CSCE 670 :: Information Storage and Retrieval :: Spring 2020

MWF 11:30am-12:20pm in ZACH 310

Instructor: James Caverlee, HRBB 403
Office Hours: TBA
Department of Computer Science and Engineering
Texas A&M University

TA: Yun He, HRBB 408D
Office Hours: TBA

Course Schedule :: Spotlight :: Project

Course Summary

What is this course about?

- One of the classic IR textbooks says that "Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)."
- Gerard Salton "the father of Information Retrieval" said that "Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information."
- More recently, Markov and de Rijke say that "IR is about technology to connect people to information. In our view, this includes search engines, recommender systems, and task-oriented dialogue systems."

In this course, we'll study the theory, design, and implementation of foundational IR systems, but also examine closely modern web search and recommender systems, including algorithms and techniques at the core of how people connect to information. Broadly, what are the principle ideas, algorithms, and systems for organizing information? By the end of the semester you will be able to:

- Define and explain the key concepts and models relevant to web search, including topics like text indexing, retrieval models, evaluation, Web crawling, link-based algorithms like PageRank, and learning to rank.
- Define and explain the key concepts and models relevant to recommender systems, including topics like collaborative filtering, matrix factorization, recommender system evaluation, and implicit recommendation.
- Design, implement, and evaluate the core algorithms underlying a fully functional web search system and recommendation system.
- Identify the salient features and apply recent research results in web search and recommender systems, including topics such as adversarial information retrieval and neural models of retrieval and recommendation.

Communication

All course communication will be via Piazza. We will post often to Piazza, so you should plan to check it often (every day).

Prerequisites

I expect all students to have had some previous exposure to basic probability, statistics, algorithms, and data structures. You should be able to design and develop large programs and learn new software libraries on your own.

Textbooks

There is no one single textbook for this course. We may read some selections from

- IIR: *Introduction to Information Retrieval*, Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze.
- MMD: *Mining of Massive Datasets*, Jure Leskovec, Anand Rajarman, and Jeff Ullman.
- SEIRP: *Search Engines: Information Retrieval in Practice*, by Croft, Metzler, and Strohman.
- DITP: *Data-Intensive Text Processing with MapReduce*, by Lin and Dyer, 2010.
- NCM: *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*, David Easley and Jon Kleinberg, Cambridge University Press. 2010.
- As well as many papers and other resources provided on the course website and/or on Piazza.

You may find some of these optional textbooks helpful, though none are required:

- *Modern Information Retrieval*, by Baeza-Yates and Ribeiro-Neto.
- *Managing Gigabytes*, by Witten, Moffat, and Bell.
- *Foundations of Statistical Natural Language Processing*, by Manning and Schutze.

It is critically important that you study the relevant course readings before class so that we can make the most of our limited class time together. I treat our class meetings as opportunities to highlight significant aspects of the material, to answer questions, to engage in discussions about particular topics, and so on. We cannot cover all of the material in class, so it is up to you to stay on top of the readings and the assignments.

Grading

The grading scale is A: 90-100, B: 80-89, C: 70-79, D: 60-69, F: 0-59. The course grading policy is as follows:

Participation (5%). Attendance in class and participation in the discussion are both important to your success in the course. We expect you to participate in online discussions on Piazza. Over the course of the semester, you should **post at least three** posts or replies to the discussion forum on Piazza. These posts can start a new thread or respond to an existing one. Since we encounter search and recommendation every day, there are ample opportunities to connect what we talk about in class to new research results, new features on existing platforms, challenges facing industry, ethical considerations, etc. Towards your participation grade, the final day to post to the discussion group is April 22. (Of course you are welcome to continue to post afterwards, but these posts will not count toward your participation grade.) Also note that your project-related posts do not count towards this participation score (e.g., posting a project proposal is a requirement of the project and does not count here).

Spotlight (10%). A spotlight is an opportunity to share a compelling aspect of web search or recommender systems -- be it, a neat feature or library you want to share via a Jupyter notebook, a discussion and brief exploration of a particular dataset, an in-depth look at a research paper, etc. The overarching goal of these spotlights is to help you transition your theoretical foundations into practice. Each spotlight must include a Jupyter notebook.

Three In-class Quizzes (30%). We will have three in-class quizzes, each worth 10% of your overall grade. Each quiz will have around 3-4 questions. For each quiz, you may bring **one** standard 8.5" by 11" piece of paper with any notes you deem appropriate or significant (front and back). No devices allowed.

- Quiz 1: February 7 (Fri)
- Quiz 2: March 18 (Wed)
- Quiz 3: April 22 (Wed)

Homework (30%). We will have several homework assignments. These will be a mix of programming assignments and problem sets. All programming assignments will be in Python; we make no expectations that you have been exposed to Python before, but we do expect you to come up to speed rapidly.

All homework assignments must be submitted by 11:59pm Central time on the due date. For the homework assignments, you may talk to any other class member or work in groups to discuss the problems **in a general way**. However, your actual detailed solution must be yours alone. If you do talk to other students, you must write on your assignment who it is that you discussed the problems with. Your submitted work must be written solely by you and not contain work directly copied from others.

Homework Collaboration Clarification: To clarify, your homework is yours alone and you are expected to complete each homework independently. Your solution should be written by you without the direct aid or help of anyone else. However, we believe that collaboration and team work are important for facilitating learning, so we encourage you to discuss problems and general problem approaches (but not actual solutions) with your classmates. If you do have a chat with another student about a homework problem, you must inform us by writing a note on your homework submission (e.g., Bob pointed me to the relevant section for problem 3). The basic rule is that no student should explicitly share a solution with another student (and thereby circumvent the basic learning process), but it is okay to share general approaches, directions, and so on. If you feel like you have an issue that needs clarification, feel free to contact either me or the TA.

Homework Plagiarism Policy: We will use the Stanford Moss system to check homework submissions for plagiarism. Students found to have engaged in plagiarism will be punished severely, typically earning an automatic F in the course and being reported to the Aggie Honor System.

Homework Late Days: For the homework assignments, you have a total of **5 late days** that you can use during the semester. However, a single assignment can be submitted **up to 3 days late** only, so we can post solutions in a timely fashion. For the purposes of the class, a late day is an indivisible 24-hour unit. Once you exhaust your 5 late days, we will not accept any late submissions.

Project (25%). For the project, you will work in teams of three or four on a problem of your choosing that is interesting, significant, and relevant to this course.

Regrade Policy: If you feel that we have made an error in grading, you may resubmit the assignment for a regrade within 7 days of receiving your graded assignment. You must include a brief written statement describing what portion has been graded in error. Note that we reserve the right to examine the entire assignment, so there is a chance we may find errors in your assignment that we missed before.

Americans with Disabilities Act (ADA) Policy Statement

The Americans with Disabilities Act (ADA) is a federal anti-discrimination statute that provides comprehensive civil rights protection for persons with disabilities. Among other things, this legislation requires that all students with disabilities be guaranteed a learning environment that provides for reasonable accommodation of their disabilities. If you believe you have a disability requiring an accommodation, please contact Disability Services, currently located in the Disability Services building at the Student Services at White Creek complex on west campus or call 979-845-1637. For additional information, visit http://disability.tamu.edu.