, clustering, Web crawling, information retrieval and search engines, social network analysis, link analysis, ranking, Web usage mining, Web personalization and recommender systems, advanced topics.

Program Outcomes

1. Be able to demonstrate an understanding of advanced knowledge of the practice of computer/software engineering, from vision to analysis, design, validation and deployment.
2. Be able to tackle complex engineering problems and tasks, using contemporary engineering principles, methodologies and tools.
3. Be able to demonstrate leadership and the ability to participate in teamwork in an environment with different disciplines of engineering, science and business.
4. Be aware of ethical, economic and environmental implications of their work, as appropriate.
5. Be able to advance successfully in the engineering profession, and sustain a process of life-long learning in engineer or other professional areas.
6. Be able to communicate effectively, in both oral and written forms.

Course Goals and Student Learning Objectives

The main focus of this course is on Web mining and its applications. More specifically, we will focus on Web usage mining techniques for Web site management, user profiling, and personalization, as well as Web content and structure mining techniques, such as Web information retrieval and link analysis, aiming at supporting search engines. The course will also include an overview of fundamental data mining techniques, focusing on those that are relevant to Web mining. Finally, we will cover current trends such as social network analysis and opinion mining from Web 2.0 sources (e.g., blogs).

This course involves a group-based term project to provide students with the opportunity to build a simplified data or web mining application, and to enhance their professional engineering skills including teamwork, technical leadership, and effective communication skills (both written and verbal).

The course also includes a set of individual assignments and survey projects to enable students to deepen their knowledge on the material.

Course Content Learning Outcomes

Upon successful completion of this course, the students will be able to:

- Discuss in-depth the fundamental data mining concepts and techniques.
- Discuss in-depth the fundamental web mining concepts and techniques.
- Explain how huge amounts of data can be processed using the MapReduce paradigm
- Explain the process of harvesting the information available on the web to build recommender systems and personalized web services.
- Explain how web search engines index, and rank web content.
- Explain how web mining can be applied to extract useful information from Web 2.0 media such as social networking web sites, blogs, reviews' sites, etc.
- Gain hands-on experience by conducting a group-based term project on designing and developing a data/web mining application, or performing an extensive analysis using data/web mining techniques.
- Effectively present and communicate the knowledge they have acquired in the course.

Required Texts/Readings

Textbook

Mining of Massive Datasets, by Anand Rajaraman and Jeffrey Ullman, available here: http://i.stanford.edu/~ullman/mmds.html

Reference books

Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, by Bing Liu
Springer (2007 or 2011 edition), ISBN: 3540378812, available from Spartan Bookstore and online

Data Mining: Concepts and Techniques, by Jiawei Han and Micheline Kamber and Jian Pei
Morgan Kaufmann, Elsevier Inc. (2011), available from Spartan Bookstore and online
(2nd edition is also acceptable)

Other Readings

Various papers and online resources that will be posted on D2L.

Classroom Protocol

You are expected to arrive in time for class. While in class you need to TURN OFF your cellphone. Please be considerate of your fellow students.

Assignments and Grading Policy

Success in this course is based on the expectation that students will spend, for each unit of credit, a minimum of forty-five hours over the length of the course (normally 3 hours per unit per week with 1 of the hours used for lecture) for instruction or preparation/studying or course-related activities including but not limited to internships, labs, clinical practice. Other course structures will have equivalent workload expectations as described in the syllabus.

Student Assessment

Individual homework assignments	5%
Pop Quizzes	10%
Term Project	25%
Midterm Exam	25%
Final Exam (comprehensive)	35%

(A+) >= 98,
(A) >= 94 and <98
(A-) >= 90 and <94
(B+) >=85 and <90
(B) >= 75 and <85
(B-) >= 70 and <75
(C+)>= 68 and <70,
(C) >=64 and <78
(C-) >= 60 and <64,
(D) >=50 and <60,
(F) < 50

- **Students must obtain a passing grade (>50%) in all components of the course in order to get a passing grade (B or better) in this class**
- **No late assignments will be accepted.** An extension will be granted only if a student has serious and compelling reasons that can be proven by an independent authority (e.g. doctor's note if the student has been sick).
- **The exam dates are final.**

Descriptions of Assignments/Exams

Exams: Exams will be a combination of multiple choice and short answer questions and will be based on the individual assignments and course material.

Class Participation: Students will be evaluated based on their participation in in-class written assignments and online discussions. For the in-class assignments, all students are required to write their names on the submitted papers. Failing to do so, even if the student was indeed present in the class, will result in no credit as the instructor is unable to verify the students' claims.

Individual Written/Programming Assignments and Pop Quizzes: Students will be provided with handouts describing the assignments and how they will be graded every week. These assignments will be in-class or take-home written assignments, in-class or take-home lab assignments, and presentation assignments for research papers or articles. Students will also have to answer to pop quizzes that will be based on the homework assignment that is due that day.

Term Project: Groups of 2-3 students will be formed to work on a term-long group project related to data or web mining. The project has deliverables throughout the semester. The quality and completeness of all the deliverables will be considered in grading the projects. All projects will be demonstrated in class.

Each group member is expected to participate in every phase of the project. The final grade of each member will be proportional to his/her participation in the group. Each member should be able to answer questions regarding the project,

present some part of the project demo, and participate in the system implementation and the writing of the technical reports.

Policy on Cheating

A student or students involved in a cheating incident involving any non-exam instrument (homework, report, or lab project) will receive an F on that instrument, and will be reported to the judicial affairs office. Whether the report will carry a recommendation for disciplinary action will be left to my judgment.

A student or students involved in a cheating incident on any quick test, the midterm exam or the final exam will receive an F in the course, and will be reported to the judicial affairs office with a recommendation for disciplinary action.

I will personally notify you of any such findings or actions. All such reports will also be brought to the attention of the Chair of the Computer Engineering department. You have certain rights of appeal, which may serve to exonerate you.
(see http://www.sjsu.edu/student_affairs/academicdishonestyrevisedpolicy.pdf)

Dropping and Adding

Students are responsible for understanding the policies and procedures about add/drop, grade forgiveness, etc. Refer to the current semester's Catalog Policies section at http://info.sjsu.edu/static/catalog/policies.html. Add/drop deadlines can be found on the current academic calendar web page located at http://www.sjsu.edu/academic_programs/calendars/academic_calendar/. The Late Drop Policy is available at http://www.sjsu.edu/aars/policies/latedrops/policy/**.** Students should be aware of the current deadlines and penalties for dropping classes.

Information about the latest changes and news is available at the Advising Hub at http://www.sjsu.edu/advising/.

University Policies

Academic integrity

Your commitment as a student to learning is evidenced by your enrollment at San Jose State University. The University's Academic Integrity policy, located at http://www.sjsu.edu/senate/S07-2.htm, requires you to be honest in all your academic course work. Faculty members are required to report all infractions to the office of Student Conduct and Ethical Development. The Student Conduct and Ethical Development website is available at http://www.sa.sjsu.edu/judicial_affairs/index.html.

Instances of academic dishonesty will not be tolerated. Cheating on exams or plagiarism (presenting the work of another as your own, or the use of another person's ideas without giving proper credit) will result in a failing grade and sanctions by the University. For this class, all assignments are to be completed by the individual student unless otherwise

specified. If you would like to include your assignment or any material you have submitted, or plan to submit for another class, please note that SJSU's Academic Policy S07-2 requires approval of instructors.

Campus Policy in Compliance with the American Disabilities Act

If you need course adaptations or accommodations because of a disability, or if you need to make special arrangements in case the building must be evacuated, please make an appointment with me as soon as possible, or see me during office hours. Presidential Directive 97-03 requires that students with disabilities requesting accommodations must register with the Disability Resource Center (DRC) at http://www.drc.sjsu.edu/ to establish a record of their disability.

Department Policy

- Students who do not provide documentation of having satisfied the class prerequisite or co-requisite requirements (if any) by the second class meeting will be dropped from the class.
- All non-proctored report (or similarly sized) assignments in courses where some of the final grade depends on prose writing will be submitted to Turnitin.com.

CMPE 239 / Web and Data Mining, Spring 2013, Course Schedule

The schedule is tentative and subject to change with fair notice. Any changes will be announced in due time in class and on the course's web site. The students are obliged to consult the most updated and detailed version of the reading material and syllabus, which will be posted on the course's web site.

Table 1 Course Schedule

Week	Date	Topics	Readings
1	1/29	Introduction	Ch.1.1 – 1.2
2	2/5	Large-scale file systems and MapReduce	Ch. 2.1, 2.2, paper on Google FS, paper on MapReduce
3	2/12	Classification: Decision Trees, Naïve Bayes, SVM	Ullman: Ch. 12.1, notes, B.Liu: Ch.3.1, 3.2 (skim), 3.6, 3.8 (skim)
4	2/19	Classification (cont'd): K-Nearest Neighborhood, Evaluation methods	Ullman: Ch. 3.5.1 - 3.5.4, 12.3.1, 12.4.1 - 12.4.4, notes/slides, Liu: 3.3, 3.9, 3.10
5	2/26	Clustering	Ullman: Ch. 7.1.1, 7.1.2, 7.2.1, 7.2.2 (skim), 7.2.3, 7.3.1- 7.3.3, 7.3.4(skim), notes/slides, Liu: Ch. 4.1, 4.2 (4.2.2 skim), 4.3 (skim), 4.4, 4.5.1, 4.8, 4.9
6	3/5	Clustering (cont'd)	Ullman: Ch. 3.5.1 – 3.5.4, 7.1.1, 7.1.2, 7.2.1, 7.2.3, 7.2.4, 7.3.1 – 7.3.3, Liu: Ch. 4.1, 4.2.1, 4.2.3, 4.4, 4.5.1, 4.5.3, slides, notes
7	3/12	MIDTERM	
8	3/19	Recommendation Systems	Ullman: Ch.9.1 - 9.3, slides/notes

Week	Date	Topics	Readings
9	3/26	SPRING BREAK	
10	4/2	Recommendation Systems (cont'd)	Ullman: Ch. 9.4, 9.5 paper of Koren et.al. on Matrix Factorization, Blog article on Netflix Challenge
11	4/9	Association Rules Class Association Rules Sequential Pattern Analysis	Liu: Ch.2.1 - 2.3, 2.4 (skim), 2.5 (skim), 2.6 (skim), notes/slides
12	4/16	Information Retrieval and Web Search Link Analysis - PageRank	Liu: Ch. 6.1, 6.2 (6.2.1-6.2.2), 6.4, 6.5, 6.6 (6.6.1 - 6.6.3), 6.7 (skim), 6.8, 6.9 (skim), 6.10 (skim) Liu: Ch. 7.1 (7.1.2), 7.3 (7.3.1-7.3.2) Lecture slides
13	4/23	Link Analysis - HITS, Spam, Spam Farms	Liu's book: Ch. 7.1 (7.1.2), 7.3 (7.3.1-7.3.2), 7.4 (7.4.1, 7.4.4) Ullman's book: Ch. 5.1, 5.4.1, 5.4.3, 5.4.4, 5.5.1 Lecture slides
14	4/30	Project presentations	
15	5/7	Project Presentations	
16	5/21	FINAL EXAM (comprehensive), 17:15 – 19:30	