CS267 Fall 2023 Sec2 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 1:30-2:45pm in MH214
Class Meets:
Sec2 MW 12:00-1:15pm in MH422

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. Heritrix.

Description

From the catalog: 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. 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 eval software.

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 21, Aug 23	Read Ch 1.1, 1.2 Introduction to IR
Week 2:Aug 28, Aug 30	Finish Ch 1
Week 3:Sep 4 (No Class), Sep 6	Read Ch 2.1-2.2, Phrase search, inverted indexes, VSM
Week 4:Sep 11(Hw1), Sep 13	Finish Ch 2 Recall and precision
Week 5:Sep 18, Sep 20	Read Ch 3 Stemming, stopping, and n-grams, will supplement with material on how to crawl
Week 6:Sep 25, Sep 27	Read Ch 4. Parts of inverted indexes and construction of them
Week 7:Oct 2(Hw2), Oct 4	Finish Ch 4
Week 8:Oct 9, Oct 11(Midterm)	Review
Week 9:Oct 16, Oct 18	Read Ch 5. Query processing techniques
Week 10:Oct 23(Hw3), Oct 25	Ch 6 Index compression,Ch 7.1, 7.2 Incremental index update
Week 11:Oct 30, Nov 1	Ch 9 Ranking functions LMJM, LMD, pseudo-relevance feedback, DFR
Week 12:Nov 6, Nov 8	Read Ch 14. Map reduce algorithms
Week 13:Nov 13(Hw4), Nov 15	Ch 15 Document Quality Measures, Web Search
Week 14:Nov 20, Nov 22 (No Class)	Ch 10.1 , 10.2 Survey Categorization and Filtering
Week 15:Nov 27, Nov 29	Vertical Search Engines
Week 16:Dec 4, Dec 6(Hw5)	Finish Vertical Search Engines, Review
	The final will be Thursday, December 14 9:45 AM-12:00 PM

Grading

HWs and Quizzes	50%
Midterm	20%
Final	30%
Total	100%

Grades will be calculated in the following manner: The person or persons with the highest aggregate score will receive an A+. 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 thirteen 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 "insurance point" towards your grade. A "insurance 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 insurance points, you can use your insurance 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 insurance point since half of what you missed (2pts) on the midterm is 1pt. In addition, to the weekly in-class exercises, one insurance point is available if in the week before the midterm you can convince me I know your name, and in the week before the final, I still know your name (Please help me improve my memory).

Links to the current list of homeworks and quizzes can be found on the left hand side 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 using the Canvas link for the name of the homework. Hardcopies or e-mail versions of your assignments will be rejected and not receive credit. Homeworks will always be due by midnight according to the Canvas server on the day their due. Late homeworks will not be accepted and missed quizzes cannot be made up; however, your lowest score amongst the first four homeworks and your quiz total will be dropped. Homework 5 can't be substituted for.

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 midterm will be during class time on: Oct 11.

The final will be: Thursday, December 14 9:45 AM-12:00 PM.

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

SJSU adheres to required safety measures from the California Department of Public Health and the Santa Clara County Public Health Department. Please refer to our SJSU Health Advisories website for the latest information and updates.

Per University Policy S16-9, relevant university policy concerning all courses, such as student responsibilities, academic integrity, accommodations, dropping and adding, consent for recording of class, etc. and available student services (e.g. learning assistance, counseling, and other resources) are listed on Syllabus Information web page (https://www.sjsu.edu/curriculum/courses/syllabus-info.php). Make sure to visit this page to review and be aware of these university policies and resources. 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. 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.