San José State University
School/Department
42669, Topics in Database System, CS 267-01, Fall, 2016

Course and Contact Information

Instructor:	Thanh Tran
Office Location:	MacQuarrie Hall 216
Telephone:	924-7227
Email:	ducthanh.tran@sjsu.edu
Office Hours:	Monday 1:30 – 3:30, please drop me email with time info and subject
Class Days/Time:	MoWe 12:00PM - 1:15PM
Classroom:	MacQuarrie Hall 222

Prerequisites:

Students should know about

- Programming (CS046A or equivalent),
- Programming languages (at least one of the following: C, C++, Java or Perl),
- Data structures and algorithms (at least CS046B or equivalent, CS146 preferred).
- Database management systems (CS157B or equivalent)

Also, some familiarity with

- Team projects,
- Distributed and parallel computing,
- Data management technologies,
- Cloud technologies,
- Artificial Intelligence and
- Big Data is beneficial.

Course Format

Technology Intensive, Hybrid, and Online Courses

This course combines theories with hands-on assignments in which students are expected to work in teams to complete real-world data analytics project(s) using open source technologies (Hadoop, HDFS, Spark, Weka, GraphLab) and/or commercial solutions provided by IBM (BigInsights, H20).

The main course components are online videos, project work, discussions of project assignments in class and guest lectures from professionals in the field.

Faculty Web Page and MYSJSU Messaging

Course materials such as syllabus, handouts, notes, assignment instructions, etc. can be found on my faculty web page at http://sites.google.com/site/kimducthanh and/or on Canvas Leaning Management System course login website at http://sjsu.instructure.com.

Course Description

General: Advanced topics in the area of database and information systems. Content differs in each offering. Possible topics include though not restricted to: Data Mining, Distributed Databases and Transaction Processing. Prerequisite: CS 157B.

Course-specific:
Making sense of and exploiting the mass amount of available data for decision-making is a critical task for organizations and companies in many industries. This course covers the main theories as well as widely used technologies for data mining, focusing on scalable solutions that are applicable to Big Data. You will go through the entire Big Data mining process, from data extraction, loading and processing to feature selection and choosing, testing as well as tuning machine learning models to accomplish various mining tasks.

Course Learning Outcomes (CLO)

This course aims to introduce students to the main theories, methods and solutions for mining Big Data, focusing on the aspects important to the application and tuning of data mining technologies.

The objective is to teach students about the main theories behind Big Data mining as well as technologies widely used in this field:

- Purposes and applications of Big Data mining
- Machine Learning concepts for Big Data mining
- Techniques and technologies for preprocessing the data and extracting features
- Techniques and technologies for performing 5 mining tasks, i.e. (0) finding similar items, and (1) classification, (2) regression, (3) clustering and (4) frequent pattern mining

Students will learn how to use and tune machine learning models for performing mining tasks over large amount of data / datasets.

Required Texts/Readings

Textbook

No textbook is required for the course.

Other Readings

Students may use the following textbooks to augment and reinforce their learning:

- Data-Intensive Text Processing with MapReduce, by Jimmy Lin and Chris Dyer
- Mining of Massive Data Sets, by Anand Rajaraman, Jure Leskovec and Jeff Ullman

Course Requirements and Assignments

The main topical requirements for the course are:

- Big Data mining applications, challenges and technologies
- Overview of the Big Data mining process
  - Extract-transform-load from external systems
  - Storage using Big Data file / database systems
  - Preprocessing: cleaning, dimensionality reduction, sampling, feature extraction
  - Task-specific model selection and tuning
  - Evaluation
- Big Data mining
  - Scalable approaches for Big Data (linear models, online learning)
  - Parallel learning with Graph-analytics systems (Pregel, GraphLab) and Dataflow-based systems (Hadoop, Haloop, Spark)
- Big Data Mining Task 1: Regression
  - Linear, logistic, Bayesian and multinomial regression
- Big Data Mining Task 2: Classification
  - Naive Bayes, kNN, Support Vector Machines, Distributed Decision Trees, Distributed Random Forest, Gradient Boosting Machines
- Big Data Mining Task 3: Clustering
  - k-means, canopy, HAC, EM-Clustering, PLSI, LDA
- (Big Data Mining Task 4: Frequent Pattern Mining)
  - Apriori, FP-tree

The course focuses on two aspects.

- **Own research / project**: In their own pursuits and experiments, students will use references provided by the instructor as well as materials they find through their own research to study new Big Data mining technologies. They present the results in class in the form of a technology overview presentation and / or tutorial, or as a written report in their technology blog. The main focus is set on the semester-long project, where students choose and experiment with several Big Data mining technologies to accomplish their project goals.
- **In-class discussion, presentation and learning**: the fundamentals are discussed in class, which comprise the main theories behind Big Data mining. This theoretical part is complemented with practical knowledge provided through technology presentations, tutorials and project reports / discussions. While 100% class attendance is not required, active participation in the technology / project discussions will be reflected in the final grade.

Project

This course is application-oriented, putting a strong emphasize on a semester-long project. The project shall be conducted collaboratively (in teams with 3 - or individually in exceptional cases. The instructor will provide project ideas and several recommended projects with different characteristics that students can choose from. Students can also propose and pursue their own project ideas.

Give a large amount of data and a specific data mining problem, the **task** is to choose, evaluate and tune machine learning models. Since current Big Data mining solutions focus on a specific set of tools / models, several technologies might have to be incorporated. While student teams are encouraged to choose an ambitious scenario (with an end user application such as "the next Facebook" in mind), they are advised to devise a contingency plan. That is, every team might have an ambitious plan A (that they might want to pursue even beyond the scope of this course) and a plan B that precisely focuses on the core course requirements.

There are two main types of **requirements**:

- The first is task coverage: in their projects, students shall go through all the main tasks involved in the Big Data mining pipeline, from data loading to feature selection to model learning, evaluation and tuning.
- The second is model coverage: for the given data mining problem, there will be always several models that are applicable. Every team member should focus on one particular model so that in the end, the project showcases the usage and comparison of at least three different models. Optionally, the team might consider advanced techniques for combining the results obtained through different models to improve the overall learning quality.

The overall project quality is measured with respect to these requirements. That is, students should prepare for these questions:

- Does the project involve all major data mining tasks?
- Does the project incorporate the use of at least three different models?
- What is the quality of the task execution / outcomes?
- What is the quality of the chosen / tuned models?

Note that in order to satisfy the course requirements, there is no need to build a fancy UI. In particular, there might be not enough time to develop a sophisticated application.

Pointers to datasets to be considered for the project are:

- UCI Machine Learning Repository: Data Sets
- Webscope from Yahoo! Labs
- Datasets for Data Mining and Data Science
- mldata

Some ideas for more general Big Data projects:

- 5 Big Data Projects That Could Impact Your Life
- 3 big ideas for big data in the public sector
- Using OpenData to help drivers navigate NYC

Specific project recommendations:

- P1: Predict which users (or information sources) one user might follow in Tencent Weibo
- P2: Predict the click-through rate of ads given the query and user information
- P3: Determine whether an author has written a given paper
- P4: Information Retrieval – Learning to Rank
- P5: Entity Resolution – Interlinking New-York Times Data
- P6: Open Task – Google: Learning from Big Data: 40 Million Entities in Context
- P7: Open Task – Yelp: Yelp Dataset Challenge

Projects done before:

- http://sjsubigdata.wordpress.com/big-data-mining/

P1, P2, P3 and P4 are well known challenges for which many solutions have already been proposed. That is, it is known what features can be used and which machine learning models shall be applied. However, some of these models are not implemented by existing Big Data mining tools . So, the task here is how to reproduce and improve upon existing solutions using available Big Data mining tools.

P5, P6 and P7 are less clear, providing more challenges and room to explore a good solution. Extra credits are available here for creative and effective solutions.

The progress of the projects will be documented in a technology blog and discussed in class on a regular basis. In particular, the blog will contain a

- **Project proposal** (1-2 pages)
    - Title, team, project idea, data mining problem, datasets to be used, main tasks to be addressed, machine learning models and technologies that might be relevant, project management (milestones, work allocation: who does what?)
- **Progress report 1 - Initial Attempt** (2-3 pages): using (a subset of) the data, obtain first initial results based on Big Data mining tool and model of your own choosing, evaluate the results, identify problems and discuss directions for subsequent work
    - Data, data mining problem, model selection, evaluation result, discussion: problems and future work
- **Progress report 2 - Data Processing and Feature Selection** (2-3 pages): first extract, integrate, clean and enrich data, then select features using appropriate technologies and tools (Pig, Hive, Talend)
    - Data Processing: initial dataset, data quality problems, data processing tasks, resulting dataset
    - Feature extraction: dataset, rationales behind feature selection, feature selection tasks, selected features
- **Progress report 3 - Model Selection and Tuning** (4-5 pages): experiment with at least 3 different models and perform model tuning
    - Data, data mining problem, model selection, model tuning, evaluation result, discussion: problems and future work
- **Final report** (8-10 pages): title, team, idea, data mining problem, datasets, project management (milestone, work allocation), project solution (feature selection, model selection, model tuning), results, conclusions, final assessment

There will be several slots for teams to present and discuss their progress in class. Also, there will be a final **project presentation**.

The breakdown of the project score (60) is as follow:

- Reports (30 points) PP (5 points) + 3 PR (15 points) + FR (10 points)
- Presentation (10 points): project reports are chosen for presentation; every team presents at least twice (5 points for each presentation)
- Project management, team communication, communication with project manager / instructor, discussion of results in class (5 points)
- Overall project quality (15 points)
    - Data cleaning, feature extraction, model selection, model tuning, model evaluation

For more precise assessment, student's individual contributions to the project shall be documented in the report and correspondingly, reflected in the project presentation. Also, every student will be asked to provide an assessment of their own and other team members' contributions (check-point / final assessment).

Technology Presentation / Blog

Every student can choose to prepare either a presentation or a written report (technology blog) of the following types:

- **Tool overview**: the goal of this is to provide an overview of the (1) machine learning tool, (2) its special support for Big Data / deal with scalability issues, if any, and (3) comparison with related tools, if applicable, and a (4) tutorial on how to use it for different tasks. The tutorial part contains detailed instructions so that fellow students can follow and learn how to use it for their projects. Fellow students should be able to successfully complete the tutorial and acquire hands-on experiences from this.

- **Model / task specific**: here, we focus on a particular task and present (1) the tool and (2) specific models and/or algorithms and/or techniques that can be used to solve this task. This should also be done as a tutorial, providing detailed instructions so that fellow students can follow and learn how to use it for their projects.
- **Critical analysis and comparison of tools**: this type requires information that goes beyond what are explicitly available. Students are expected to identify the criteria used for the comparison, critically analyze the technologies, derive conclusions and if necessary, run experiments to verify them.
- **Own tool implementation**: for this, students implement an existing ML algorithm using a Big Data processing platform of choice.

In order to make the class more discussion-oriented, the presenter is encouraged to ask questions and to spur discussion. For this, try to avoid tough questions. As a rule of thumb, prepare questions that are more open to facilitate interesting exchange and class discussion.

The breakdown of the score (20) is as follow:

- Management / pacing (5): all reports / presentations require instructor's guidance and final approval; the highest score can be achieved when the result is delivered as requested by the instructor and requires the least amount of guidance
- Quality of the report / presentation (15)

Presentations / blogs done before: the presentation / blog chosen should be completely (a) different from or (b) based on works done before. When choosing option (b), the presentation should have clear improvements (structure, content, presentation) over the previous one + have at least 20% additional materials.

- http://sjsubigdata.wordpress.com/big-data-mining/

Final Examination or Evaluation

The exam will test the student's knowledge of the key concepts in the course. For preparation, the instructor will release candidate questions 2 weeks before the scheduled exam.

Grading Information

Determination of Grades

The tentative components of the final grade are:

- Project 60%
- Optional: technology presentation / blogging 20%
- Final Exam: 40% (max) OR 20% plus presentation / blogging score

A more fine-grained breakdown of these components are given in the project / technology presentation sections.

Late progress will result in lower score (see scores for pacing etc.).

Each assignment, project, and exam will be scored (given points) but not assigned a letter grade.
Final individual class letter grades will be assigned based on the class curve.
Your final class grade can be adjusted up or down depending on your level and quality of class / project participation.

Classroom Protocol

Attendance is not required. Late attendance or any behavior (such as usage of cell phone) that leads to distraction should be avoided.

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/

42669 / Topics in Database System, Fall 2016, Course Schedule

This schedule is subject to change with fair notice. Please check your Canvas (https://sjsu.instructure.com) for announcements.

Course Schedule

Week	Date	Topics, Readings, Assignments, Deadlines
1	08/24/2016	Introduction to Big Data Mining Readings: https://drive.google.com/open?id=0BzbxhrIWrCCDeF9JbFl6ZGNQN3M Assignments: form team
1	08/29/2016	Big Data Mining projects Readings: https://drive.google.com/open?id=0BzbxhrIWrCCDeF9JbFl6ZGNQN3M Assignments: discuss project ideas
2	08/31/2016	Project Proposal Discussions
2	09/07/2016	Project Proposal Presentations
3	09/12/2016	Big Data Mining tools 1: Weka, Spark, H20, R Readings: Overview of tools http://www.kdnuggets.com/2015/06/data-mining-data-science-tools-associations.html, http://www.datamation.com/data-center/slideshows/8-open-source-big-data-mining-tools.html, http://thenewstack.io/six-of-the-best-open-source-data-mining-tools/ Tool-specific documentation RapidMiner: http://docs.rapidminer.com/studio/getting-started/

Week	Date	Topics, Readings, Assignments, Deadlines
		Weka: http://www.cs.waikato.ac.nz/ml/weka/documentation.html R: http://www.rdatamining.com/docs, http://www.tutorialspoint.com/r/ Spark: https://www.infoq.com/articles/apache-spark-introduction, http://www.kdnuggets.com/2015/11/petrov-apache-spark-machine-learning-large-data.html, http://spark.apache.org/docs/latest/index.html, http://spark.apache.org/docs/latest/mllib-guide.html For advanced stuff: http://www.kdd.org/kdd2015/tutorial.html Assignments: build your first BDM model, work on Project Report 1
3	09/14/2016	Big Data Mining tools 2: Weka, Spark, H20, R
4	09/19/2016	Big Data Mining tools 3: Weka, Spark, H20, R
4	09/21/2016	Project Report 1 Presentations
5	09/26/2016	Project Report 1 Presentations
5	09/28/2016	Project Report 1 Presentations
6	10/03/2016	Big Data Mining process – a quick run-through Readings: https://drive.google.com/open?id=0BzbxhrIWrCCDYVRORHM1LXNyTVU, Data exploration: https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/, http://ampcamp.berkeley.edu/big-data-mini-course/data-exploration-using-spark.html, http://blog.cloudera.com/blog/2016/06/how-to-analyze-fantasy-sports-with-apache-spark-and-sql-part-2-data-exploration/ Model Evaluation: https://spark.apache.org/docs/latest/mllib-evaluation-metrics.html Assignment: start working on Project Report 2
6	10/05/2016	Big Data Mining process – a quick run-through
7	10/10/2016	Parallel computing for BDM Readings: https://drive.google.com/open?id=0BzbxhrIWrCCDTnc4OFgzRW9UY0E,
7	10/12/2016	Parallel computing for BDM
8	10/17/2016	Big Data Processing Tools: Hadoop, Pig, Hive Readings:

Week	Date	Topics, Readings, Assignments, Deadlines
		http://hortonworks.com/hadoop-tutorial/hello-world-an-introduction-to-hadoop-hcatalog-hive-and-pig/ Hive and Pig: https://www.dezyre.com/article/difference-between-pig-and-hive-the-two-key-components-of-hadoop-ecosystem/79 ETL: https://blogs.aws.amazon.com/bigdata/post/Tx2D93GZRHU3TES/Using-Spark-SQL-for-ETL
8	10/19/2016	Project Report 2 Presentations
9	10/24/2016	Unsupervised Learning Readings: https://drive.google.com/open?id=0BzbxhrIWrCCDb0YzZ2YxN3JFcjg Assignment: start working on Project Report 3
9	10/26/2016	Supervised Learning – Linear Models Readings: https://drive.google.com/open?id=0BzbxhrIWrCCDOXNIYklHa1FrbEE
10	10/31/2016	Linear Model Extensions for Dealing with Non-Linearity Readings: https://drive.google.com/open?id=0BzbxhrIWrCCDOXNIYklHa1FrbEE
10	11/02/2016	Non-linear Models: Decision Tree Readings: https://drive.google.com/open?id=0BzbxhrIWrCCDaHJJQWlIZ2dUams
11	11/072016	Non-linear Models: Instance-based Readings: https://drive.google.com/open?id=0BzbxhrIWrCCDcVJ6S0tua2hBdmM
11	11/09/2016	Non-linear Models: Statistical Learning Readings: https://drive.google.com/open?id=0BzbxhrIWrCCDZHVUS1RQeTB5X00
12	11/14/2016	Non-linear Models: Ensemble Methods

Week	Date	Topics, Readings, Assignments, Deadlines
		Readings: https://drive.google.com/open?id=0BzbxhrIWrCCDMTR2cFFCLUdZZFE
12	11/16/2016	Non-linear Models: Random Forest Readings: https://drive.google.com/open?id=0BzbxhrIWrCCDMTR2cFFCLUdZZFE
13	11/21/2016	Multi-class Learning Readings: https://drive.google.com/open?id=0BzbxhrIWrCCDWHVEaFFVeS15RHc
13	11/23/2016	Learning to Rank Readings: https://drive.google.com/open?id=0BzbxhrIWrCCDdDhmRWVlQTc0N1E
14	11/28/2016	Project Report 3 Discussion
14	11/30/2016	Project Report 3 Presentations
15	12/05/2016	Project Report 3 Presentations
15	12/07/2016	Project Report 3 Presentations
16	12/12/2016	Exam Preparation & QA
Final Exam	16/12/2016	9:45-12:00