CS267 Fall 2018 Sec1 Home Page/Syllabus

Topics in Database Systems

Instructor: Chris Pollett
Office: MH 214
Phone Number: (408) 924 5145
Email: chris@pollett.org
Office Hours: MW 4:30-5:45pm
Class Meets:
Sec1 MW 3:00pm-4:15pm in MH223

Prerequisites

To take this class you must have taken: CS157B with a grade of C- or better.

Texts and Links

Required Texts:	Information Retrieval: Implementing and Evaluating Search Engines. Buttcher, Clarke, and Cormack
Online References and Other Links:	Yioop! Open Source Search Engine. Nutch. Wumpus. Heritrix.

Catalog Description

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.

Section-Specific Description

For this section, we will study information retrieval systems. Information Retrieval is the study of how to represent, search, and manipulate large collections of text and human data. Modern search engines such as Google, Bing, Baidu, Yandex are probably the most familiar examples of IR systems. Other examples are digital libraries (Melvyl), e-mail, and technical report systems, plagiarism systems such as turnitin.com, and even desktop search systems. Such systems are databases; however, the typical implementations of their building blocks such as indices, ordering result sets, and so on differs from conventional databases. The focus of this class is on implementation techniques for information retrieval systems, and also on measuring how effective the results returned from such systems are.

Course Learning Outcomes (CLOs)

By the end of this course, a student should be able to:

CLO1 -- Code a basic inverted index capable of performing conjunctive queries.

CLO2 -- Be able to calculate by hand on small examples precision (fraction relevant results returned), recall (fraction of results which are relevant), and other IR statistics.

CLO3 -- Be able to explain where BM25, BM25F and divergence from randomness statistics come from.

CLO4 -- Give an example of how a posting list might be compressed using difference lists and gamma codes or Rice codes.

CLO5 -- Demonstrate with small examples how incremental index updates can be done with log merging.

CLO6 -- Be able to evaluate search results by hand and using TREC evalsoftware.

CLO7 -- Know at least one Map Reduce algorithm (for example to calculate page rank).

Course Schedule

Below is a tentative time table for when we'll do things this quarter:

Week 1: Aug 20, Aug 22 (First Day)	Read Ch 1.1, 1.2 Introduction to IR
Week 2: Aug 27, Aug 29	Finish Ch 1
Week 3: Sep 3 (Labor Day), Sep 5	Read Ch 2.1-2.2, Phrase search, inverted indexes, VSM
Week 4: Sep 10, (HW1 due) Sep 12	Finish Ch 2 Recall and precision
Week 5: Sep 17, Sep 19	Read Ch 3 Stemming, stopping, and n-grams, will supplement with material on how to crawl
Week 6: Sep 24, Sep 26 (HW2 due)	Alexis Rossi from Internet Archive will speak Sep 24. Read Ch 4. Parts of inverted indexes and construction of them
Week 7: Oct 1, Oct 3 (Midterm 1)	Review
Week 8: Oct 8, Oct 10	Finish Ch 4
Week 9: Oct 15, Oct 17 (HW3 due)	Read Ch 5. Query processing techniques
Week 10: Oct 22, Oct 24	Finish Ch5. Start Ch 6. Index compression
Week 11: Oct 29, Oct 31	More Ch 6
Week 12: Nov 5, Nov 7	Finish Ch 6.
Week 13: Nov 12 (Campus Closed), Nov 14	Review
Week 14: Nov 19, (Midterm 2) Nov 21 (Thanksgiving)	Exam

Week 15: Nov 26, Nov 28	Read 7.1, 7.2, Incremental index updates, Read Ch 9. DFR
Week 16: Dec 3, Dec 5 (HW4 due)	Read Ch 14. Map reduce algorithms
Week 17: Dec 10 (Last Day), Dec 12	Finish Ch 14
	The final will be Friday, December 14 from 12:15-2:30pm

Grading

HWs and Quizzes	50%
Midterm 1	15%
Midterm 2	15%
Final	20%
Total	100%

Grades will be calculated in the following manner: The person or persons with the highest aggregate score will receive an A+. Since this is a graduate class, the curve will be slightly higher than for an undergrad course taught by me. A score of 55 will be the cut-off for a B-. The region between this high and low score will be divided into five equal-sized regions. From the top region to the low region, a score falling within a region receives the grade: A, A-, B+, B, B-. If the boundary between an A and an A- is 85, then the score 85 counts as an A-. Scores below 55 but above 50 receive the grade D. Those below 50 receive the grade F.

If you do better than an A- in this class and want me to write you a letter of recommendation, I will generally be willing provided you ask me within two years of taking my course. Be advised that I write better letters if I know you to some degree.

Course Requirements, Homework, Quiz Info, and In-class exercises

This semester we will have five homeworks, weekly quizzes, and weekly in-class exercises.

Every Monday this semester, except the first day of class, the Midterm Review Day, and holidays, there will be a quiz on the previous week's material. The answer to the quiz will either be multiple choice, true-false, or a simple numeric answer that does not require a calculator. Each quiz is worth a maximum of 1pt with no partial credit being given. Out of the total of twelve quizzes this semester, I will keep your ten best scores.

On Wednesday's, we will spend 15-20 minutes of class on an in-class exercise. You will be asked to post your solution to these exercises to the class discussion board. Doing so is worth 1 "pre-point" towards your grade. A "pre-point" can be used to get one missed point back on a midterm or final, up to half of that test's total score. For example, if you scored 0 on the midterm and have 10 pre-points, you can use your pre-points, so that your midterm score is a 10. On the other hand, if you score 18/20 on the midterm, you can use at most 1 pre-point since half of what you missed (2pts) on the midterm is 1pt.

Links to the current list of homeworks and quizzes can be found on the left hand frame of the class

homepage. After an assignment has been returned, a link to its solution (based on the best student solutions) will be placed off the assignment page. Material from assignments may appear on midterms and finals. For homeworks you are encouraged to work in groups of up to three people. **Only one person out of this group needs to submit the homework assignment; however, the members of the group need to be clearly identified in all submitted files.**

Homeworks for this class will be submitted and returned completely electronically. To submit an assignment click on the submit homework link for your section on the left hand side of the homepage and filling out the on-line form. Hardcopies or e-mail versions of your assignments will be rejected and not receive credit. Homeworks will always be due by the start of class on the day their due. Late homeworks will not be accepted and missed quizzes cannot be made up; however, your lowest score amongst the five homeworks and your quiz total will be dropped.

When doing the programming part of an assignment please make sure to adhere to the specification given as closely as possible. Names of files should be as given, etc. Failure to follow the specification may result in your homework not being graded and you receiving a zero for your work.

Classroom Protocol

I will start lecturing close to the official start time for this class modulo getting tangled up in any audio/visual presentation tools I am using. Once I start lecturing, please refrain from talking to each other, answering your cell phone, etc. If something I am talking about is unclear to you, feel free to ask a question about it. Typically, on practice tests days, you will get to work in groups, and in so doing, turn your desks facing each other, etc. Please return your desks back to the way they were at the end of class. This class has an online class discussion board which can be used to post questions relating to the homework and tests. Please keep discussions on this board civil. This board will be moderated. Class and discussion board participation, although not a component of your grade, will be considered if you ask me to write you a letter of recommendation.

Exams

The midterms will be during class time on: Oct 3 and Nov 19.

The final will be: Friday, December 14 from 12:15-2:30pm.

All exams are closed book, closed notes and in this classroom. You will be allowed only the test and your pen or pencil on your desk during these exams. The final will cover material from the whole semester although there will be an emphasis on material after the last midterm. No make ups will be given. The final exam may be scaled to replace a midterm grade if it was missed under provably legitimate circumstances. These exams will test whether or not you have mastered the material both presented in class or assigned as homework during the quarter. My exams usually consist of a series of essay style questions. I try to avoid making tricky problems. The week before each exam I will give out a list of problems representative of the level of difficulty of problems the student will be expected to answer on the exam. Any disputes concerning grades on exams should be directed to me, Professor Pollett.

Regrades

If you believe an error was made in the grading of your program or exam, you may request **in person** a regrade from me, Professor Pollett, during my office hours. **I do not accept e-mail requests for regrades.** A request for a regrade must be made no more than a week after the homework or a midterm is returned. If you cannot find me before the end of the semester and you would like to request a regrade of your final, you may see me **in person** at the start of the immediately following semester.

University Policies and Procedures

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/. Below are some brief comments on some of these policies as they pertain to this class.

Academic Integrity

For this class, you should obviously not cheat on tests. For homeworks, you should not discuss or share code or problem solutions between groups! At a minimum a 0 on the assignment or test will be given. A student caught using resources like Rent-a-coder will receive an F for the course. Faculty members are required to report all infractions to the Office of Student Conduct and Ethical Development.

Accommodations

If you need a classroom accommodation for this class, and have registered with the Accessible Education Center, please come see me earlier rather than later in the semester to give me a heads up on how to be of assistance.