CSC 560: Special Topics in Databases
Web Mining
Fall 2009
Course Syllabus

September 20, 2009

Instructor: Alexander Dekhtyar
email: dekhtyar@calpoly.edu
office: 14-215

What	When		Where
Lecture	MW	4:10 – 6:00pm	14-232B
Final Exam time	December 7, 2009 (Monday)	4:10 - 7:00pm	14-232B

Note: the class will not have a written final exam, but the exam time will be used for student presentations.

Office Hours

	When	Where
Monday	8:30am - 9:30am	14-215
Wednsday	8:30am - 9:30am	14-215
Thursday	9:00am - 12:00pm	14-215

Additional appointments can be scheduled by emailing the instructor at *dekhtyar@calpoly.edu*.

Description

This course has two main objectives:

- Study a number of advanced data mining methods;

- Study data mining techniques and methods designed specifically for analysis of data on the World Wide Web.

Texbook

Most of the course material will come from:

- Bing Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, Springer, 1st ed. 2007. ISBN: 978-3540378815.

In addition, we will be reading a number of papers.

Topics

Week	DOW	Date	Lecture	Assignment
Week 1	Wednesday	23-Sep	Syllabus, Intro to Data Mining, Web Mining	Stage 1: team formation
Week 2	Monday	28-Sep	Recap: Classification	Stage 1: due/ Stage 2: proposal
	Wednesday	30-Sep	Recap: Clustering	
Week 3	Monday	5-Oct	Web Mining: Intro	
	Wednesday	7-Oct	Web Structure Mining	
Week 4	Monday	12-Oct	Web Content Mining	
	Wednesday	14-Oct	FURLOUGH	
Week 5	Monday	19-Oct	Web Usage Mining	Stage 2: due
	Wednesday	21-Oct	Proposals	Stage 3: project implementation
Week 6	Monday	26-Oct	Classification: Support Vector Machines	
	Wednesday	28-Oct	Classification: Support Vector Machines	
Week 7	Monday	2-Nov	Generative models, E-M algorithms	
	Wednesday	4-Nov	User Modeling/Opinion Mining/Recommendations	
Week 8	Monday	9-Nov	Student Presentations	Stage 3: status update
	Wednesday	11-Nov	*Veteran's Day* (no class)	
Week 9	Monday	16-Nov	FURLOUGH	
	Wednesday	18-Nov	Student Presentations	
Week 10	Monday	23-Nov	Student Presentations	
	Wednesday	25-Nov	Thanksgiving	
Week 11	Monday	30-Nov	Student Presentations	
	Wednesday	2-Dec	Student Presentations	Stage 3: reports due

Furloughs

During this academic year all Cal Poly faculty is observing a furlough.

Each full-time faculty member is required to observe **six days of furlough** during the Fall quarter.

I will be observing the following days:

No.	Date	Day of Week	Effect
1.	September 29	Tuesday	*no effect*
2.	**October 14**	Wednsday	no class, no office hours
3.	**October 23**	Friday	no class
4.	**November 16**	Monday	no class, no office hours
5.	November 27	Friday	*no effect* (Friday after Thanksgiving)
6.	December 3	Thursday	*no office hours*

I will be off-campus and unavailable for emails on furlough days.

Note. Please be aware that while the faculty are furloughed, the students are not. Unless otherwise announced in advance, the lectures are cancelled on furlough days, however, *Assignments may be due for electronic submission on faculty furlough days.*

Grading

Team Project	50%
Course Presentation	30%
Other assignments	20%

Course Policies

Prerequisites

The sole prerequisite for this course is CSC 365, Introduction to Databases. CSC 468 is **NOT a prerequisite**!

Some of you have taken CSC 466, Knowledge Discovery From Data, which can be viewed as a precursor to this course. To allow for extra enrollment CSC 466 is **NOT a prerequisite** to this course either. We will spend some time early in the quarter to recap CSC 466 material that is necessary for the course, but beyond that, we will concentrate on new material that should be equally accessible with and without CSC 466 experience.

Exams

The course has **no exams.** Instead, there is a quarter-long team project with a number of deliverables (proposal, presentation, report), which accounts for 50% of the course grade.

The final exam time is reserved for final team presentations.

In-Class Presentation

Each student in the course will at some point prepare and give a presentation on a topic related to the area of Data Mining and Knowledge Discovery from Data. This assignment will also have a number of deliverables (topic choice, bibliography, wiki page/lecture notes, presentation) spread over the quarter.

Two possible styles of presentation: (a) pick a sub-area or a problem in the area of KDD and prepare a survey of the state-of-the-art in it; (b) pick a specific paper and discuss in-depth its contribution to the field.

While each student will be judged individually on this assignment, I will allow collaboration (within reason - two-three people per serious topic is ok) in preparation of this assignment. Note, that if you are working on this assignment jointly with someone, you will have to cover proportionally more material.

Homeworks, Programming Assignments

While the course does not have an intensive lab-based hands-on component, we may have one-to-three small programming assignments in the course to ensure that you've had a chance to implement some of the important data mining techniques. (The only other programming in the course may/will come from your project).

Communication

The class will have an official mailing list. The email address for the mailing list is *csc-560-01-2098@calpoly.edu*. All students enrolled in the class are automatically subscribed to the mailing list (using the email addresss that the CS department has on file).

I encourage questions during classtime and questions via email. My answers to email questions may be broadcast to the entire class via the mailing list, if the answer may be relevant to everyone (e.g. a correction in a text of a handout, or a clarification of a homework problem), and may also appear on the web page. The questions can also be posted to the mailing list directly. The mailing list will also be used for all annoucements related to the course. It is your responsibility to read your class-related email. Failure to read email posted to the mailing list cannot be used as an excuse in the class.

Web Page

Class web page can be found at

http://www.csc.calpoly.edu/∼dekhtyar/560-Fall2009

Through this page you will be able to access all class handouts including homeworks, lab assignments, project information, lab/project data and lecture notes.

Links to additional information, and notes and announcements will also be posted.

Wikis

The course has its own wiki page at

http://wiki.csc.calpoly.edu/csc560

Note, that this is the generic wiki for all CSC 560 classes. All your material will be put under the Fall 2009 heading.

All data that we will come across throughout the course should/will be posted to the datasets wiki:

http://wiki.csc.calpoly.edu/datasets

You will have read-only access to the parts of the wiki relevant to the course. For write-access, contact me.

Academic Integrity

University Policies

Cal Poly's Academic Integrity policies are found at

http://www.academicprograms.calpoly.edu/academicpolicies/Cheating.htm

In particular, these policies define *cheating* as (684.1)

> "*...obtaining or attempting to obtain, or aiding another to obtain credit for work, or any improvement in evaluation of performance, by any dishonest or deceptive means. Cheating includes, but is not limited to: lying; copying from another's test or examination; discussion of answers or questions on an examination or test, unless such discussion is specifically authorized by the instructor; taking or receiving copies of an exam without the permission of the instructor; using or displaying notes, "cheat sheets," or other information devices inappropriate to the prescribed test conditions; allowing someone other than the officially enrolled student to represent same.*"

Plagiarism, per University policies is defined as (684.3)

> "... the act of using the ideas or work of another person or persons as if they were one's own without giving proper credit to the source. Such an act is not plagiarism if it is ascertained that the ideas were arrived through independent reasoning or logic or where the thought or idea is common knowledge. Acknowledgement of an original author or source must be made through appropriate references; i.e., quotation marks, footnotes, or commentary."

University policies state (684.2): "Cheating requires an "F" course grade and further attendance in the course is prohibited." (appeal process is also outlined, see the web site above for details.). Plagiarism, per university policies (684.4) can be treated as a form of cheating, although a level of discretion is given to the instructor, allowing the instuctor to determine the causes of plagiarism and effect other means of remedy. It is the obligation of the instructor to inform the student that a penalty is being assessed in such cases.

Course Policies

All homeworks are to be completed by each student **individually**. Lab assignments are to be completed by the appropriate units (individual, pair, group), and no code/solution-sharing between units is permitted. Students are encouraged to discuss class content among themselves but NOT in a manner that constitutes plagiarism and cheating as defined above (e.g., you can solve together a problem from the textbook that had not been assigned in the homework, but you should solve assigned problems individually).