EECS 485 - Web Databases & Information Systems
The University of Michigan
Fall 2013

Lectures	Mon, Wed 10:30AM - 12:00PM	**Professor**	Michael Cafarella
Location	1013 DOW	**Office**	4709 BBB
GSI	Jun Chen	**Professor Office Hours**	Monday 12:30-1:30 or by appt
GSI email	chjun@umich.edu	**Professor Email**	michjc@umich.edu

Course Description

This course is a contemporary exploration of modern web-based information systems. It will integrate concepts from multiple computer science topics used in the design, development, and deployment of web-based applications, services, and knowledge systems. While broad in scope, it will also cover several key concepts in depth, including: web networking protocols, web databases and applications, web services, web search, web-relevant security issues, web infrastructure, and web-relevant data mining. Students will learn how to incorporate these concepts into an engineering process that includes design, analysis, development and testing, using technologies such as HTTP, XML, JavaScript, AJAX, and others.

Students will form teams to implement assignments on Linux-based Apache web servers using open-source components. These assignments will culminate in students implementing their own large-scale Web search engine, roughly comparable to Google or Bing. At the end of this course, students will understand the science behind web-based information systems and the engineering principles for building them. This is a 4-credit course and satisfies the Software Area Kernel Requirement for MS and Ph.D. students in CSE.

Fully understanding the Web requires background from many different aspects of computer science. This course will try to bring together this disparate material and make you think about how these should work together to create a usable, efficient, and secure distributed information system. Most students will probably be familiar with a portion of the class material, but very few students will have background in all of the topics covered. (If you're among the lucky few, you might consider taking something else!)

Objectives

This course is about the design and development of information systems in wide area networks. Its primary goal is to take a holistic view of modern web systems and their constituent technologies. By the end of this course, successful students will be able to:

- Understand how n-tiered architectures can be used to implement secure, scalable systems
- Design and develop database-driven websites and applications
- Understanding XML as a messaging and data exchange mechanism
- Utilize JavaScript to improve database-driven websites
- Analyze server logs to understand system performance and user behavior
- Understand designs for modern search engines and datacenters
- Understand Web "semantic systems," such as auctions, recommendation systems, and

search ranking.

- Understand critical components of the modern Web infrastructure: DNS, Content Delivery Networks, etc.

Prerequisites

The formal prerequisite for this class is EECS 281 OR (EECS 282 and EECS 382). In past years EECS 484 was a requirement, but no more.

Less formally, a working knowledge of databases and SQL is required. If you have not taken EECS 484, you may have to do additional reading on your own.

There will be a substantial amount of programming in this class, and programming will not be a major topic of lectures. So you are expected to have "programming maturity." That is, you are deeply familiar with at least one programming language that is suitable for software engineering, such as Java, C, C++, Python, etc. You should be willing and able to pick up other similar languages. You should be able to understand and use new APIs by reading manuals and other relevant documentation. Students who may be unsure about their qualifications should approach the instructor with any questions.

By the end of class you will have written programs in three different languages: a Web development language, JavaScript, and Java/C++.

Requirements

Achieving the course objectives will require a significant amount of learning outside of class. Lectures will cover key topics and help integrate concepts, but will not necessarily cover all implementation details. Discussion sessions will be dedicated to development techniques, assignment details, and current lecture topics. Although significant support will be available, student teams will be required to research various technologies and development techniques to complete their assignments.

Coursework will consist of a sequence of 6 programming assignments, the last of which will integrate all of the previous components. This overall project will be a Web search engine capable of scaling to roughly the level of Google in 2004. (Google in 2004 was pretty good!) The project is intended to give enough flexibility that you can show off your creativity as well as what you learned in class.

All of the programming assignments are done as part of a group of 3 students. You will decide your fellow group members during the first week of class, and your group is intended to remain fixed for the duration of the class. If you do not know anybody in class, or have difficulty finding group mates, contact the instructor or GSI as soon as possible and we will help you create a group.

All team members will get the same grade on their joint work, except under very unusual (and generally unpleasant) circumstances. Some students are uncomfortable with the idea that team members will help determine each others' grades. This is unavoidable in EECS 485; if group work makes you uncomfortable, you will probably be happier taking a different class.

An important part of the course is working within a team environment to solve hard problems. As such, we may sometimes ask you to write a brief report that describes the contributions of each team member for a given assignment. The results may impact your grade on that assignment.

There will be one midterm and a final examination.

Textbook

There is no comprehensive textbook. Readings for this course are provided via CTools in the Resources/Readings directory. They are a set of PDF files specific to each topic along with links to outside information sources. These readings are required for the class and you will be expected to know their content. The lecture notes will also be made available shortly before each class.

The only book you should buy is a book about JavaScript. It has useful content you can't find in any other format, and it's reasonably priced:
JavaScript: The Good Parts, 1st Edition, by Douglas Crockford. O'Reilly, 2008.

If you want to have a supplemental text for the programming assignments (especially the first three), here are a few relatively inexpensive options. These are optional texts and are not required, but you might find them useful:

Web Database Applications with PHP and MySQL, by David Lane and Hugh Williams. O'Reilly, 2009.

PHP and MySQL Web Development, 4th Edition, by Luke Welling and Laura Thomson. Addison-Wesley, 2009.

Learning PHP, MySQL, JavaScript, and CSS, by Robin Nixon. O'Reilly, 2012.

Some of these texts are available in electronic format at lower cost (e.g., from Amazon or the O'Reilly website). Also, some of these texts may be available for free online for Michigan students. Check out http://proquest.safaribooksonline.com.

Assignments

Programming assignments are cumulative, with subsequent assignments depending upon previous ones. Make sure not to fall behind, or you will be in serious trouble. If you have an illness, a family situation, or other emergency, figure out a fair way to manage the load with your project partners.

There will be weekly written review questions for lecture material that will be ungraded. They are intended as a study aide and fodder for discussion section.

Exams

There will be a midterm exam and a final. The midterm will cover topics discussed in roughly the first half of the semester, and the final exam will be comprehensive. There will be no make-up exams. Make sure not to miss the midterm or final. The final exam will be held during the standard scheduled exam period for this class: Tuesday, December 17, 1:30PM - 3:30PM.

Tentative Lecture and Discussion Schedule

Note that this schedule only lists the first two discussion sections, but they will generally be held each week. After the first few weeks, each discussion section will cover questions on the pending programming assignment or the lecture-based ungraded review questions.

Note: This syllabus is subject to revision at any time.

1	Wed, Sep 4	Welcome & Web Overview	
DISC	Fri, Sep 6	Intro to Apache, MySQL	
2	Mon, Sep 9	Networking 1: TCP/IP	Prog1 OUT
3	Wed, Sep 11	Networking 2: TCP/IP cont'd, HTTP	
DISC	Fri, Sep 13	Intro to PHP	
4	Mon, Sep 16	Dynamic Page Generation	
5	Wed, Sep 18	Sessions and Personalization	Prog1 IN, Prog2 OUT
6	Mon, Sep 23	Web Security	
7	Wed, Sep 25	Public Key Cryptography	
8	Mon, Sep 30	JavaScript 1	Prog2 IN, Prog3 OUT
9	Wed, Oct 2	JavaScript 2	
10	Mon, Oct 7	XML and JSON	
11	Wed, Oct 9	NoSQL	
--	Mon, Oct 14	FALL STUDY BREAK	
12	Wed, Oct 16	Web Search 1: Information Retrieval	Prog3 IN, Prog4 OUT
13	Mon, Oct 21	Web Search 2: Information Retrieval	
--	Wed, Oct 23	**MIDTERM EXAM**	
14	Mon, Oct 28	Web Search 3: Link Analysis	
15	Wed, Oct 30	Web Search 4: Deduplication, LSH, etc.	
16	Mon, Nov 4	The Google File System	Prog4 IN, Prog5 OUT
17	Wed, Nov 6	MapReduce and Big Data	
18	Mon, Nov 11	Data Mining 1: Supervised Learning	
19	Wed, Nov 13	Data Mining 2: Unsupervised Learning	
20	Mon, Nov 18	Data Mining 3: Rule Induction	Prog5 IN, Prog6 OUT
--	Wed, Nov 20	THANKSGIVING MYSTERY TOPIC	
21	Mon, Nov 25	Data Mining 4: Text Mining	
22	Wed, Nov 27	DNS and Content Delivery Networks	
23	Mon, Dec 2	Auctions	
24	Wed, Dec 4	Recommendation Systems	
25	Mon, Dec 9	Inside The Datacenter	
26	Wed, Dec 11	Topic Review; Future of the Web	Prog6 IN

Discussion sections will be held at the following times:

- Thursdays at 4:30 (Dow 1006)
- Fridays at 10:30 (EECS 1500)
- Fridays at 12:30 (Dow 1010)
- Fridays at 2:30 (Dow 3150)

In general you should attend the discussion section for which you are registered. If you are sick or have a game when your section is scheduled to meet, you should attend a different section that week.

Grading

Your grade for the class will be determined by the following weighting:

Evaluation	**Percentage of Grade**
Programming Assignments	50
Midterm Exam	20
Final Exam	30
Total	**100**

In the event there are grading errors, please bring them to the attention of the GSI or instructor promptly. Please realize that we are careful in terms of applying a uniform grading policy, and so will not be able to make changes unless you have a particular special circumstance. No regrade requests will be entertained more than two weeks past the time the graded assignment or exam was returned to you -- it is very hard for us to grade consistently if we have to go back to things we did weeks ago and no longer have the material fresh in our heads.

Due Dates

Assignments are due at 11:55pm on the date indicated in the above schedule.

You may take a total of 4 late days total over all group programming assignments (except the last assignment), subject to a maximum of 2 late days for any assignment. These late days are intended to account for the inevitable illnesses, family visits, demands from other classes, etc. If you have a lot of assignments due the same day, just take a late day! Sick and can't work? Use a late day! That's what they're for.

Every additional late day beyond 4 will cost you 1% of the final grade (total for the course). Any assignment completed more than two days late will earn a zero for that assignment. All students in a group will be allocated late days and penalties together.

Every now and then there will be a severe emergency that cannot be covered by your late days, in which case you should approach us and ask (as early as possible). But we expect these situations to be very rare. Plan for contingencies. Allow some slack in your schedule.

Please note: *Your late days cannot be applied to the exams or assignment #6.* Each year some students save their late days to spend on the last assignment, only to realize in horror that their late days can no longer be used. Don't let that happen to you.

Programming Projects

The project in this class is cumulative: every piece will fit into the final system that you build. This is intended to be a major software engineering project on your part, and should result in a working search engine that would be genuinely useful to a general audience. There are some major features that are required, but you will also have the opportunity to exercise your creativity by adding lots of bells and whistles.

Keep in mind that some of the most interesting sites on the Web were created by just a couple of people. This is your chance to build something great.

Office Hours

The instructor, GSI, and IAs all hold regular office hours to help you with your questions.

Mike's office hours will be in BBB 4709 on Mondays 12:30-1:30pm, and by appointment.

The other office hours are as follows:
Monday: Matt 3-5
Tuesday: Jun 10-12
Wednesday: Otto 12-2
Thursday: Jun 3-4, Matt 4-5
Friday: Otto 1:30-2:30

The GSI and IAs' hours will be held in the CSE learning center (BBB 1637).

Field Trip!

As far as we know, we are the only class in EECS with a field trip! We will be visiting a large University of Michigan datacenter located off-campus in Ann Arbor (accessible via public transportation). You will see physical infrastructure that is similar, though not identical, to those used by large Web service providers like Google and Facebook. We will see the computers (thousands of them), electrical equipment, security arrangements, backup generators, and other items. It may sound a bit boring to go look at a bunch of hardware, but most people find the experience quite surprising and interesting. Attending the field trip is not mandatory, but is both highly encouraged and awesome. It will take place sometime in late November or early December. Details to follow.

Plagiarism

You may seek advice from any source -- the web, your classmates, others. However, every line of code you write must be your own. No copying of code from other sources, whether internal or external to the class, is permitted. Third party libraries and tools may be used, with adequate attribution, provided that the use of such library or tool does not render the assignment trivial; consult with the GSI or professor to make sure you have approval in advance. Violating the plagiarism rules may bring about a range of academic penalties, including (but not limited to) losing credit on the assignment, losing credit for the course, and even expulsion from the university.

Online Discussion

We will have an online discussion board for the class, reachable from the class website on CTools. If you have questions outside of class or office hours, please post them here. If you see a question you can answer, or simply have something interesting and relevant to say, please post away (keeping in mind the plagiarism rules above). We hope we can keep a lively conversation going on the board.

Accommodations for Students with Disabilities

If you think you need an accommodation for a disability, please let the instructor know as early as possible. Some aspects of this course may be modified to facilitate your participation and progress. As soon as you make the instructor aware of your needs, we can work with the Office of Services for Students with Disabilities (SSD) to help determine appropriate accommodations. SSD (734-763-3000; http://www.umich.edu/ sswd/) typically recommends accommodations through a Verified Individualized Services and Accommodations (VISA) form. We will treat any information you provide as private and confidential.

Some special arrangements take time to provision, so do let us know as soon as you can.