San José State University
Computer Engineering Department
CMPE 256, Large-Scale Analytics, Section 1, Spring 2019

Course and Contact Information

Instructor:	Magdalini Eirinaki
Office Location:	ENG 283F
Telephone:	(408) 924 3828
Email:	magdalini.eirinaki@sjsu.edu
Office Hours:	Thursday, 1 – 3 pm Always check with the CMPE website and the class Canvas site for the most up to date office hours at http://cmpe.sjsu.edu/content/office-hours.
Class Days/Time:	Tuesday & Thursday, 10:30 - 11:45 am
Classroom:	ENG 301
Prerequisites:	Classified graduate standing (all students) and 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 Canvas Leaning Management System course login website at http://sjsu.instructure.com. You are responsible for regularly checking with the messaging system through MySJSU at http://my.sjsu.edu (or other communication system as indicated by the instructor) to learn of any updates.

Course Description

Data mining and machine learning algorithms and applications for large amounts of data. Information retrieval and search engines, social network analysis, link analysis, ranking, Web personalization, recommender systems, opinion mining and sentiment analysis, advanced topics.

Course Goals

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 machine learning algorithms and applications for large amounts of data. We will cover a broad range of data analytics, data and Web mining and machine learning algorithms and techniques, presented through their use in applications such as recommender systems, social network analysis, web search, opinion mining and sentiment analysis, etc. The lectures will revolve around the fundamental concepts of the areas covered as well as case studies, 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 large-scale data mining and machine learning 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 web mining can be applied to extract useful information from Web 2.0 media such as social networking web sites, blogs, reviews' sites, etc.
- Explain how classification and text mining techniques can be applied to documents in order to perform sentiment analysis.
- Explain how graph mining algorithms can be applied to rank search, identify communities, identify influencers in a social network etc.
- Quickly get accustomed to any data mining/data analysis software application and be able to use it.
- 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 would like to get some background knowledge or seek more details on the topics covered in class.

Textbooks

Recommender Systems: The textbook, by Charu Aggrawal
Springer, 2016, ISBN 978-3-319-29659-3
(available in pdf from the SJSU Library at: http://library.sjsu.edu)

Social Media Mining, An Introduction, by Reza Zafarni, Mohammad Ali Abasi, Huan Liu
Cambridge University Press, 2014 (available to download at: http://dmml.asu.edu/smm/SMM.pdf)

Networks, Crowds, and Markets, by D. Easley and J. Kleinberg
Cambridge University Press, 2010 (available to download at:
https://www.cs.cornell.edu/home/kleinber/networks-book/networks-book.pdf)

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)

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)

Other Readings

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

Lecture slides (available via Canvas)

Reference textbooks (not required)

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)

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), ISBN: 9780123814791
(available as ebook from the SJSU Library)

Other technology requirements / equipment / material

Programming languages, platforms, as well as software applications and tools, such as Jupyter notebook, Spark (MLLib), Mahout, Python (sklearn, networkx), Gephi, GraphX, Google CloudML etc. that will be required for this class are either free to download, or the instructor will provide the students with academic licenses/credits. 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 activities	5%
Homework assignments	5%
Quizzes	10%
Programming assignments/Competitions	10%
Term Project	25%
Midterm Exam	15%
Final Exam (comprehensive)	30%

Descriptions of Assignments/Exams

In-class activities: Students will be evaluated based on their *participation* in in-class activities. 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 names appear 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. As the name implies, credit will be given only to those present when the activity took place in the classroom. No make-ups or remote participation is allowed.

Individual Written/Programming (Homework) Assignments and Quizzes: Students will be provided with details 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/or presentation assignments for research papers or articles. Students will also have to answer to quizzes that will be based on the homework assignment that is due that day. The worst quiz grade will not be counted towards the final pop quiz grade of each student ("worst-one out policy").

Programming Assignments/Competitions: Students will participate in one or more competition-like programming assignments related to the contents of the class. The students will have to implement a solution to a given problem and will be evaluated against a baseline with a given related metric. An accompanying report will be submitted. Credit will be given to those who successfully beat the baseline and present their solution with a report. Extra credit might be given to top-performing submissions.

Term Project: Groups of 3 students will be formed to work on a term-long group project related to the topics taught in class. 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. Grading will be rubric-based.

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. The final exam is comprehensive and the date is determined by the University's Final Examination Schedule.

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 <68
(C-) >= 60 and <64,
(D) >=50 and <60,
(F) < 50

- 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).
- No make-up or extra credit assignments will be given to those who miss deadlines or fail in one of the components in the class.
- 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

Per University Policy S16-9, university-wide policy information relevant to all courses, such as academic integrity, accommodations, etc. will be available on Office of Graduate and Undergraduate Programs' Syllabus Information web page at http://www.sjsu.edu/gup/syllabusinfo/"

Department Policies

- 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.
- Major exams in this class may be video recorded to ensure academic integrity. The recordings will only be viewed if there is an issue to be addressed. Under no circumstances will the recordings be publicly released.

CMPE 256-01 / Large-Scale Analytics, Spring 2019, 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

Week	Date	Topics, Readings, Assignments, Deadlines
1	1/24	Introduction to CMPE 256
2	1/29	Introduction to machine learning and large-scale analytics
	1/31	Building blocks: Data preparation
3 - 4	2/5	Scaling in/out/up Dimensionality reduction, MapReduce Brief introduction to Mahout and Spark frameworks
	2/7	
	2/12	
4 - 8	2/14	Recommendation systems – Data engineering, Data preprocessing, Content-based Collaborative Filtering, User- and Item-based Collaborative Filtering, Evaluation
	2/19	
	2/21	
	2/26	
	2/28	
	3/5	
	3/7	
	3/12	
8	3/14	MIDTERM
9 - 10	3/19	Recommendation systems - Latent factor Collaborative Filtering/Matrix Factorization Case study: Netflix recommendations
	3/21	
	3/26	
10 - 12	3/28	Graph mining – Graph theory basics, Social Network Analysis & Mining, Link analysis, ranking algorithms, Identification of influencers, community detection Case study: Facebook romantic relationships SPRING BREAK: 4/1 – 4/5
	4/9	
	4/11	
	4/16	
	4/18	
13	4/23	Fairness and bias in Machine Learning
13-14	4/25	Bringing all together / Applications: trust and influence propagation in social

Week	Date	Topics, Readings, Assignments, Deadlines
	4/30	networks, social recommender systems, opinion mining, etc. (as time allows)
14-15	5/2	Project Presentations
	5/7	
	5/9	
Finals Week	Thursday, May 16	FINAL EXAM 9:45 am – 12 (noon)