San José State University
Computer Engineering Department
CMPE 239, Web and Data Mining, Section 1, Fall 2015

Course and Contact Information

Instructor:	Magdalini Eirinaki
Office Location:	ENG 283F
Telephone:	(408) 924 3828
Email:	magdalini.eirinaki@sjsu.edu
Office Hours:	Tuesday, 4 – 5 PM Thursday, 2 – 3 PM Always check with the CMPE website for the most up to date office hours at http://cmpe.sjsu.edu/content/office-hours.
Class Days/Time:	Tuesday, 6 – 8:45 PM
Classroom:	ENG 337
Prerequisites:	CMPE 272 or instructor's consent (for non-MS SE students)

Course Format

This course requires the student to have a personal computer that is installed with a modern operating system. The lectures will be delivered in the classroom, however the students might be asked to use their laptops or smart devices during the class, or offline in order to participate in the class assignments.

Faculty Web Page and MYSJSU Messaging

Course materials such as syllabus, handouts, notes, assignment instructions, etc. can be found on the Canvas learning management system course website. You are responsible for regularly (at least once a day) checking with the messaging system through MySJSU and Canvas to learn of any updates.

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.

Learning Outcomes

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

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 Learning Outcomes (CLO)

The main focus of this course is on data mining and its applications on the Web. More specifically, we will cover a broad range of web and data mining algorithms and techniques, presented through their use in web-based applications, such as recommender systems, social network mining, web search, opinion mining and sentiment analysis, etc. The lectures will revolve around the fundamental concepts of the areas covered, providing the students the opportunity to gain deep understanding and be able to apply them in real-life scenarios. This will be achieved through multiple in-class and homework assignments that will be required throughout the class. Some data analytics and data mining tools will be used for demonstration purposes, however mastering specific tools is not the primary objective of the lectures. Instead, the students will be able to gain hands-on experience on such technologies via the group-based term project, where they will be expected to build a prototype web mining application, and to enhance their professional engineering skills including teamwork, technical leadership, and effective communication skills (both written and verbal).

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

- Discuss and apply fundamental data mining concepts and techniques such as classification and clustering algorithms, as well as frequent pattern mining.
- Discuss and apply fundamental web mining concepts and techniques using state-of-the-art technologies and platforms.
- Explain how highly unstructured datasets, such as those collected by web applications can be used in order to build recommender systems and personalized web services and be able to identify the proper techniques and algorithms in order to prepare, preprocess, analyze and mine them.
- Explain how classification and text mining techniques can be applied to documents in order to perform sentiment analysis.
- Explain how web search engines index and rank web content using graph mining algorithms.
- 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.
- Quickly get accustomed to any data mining/data analysis software application and
- 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

This class does not have a single textbook. Instead, the students have to study material coming from various books, papers and other resources, all of which are free to download (for academic use). It is each student's responsibility to consult with the updated syllabus on Canvas in order to identify which readings cover the concepts that are taught each week.

A list of reference textbooks is also provided for those who'd like to get some background knowledge or seek more details on the topics covered in class.

Textbooks

[MMDS] *Mining of Massive Datasets,* by Jure Leskovec, Anand Rajaraman and Jeffrey Ullman, 2nd edition, Cambridge University Press, December 2014 (download from http://infolab.stanford.edu/~ullman/mmds/book.pdf)

[ISLR] *An Introduction to Statistical Learning with Applications In R,* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, Springer Texts in Statistics, 2013 (download from http://www-bcf.usc.edu/~gareth/ISL/)

[DMA] *Data Mining and Analysis: Fundamental Concepts and Algorithms,* by Mohammed J. Zaki, Wagner Meira, Jr., Cambridge University Press, May 2014 (download from http://www.dataminingbook.info/pmwiki.php/Main/BookDownload)

[SAOM] *Sentiment Analysis and Opinion Mining,* by Bing Liu, Morgan & Claypool publishers, May 2012 (download the pre-publication version from http://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf)

[OMSA] *Opinion Mining and Sentiment Analysis,* by Bo Pang and Lillian Lee, in Foundations and Trends in Information Retrieval, Vol.2, No. 1-2 (2008 (download the pre-publication version from http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf)

Other Readings

Papers, tutorial slides, articles and all other material that will be made available via Canvas

Lecture slides (available via Canvas)

Reference textbooks (not required)

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

Recommender Systems Handbook, Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors, Springer US, 2011, ISBN: 978-0-387-85819-7

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

Open Intro Statistics, by David M. Diez, Christopher D. Barr, and Mine Cetinkaya-Rundel, 3rd edition, 2015 (https://www.openintro.org/stat/textbook.php)

Other equipment / material requirements

Programming languages, platforms, as well as software applications and tools, such as Spark, Mahout, R/RStudio, Python, WEKA, Tableau, etc. that will be required for this class are either free to download, or the instructor will provide the students with academic licenses. Students will be informed in class and via Canvas ahead of time in order to install all required software.

Course Requirements and Assignments

SJSU classes are designed such that in order to be successful, it is expected that students will spend a minimum of forty-five hours for each unit of credit (normally three hours per unit per week), including preparing for class,

participating in course activities, completing assignments, and so on. More details about student workload can be found in University Policy S12-3 at http://www.sjsu.edu/senate/docs/S12-3.pdf.

Student Assessment

In-class & online activities	5%
Short story assignment	5%
Individual homework assignments	5%
Pop Quizzes	10%
Term Project	25%
Midterm Exam	20%
Final Exam (comprehensive)	30%

Descriptions of Assignments/Exams

In-class & social network assignments: Students will be evaluated based on their participation in in-class and social network assignments. All students are required to write their names on the submitted work and/or submit their answers online using their unique IDs, shared with the instructors. Failing to do so, even if the student was indeed present in the class, will result in zero credit as the instructor is unable to verify the student's claims. Moreover, students whose name appears on submitted work, but were not in class, as well as the students who submitted their name on their behalf are violating the academic integrity policy and will be reported immediately to the office of Student Conduct and Ethical Development.

Short story assignment: Each student will be assigned a week during which they have to find, and present in class, a news story (e.g. about a new algorithm/technology etc.) that is related to the topics covered in class. This story has to have some technical background and the student will be asked to elaborate and analyze the elements pertaining to the class as well as carry a discussion with the rest of the group based on it. The peers will be able to provide feedback, taken into account for the grade.

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. The worst pop quiz grade will not be counted towards the final pop quiz grade of each student ("worst-one out policy").

Term Project: Groups of 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. The project details will be announced by the instructor and posted on the course's web site well before the deadlines.

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, as assessed by the instructor and the student's peers. 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. The term project will be graded on the basis of the following three components: a) project implementation, b) project report, c) project demonstration.

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

NOTE that University policy F69-24 at http://www.sjsu.edu/senate/docs/F69-24.pdf states that "Students should attend all meetings of their classes, not only because they are responsible for material discussed therein, but because active participation is frequently essential to insure maximum benefit for all members of the class. Attendance per se shall not be used as a criterion for grading."

Grading Policy

The final grades will be calculated based on the following:

(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 60% or more 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.

All students have the right, within a reasonable time, to know their academic scores, to review their grade-dependent work, and to be provided with explanations for the determination of their course grades.

Classroom Protocol

You are expected to arrive in time for class. While in class you need to turn off your cellphone unless directed otherwise by your instructor. Laptop/tablet/smart phone use is allowed only for activities related to the class. Please be considerate of your fellow students.

University Policies

General Expectations, Rights and Responsibilities of the Student

As members of the academic community, students accept both the rights and responsibilities incumbent upon all members of the institution. Students are encouraged to familiarize themselves with SJSU's policies and practices pertaining to the procedures to follow if and when questions or concerns about a class arises. See University Policy S90–5 at http://www.sjsu.edu/senate/docs/S90-5.pdf. More detailed information on a variety of related topics is available in the SJSU catalog, at http://info.sjsu.edu/web-dbgen/narr/catalog/rec-12234.12506.html. In general, it is recommended that students begin by seeking clarification or discussing

concerns with their instructor. If such conversation is not possible, or if it does not serve to address the issue, it is recommended that the student contact the Department Chair as a next step.

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 year calendars document on the Academic Calendars webpage at http://www.sjsu.edu/provost/services/academic_calendars/. 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/.

Consent for Recording of Class and Public Sharing of Instructor Material

University Policy S12-7, http://www.sjsu.edu/senate/docs/S12-7.pdf, requires students to obtain instructor's permission to record the course and the following items to be included in the syllabus:

- "Common courtesy and professional behavior dictate that you notify someone when you are recording him/her. You must obtain the instructor's permission to make audio or video recordings in this class. Such permission allows the recordings to be used for your private, study purposes only. The recordings are the intellectual property of the instructor; you have not been given any rights to reproduce or distribute the material."
    - It is suggested that the greensheet include the instructor's process for granting permission, whether in writing or orally and whether for the whole semester or on a class by class basis.
    - In classes where active participation of students or guests may be on the recording, permission of those students or guests should be obtained as well.
- "Course material developed by the instructor is the intellectual property of the instructor and cannot be shared publicly without his/her approval. You may not publicly share or upload instructor generated material for this course such as exam questions, lecture notes, or homework solutions without instructor consent."

Academic integrity

Your commitment, as a student, to learning is evidenced by your enrollment at San Jose State University. The University Academic Integrity Policy S07-2 at http://www.sjsu.edu/senate/docs/S07-2.pdf 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.sjsu.edu/studentconduct/.

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 at http://www.sjsu.edu/president/docs/directives/PD_1997-03.pdf requires that students with disabilities requesting accommodations must register with the Accessible Education Center (AEC) at http://www.sjsu.edu/aec to establish a record of their disability.

Accommodation to Students' Religious Holidays

San José State University shall provide accommodation on any graded class work or activities for students wishing to observe religious holidays when such observances require students to be absent from class. It is the responsibility of the student to inform the instructor, in writing, about such holidays before the add deadline at the start of each semester. If such holidays occur before the add deadline, the student must notify the instructor, in writing, at least three days before the date that he/she will be absent. It is the responsibility of the instructor to make every reasonable effort to honor the student request without penalty, and of the student to make up the work missed. See University Policy S14-7 at http://www.sjsu.edu/senate/docs/S14-7.pdf.

CMPE 239-01 / Web and Data Mining, Fall 2015, Course Schedule

The schedule (and related dates/readings/assignments) is tentative and subject to change with fair notice. In case of guest lectures the syllabus will be updated accordingly. Any changes will be announced in due time in class and on the course's web site (Canvas). The students are obliged to consult the most updated and detailed version of the reading material and syllabus, which will be posted on Canvas.

Course Schedule

Module (Weeks)	Date	Topics	Readings	Assignments
1 (1)	8/25	Introduction	[MMDS] Ch.1 [DMA] Ch. 1.1-1.3	
2 (2-3)	9/1 9/8	Prediction & Classification (part I): Regression, Decision Trees, K-NN	[ISLR] Ch. 2.1.3-2.1.5, 3.1, 3.5, 4.1, 8 [DMA] Ch. 19 [MMDS] Ch. 3.5.1-3.5.4	HW 1
3 (3-6)	9/8 9/15 9/22 9/29	Recommender Systems: Content-based Recommendations, Collaborative Filtering, Latent Factor Models, Evaluation Methods	[MMDS] Ch. 3.3, 3.5.1 – 3.5.4, Ch.9.1 - 9.3 [ISLR] Ch. 5 [DMA] Ch. 22 [Lops et al.] paper on Content-based Recommender Systems [Koren et al.] paper on Matrix Factorization [Linden et al.], [Sarwar et al.] and [Derosiers and Karypis] papers on Item-to-Item Collaborative filtering [Amatrian] paper on Netflix Recommendations	HW 2
4 (7)	10/6	Recommender Systems & Big Data: Optimization, Dimensionality Reduction, MapReduce	[MMDS] Ch. 2.1, 2.2, 9.4 [Dean and Ghemawat] paper on MapReduce	HW 3
	10/13	MIDTERM		
5 (9, 10)	10/20 10/27	Mining Relationships (part I): Clustering Guest lecture (tentative)	[MMDS] Ch. 3.5.1 – 3.5.4, 7.1 – 7.4 [DMA] Ch. 15 [ISLR] Ch. 10.3	HW 4
6 (11, 12)	11/3 11/10	Prediction & Classification (part II): Naïve Bayes, Opinion Mining & Sentiment Analysis	[OMSA] Ch. 1 – 4 [SAOM] Ch. 1 – 3	HW 5
7 (13)	11/17	Mining Relationships (part II): Frequent Itemsets, Association Rules	[MMDS] Ch. 6.1, 6.2 [DMA] Ch.8, Ch.10.1, 10.2	

Module (Weeks)	Date	Topics	Readings	Assignments
8 (14)	11/24	Graph Mining: Web search and ranking, Social network graph mining	[MMDS] Ch. 5.1, 5.4, 5.5, 10.1, 10.2 [Gupta et al.] paper on Twitter graph mining	HW 6
9 (15, 16)	12/1 12/8	Project Presentations		
	12/15	FINAL EXAM (comprehensive) 17:15 – 19:30		