CSCE 4200/5200 Information Retrieval and Web Search

Instructor Contact

Name: Dr. Ting Xiao
Email: ting.xiao@unt.edu
Office Location: F213
Office Hour: 4:15 - 5:15 pm Monday

TA Name: Beilei Jiang
TA Email: beileijiang@my.unt.edu
Office Location: F232
TA Office Hours: 10:30 am - 12:00 pm Wednesday

Communication Expectations: Students will be expected to regularly check their university email and the course canvas page for relevant updates or announcements. If there are personal concerns or questions, it is suggested that students discuss before or after class or during office hours, and at other times send the instructor an email. You can expect email responses within 24 hours M-F with possible delays over the weekend. Feedback on assignments and grades will be approximately one week after assignments are due.

Course Description

This course will introduce students to text-based information retrieval (IR) techniques, i.e. search engines (e.g., Google, Yahoo, etc.). Various IR models such as the Boolean model, vector space model, probabilistic models and machine learning models will be studied. Efficient indexing techniques for both general document collections and specialized collections (strings, digital library) will be examined. The course will also cover web search engines techniques, such as hyperlink analysis (e.g., PageRank (used in Google search)), Web technologies and representations, and query languages, with a focus on techniques that can be used to access, retrieve, organize, and present information. Students will work on programming projects to gain hands-on experience in building an IR system.

Materials

Much of the material in the course is based on the book and lecture material of Chris Manning at Stanford University from their course of the same name.

- Textbook information: Introduction to Information Retrieval by C. Manning, P. Raghavan, and H. Schütze, Cambridge University Press. Note, a free online version of this book is available at http://informationretrieval.org

- Supplementary material:
  Additional web content from Stanford’s CS 276: Information Retrieval and Web Search is optional
  http://web.stanford.edu/class/cs276/

Course Structure

Time: MW 5:30-6:50pm
Location: Discovery Park B142
Credit hours: 3
15 Weeks: August 26 - December 11, 2019

Tentative topics (with chapter numbers)
01 Boolean retrieval
02 The term vocabulary & postings lists
03 Dictionaries and tolerant retrieval
04 Index construction
06 Scoring, term weighting & the vector space model
07 Computing scores in a complete search system
08 Evaluation in information retrieval
11 Probabilistic information retrieval
13 Text classification & Naive Bayes
14 Vector space classification
15 SVM & ML on documents
16 Flat clustering
17 Hierarchical clustering
21 Link Analysis (PageRank)

Course Prerequisites

CSCE 3110: Data Structures and Algorithms, for CSCE 4200

Programming experience (Python preferred). Introductory courses on data structures and algorithms. Linear algebra and probability theory preferred.

Course Objectives

By the end of this course, students will be able to:

1. Perform efficient text-based indexing
2. Use traditional and machine learning-based ranking approaches
3. Perform document clustering and classification
4. Understand information retrieval techniques for the web, including link-based algorithms

Teaching Philosophy

There are facts about information retrieval that you will pick up in this class, but the best long-lasting benefits come from the implicit learning that will happen. You will understand a new approach to organizing information that is readily applicable to more than text processing. And work hard and creatively on the projects and assignments and your programming ability will improve. Though the title is Information Retrieval, the skills and approaches from this class will help you well beyond this semester.

Technical Requirements & Skills

Minimum Technology Requirements

- Computers are required
  - You can choose to use your own laptop, or use the machines in the lab
  - In-class exams: These will be done individually on your computer in class. You will be expected to connect to the UNT wireless network
  - There will be in-class activities along with lecture that are not required, but will help in understanding and applying the material
- Canvas Technical Requirements (https://clear.unt.edu/supported-technologies/canvas/requirements)

Computer Skills & Digital Literacy

- Using Canvas
- Download and install Anaconda Python version 3 and open a Jupyter notebook

Netiquette

Netiquette, or online etiquette, refers to the way students are expected to interact with each other and with their instructors online. Here are some general guidelines:

- Treat your instructor and classmates with respect in email or any other communication.
- Use clear and concise language.
- Remember that all college level communication should have correct spelling and grammar (this includes discussion boards).
- Be cautious when using humor or sarcasm as tone is sometimes lost in an email or discussion post and your message might be taken seriously or sound offensive.
- Be careful with personal information (both yours and other’s).
- Do not send confidential information via e-mail

Course Requirements

The following activities are graded, with anticipated point values shown. Note, point values shown are only approximate to give a sense of expected value and effort for each.

- Assignments: 40 pts (10 pts each)
- Midterm: 15 pts

- Projects: 20 pts
    - Brainstorming - 1 pt
    - Proposal - 2 pts
    - Update - 2 pts
    - Presentation - 5 pts
    - Final Report - 5 pts
    - Group Self-Assessment - 5 pts
- Final Exam: 25 pts

Grading

Grades are determined by a simple points system, with a total of at least 100 pts as the goal though more than 100 points are likely. The expected distribution of points is given above, with the exact scale determined by point values given for each assignment, quiz, or exam - this is subject to minor modification based on actual points given. Note, due to the nature of the course, exams and quizzes are a significant means of establishing your final grade, so please complete the assignments in a timely way and study appropriately prior to each quiz and exam.

Grading Scale: Points needed to get each grade: A=90.0, B=80.0, C=70.0, D=60.0. Don't expect this scale to change. If class grades are low (I expect the vast majority of students will get A's and B's), extra quizzes or assignments will be given to add points. (Note, **these grades are based on points and not percentages, so if 120 points are given, you only need 93/120 to get an A!**)

CSCE 4200 vs 5200 grading: In line with the added expectations for the graduate version of the course, **students enrolled in CSCE 4200 will receive 5 additional bonus points**.

Course Evaluation

Student Perceptions of Teaching (SPOT) is the student evaluation system for UNT and allows students the ability to confidentially provide constructive feedback to their instructor and department to improve the quality of student experiences in the course. Spot evaluations will be available **November 18 - December 5th, 2019**

Course Policies

Assignment Policy

Assignments are generally due within a week after they are assigned (e.g., Monday assignments are due the following monday before class) unless otherwise specified, and are to be turned in through Canvas**.** Assignments, project reports and presentation slides are to be turned in as PDF. Code is to be turned in with both Jupyter notebook and PDF form, along with any files necessary to run your assignment. Results should be presentable, with appropriate comments for someone to follow what you have done. Assignments are to be turned in individually, although students are encouraged to work together extensively in all ways except direct copy/paste of individual work - do keep in mind that only applied to assignments, not quizzes or exams. It is important to keep up given the pace of new assignments. See the late policy below.

There will be no extra credit or special assignments for individuals in this class, however, extra opportunities for points given to the entire class will be considered toward the end of the course if scores are low.

If during an online quiz or exam there is a technical error which affects your ability to complete the assignment, you are immediately to let the quiz or exam proctor know and the instructor will discuss ways to allow you to resume the test without giving an unfair advantage. In the event of any unexpected server outage or any unusual technical difficulty which prevents students from completing a time sensitive assessment activity, the instructor will extend the time windows and provide an appropriate accommodation based on the situation. Students should immediately report any problems to the instructor.

Examination Policy

Quizzes and exams will primarily be on the computer. If you would like to use a laptop be sure to bring it on the appropriate days. Quizzes and exams are "open book", open note, and open internet unless otherwise specified in advance. Quizzes and exams must be taken in the classroom unless special accommodations have been made through the Office of Disability Accommodation (ODA). Another other accommodations must be given by prior arrangement with the instructor, otherwise documentation proving an extenuating circumstance will have to be provided after the missed exam. Time will be limited, and all work will be individual. You are strongly encouraged to attempt to solve the tasks iteratively and incrementally - write code that works first, but works poorly, and improve from there, rather than write perfect code top to bottom. Exams will focus on the most recent material but are expected to be cumulative in scope.

Instructor Responsibilities and Feedback

The instructor's goals are not only to provide a series of organized facts and examine you, rather the ultimate goal is to help the students grow and learn using the content as a means to achieving that goal. To that end, you are encouraged to discuss any ways in which the course or the instructors experience may help you achieve your professional aspirations. You will be provided with clear instructors for all projects and assignments, grading rubrics, and course content to focus on learning and adapting the information rather than guess expectations.

The instructor will try to respond to all questions by email within 24 hours during business days Mon - Fri, and graded feedback will regularly come within one week of the assignment or exam due date.

Late Work

If assignments or project work are turned in after the due date, this places an undue burden on the instructor and the TA, especially when this policy is abused. This is heightened given the pace of this course. In general, late assignments will not be accepted, though you are encouraged to discuss with the instructor if there are extenuating circumstances, in which case a point reduction is at the discretion of the instructor.

Attendance Policy

You are expected to attend all lectures and to complete all readings on time, however, this course does not use participation points and there is no penalty for missing days without exams, or group project efforts. You are responsible for keeping up with the material covered in the class if you are not present. If a class is missed, you are expected to proactively reach out to classmates or the instructor about in-class announcements.

Class Participation

Individual attendance is not required except on group project and exam days. There is no direct participation grading, but in the past there has been a strong correlation between engagement and accomplishment in courses - especially for those that are struggling with the material. Feel free to prioritize your time, but prioritize wisely.

Syllabus Change Policy

Any substantial changes to the syllabus after the first week will be highlighted in red on the online platform. Approximate point values are expected to vary but will be fixed when the assignment or exam is given.

UNT Policies

Academic Integrity Policy

Academic Integrity Standards and Consequences. According to UNT Policy 06.003, Student Academic Integrity, academic dishonesty occurs when students engage in behaviors including, but not limited to cheating, fabrication, facilitating academic dishonesty, forgery, plagiarism, and sabotage. A finding of academic dishonesty may result in a range of academic penalties or sanctions ranging from admonition to expulsion from the University.

ADA Policy

UNT makes reasonable academic accommodation for students with disabilities. Students seeking accommodation must first register with the Office of Disability Accommodation (ODA) to verify their eligibility. If a disability is verified, the ODA will provide a student with an accommodation letter to be delivered to faculty to begin a private discussion regarding one's specific course needs. Students may request accommodations at any time, however, ODA notices of accommodation should be provided as early as possible in the semester to avoid any delay in implementation. Note that students must obtain a new letter of accommodation for every semester and must meet with each faculty member prior to implementation in each class. For additional information see the ODA website (https://disability.unt.edu/).

Emergency Notification & Procedures

UNT uses a system called Eagle Alert to quickly notify students with critical information in the event of an emergency (i.e., severe weather, campus closing, and health and public safety emergencies like chemical spills, fires, or violence). In the event of a university closure, please refer to Blackboard for contingency plans for covering course materials.

Retention of Student Records

Student records pertaining to this course are maintained in a secure location by the instructor of record. All records such as exams, answer sheets (with keys), and written papers submitted during the duration of the course are kept for at least one calendar year after course completion. Course work completed via the Blackboard online system, including grading information and comments, is also stored in a safe electronic environment for one year. Students have the right to view their individual record; however, information about student's records will not be divulged to other individuals without proper written consent. Students are encouraged to review the Public Information Policy and the Family Educational Rights and Privacy Act (FERPA) laws and the University's policy. See UNT Policy 10.10, Records Management and Retention for additional information.

Acceptable Student Behavior

Student behavior that interferes with an instructor's ability to conduct a class or other students' opportunity to learn is unacceptable and disruptive and will not be tolerated in any instructional forum at UNT. Students engaging in unacceptable behavior will be directed to leave the classroom and the instructor may refer the student to the Dean of Students to consider whether the student's conduct violated the Code of Student Conduct. The University's expectations for student conduct apply to all instructional forums, including University and electronic classroom, labs, discussion groups, field trips, etc. Visit UNT's Code of Student Conduct (https://deanofstudents.unt.edu/conduct) to learn more.

Access to Information - Eagle Connect

Students' access point for business and academic services at UNT is located at: my.unt.edu. All official communication from the University will be delivered to a student's Eagle Connect account. For more information, please visit the website that explains Eagle Connect and how to forward e-mail Eagle Connect (https://it.unt.edu/eagleconnect).

Student Evaluation Administration Dates

Student feedback is important and an essential part of participation in this course. The student evaluation of instruction is a requirement for all organized classes at UNT. The survey will be made available during weeks 13, 14 and 15 of the long semesters to provide students with an opportunity to evaluate how this course is taught. Students will receive an email from "UNT SPOT Course Evaluations via IASystem Notification" (no-reply@iasystem.org) with the survey link. Students should look for the email in their UNT email inbox. Simply click on the link and complete the survey. Once students complete the survey they will receive a confirmation email that the survey has been submitted. For additional information, please visit the SPOT website (http://spot.unt.edu/) or email spot@unt.edu.

Getting Help

Technical Assistance

UIT Help Desk (http://www.unt.edu/helpdesk/index.htm)
Email: helpdesk@unt.edu
Phone: 940-565-2324
In Person: Sage Hall, Room 130
Walk-In Availability: 8am-9pm
Telephone Availability:

- Sunday: noon-midnight
- Monday-Thursday: 8am-midnight
- Friday: 8am-8pm
- Saturday: 9am-5pm

Laptop Checkout: 8am-7pm

Student Support Services

- Registrar (https://registrar.unt.edu/registration)
- Financial Aid (https://financialaid.unt.edu/)
- Student Legal Services (https://studentaffairs.unt.edu/student-legal-services)
- Career Center (https://studentaffairs.unt.edu/career-center)
- Multicultural Center (https://edo.unt.edu/multicultural-center)
- Counseling and Testing Services (https://studentaffairs.unt.edu/counseling-and-testing-services)
- Student Affairs Care Team (https://studentaffairs.unt.edu/care)
- Student Health and Wellness Center (https://studentaffairs.unt.edu/student-health-and-wellness-center)
- Pride Alliance (https://edo.unt.edu/pridealliance)

Academic Support Services

- Academic Resource Center (https://clear.unt.edu/canvas/student-resources)
- Academic Success Center (https://success.unt.edu/asc)
- UNT Libraries (https://library.unt.edu/)
- Writing Lab (http://writingcenter.unt.edu/)
- MathLab (https://math.unt.edu/mathlab)