General information

Last edited: September 3, 2024, 1:48 PM

COMPSCI 446 Search Engines
Fall 2024

Credit Hours: 3 credits
Prerequisites: CMPSCI 240 or CMPSCI 383, or equivalent.
Gradescope entry code: 5KD76X (linked to Canvas so you should be signed up)
Piazza access code: 6yqjo7rajnf

Instructor

- James Allan
- https://cs.umass.edu/~allan
- allan@cs.umass.edu (preferred)
- CS Building, room 370
- 413-545-3240 (unreliable so usually voicemail only)

Class meetings:

- Tuesdays and Thursday 4:00-5:15pm
- Hasbrouck Laboratory 134

Office Hours:

- Currently TBD; will be posted on Moodle once set

Course Objectives

Information Retrieval (IR) is the theory and practice that underlies technologies such as search engines. It deals with models and methods for representing, indexing, searching, browsing, and summarizing information in response to a person's information need. This course provides an overview of the important issues in information retrieval, and how those issues affect the design and implementation of search engines. The course emphasizes the technology used in Web search engines, and the information retrieval theories and concepts that underlie all search applications. Mathematical experience (as provided by CMPSCI 240) is required. You should also be able to program in Python (though we will attempt to support Java to the extent possible).

NB: This course may be considered programming intensive by some. If you have heard about past versions of the course, this semester we are using Jupyter notebooks rather than asking you to program from scratch in an IDE. That does push everyone toward Python because other languages (e.g., Java) are not well supported.

Learning Outcomes

At the end of this course you:

1. will understand the basic computational models for representing text and information needs (queries) and how those models allow us to rank documents by their likelihood of being relevant to the information need;
2. will understand how to implement a basic working search engine, based on your ability to select the appropriate data structures and algorithms to enable building a performant system;
3. will understand the key ideas of how search engines are evaluated in the laboratory and in commercial settings; and,
4. will be able to use the techniques to solve other IR related problems, such as computing PageRank, performing evaluation of a retrieval system using a variety of evaluation metrics, perform clustering and classification on document collections and more.

Course materials and texts

The following text is required for this course:

- B. Croft, D. Metzler, and T. Strohman, *Search Engines: Information Retrieval in Practice*. Addison Wesley, February 2009. Available for free download at https://ciir.cs.umass.edu/irbook. It is also available from Amazon for about US$73 new or US$38 used (as of August 2024). It may, of course, be available from other sellers for less.

You may find the following textbook useful for understanding some of the material, but it is not required:

- C. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008. The authors of this text maintain a web site with information about the book, including a couple of on-line versions of the text. It is also available from Amazon for about US$58 new or US$30 used (as of August 2024). It may, of course, be available from other sellers for less.

List of Topics

The following topics will be covered in this course, though not in the order listed (which corresponds to the textbook's ordering). Some topics may be omitted in response to student interest and class discussion. Topics will definitely be reordered.

1. Search Engines and Information Retrieval
2. Architecture of a Search Engine

3. Acquiring Data
    - Crawling the Web
    - Document Conversion
    - Storing the Documents
    - Detecting Duplicates, removing noise

4. Processing Text
    - Text Statistics, document parsing
        - Tokenizing, stopping, stemming, phrases, structure, links, internationalization
        - Named Entity Recognition

5. Ranking with Indexes
    - Abstract Model of Ranking
    - Inverted indexes, MapReduce
    - Query Processing
        - Document-at-a-time evaluation
        - Term-at-a-time evaluation
        - Optimization techniques, structured queries, distributed evaluation, caching

6. Queries and Interfaces
    - Information Needs and Queries
    - Query Transformation and Refinement
        - Stopping and Stemming Revisited
        - Spell Checking and Query Suggestions
        - Query Expansion, Relevance Feedback
        - Context and Personalization
    - Displaying the Results
        - Result Pages and Snippets
        - Advertising and Search
        - Clustering the Results
        - Translation

7. Retrieval Models
    - Traditional Retrieval Models
        - Boolean, Vector Space Models
    - Probabilistic Models
        - Information Retrieval as Classification
        - The BM25 Ranking Algorithm
    - Ranking based on Language Models
        - Query Likelihood Ranking
        - Relevance Models and Pseudo-Relevance Feedback
    - Complex Queries and Combining Evidence
        - The Inference Network Model
        - The Galago Query Language
    - Web search
    - Machine Learning and IR

8. Evaluating Search Engines
    - Test collections
    - Query logs
    - Effectiveness Metrics
        - Recall and Precision
        - Averaging and interpolation
        - Focusing on the top documents
    - Training, Testing, and Statistics
        - Significance tests
        - Setting parameter values

9. Classification and Clustering

10. Social Search
    - User tagging
    - Filtering and recommending

11. Deep Learning for IR
    LLMs for IR

Grading Criteria

Your final grade in this class will be based upon the following:

- In-class exercises (X1 to X13): 5%
- At-home low-stakes reviews (R1 to R13): 10%
- Programming assignments (P1 to P3): 45%
- Midterm and final exams: 20% each: 40%

The _in-class exercises (X*) are 15- to 20_-minute small-group exercises during class where you work in groups of 2-4 students to respond to a short prompt related to the material in class and/or recent readings. Their purpose is to allow you to stop and think about the lecture content and how it relates to other material you've been learning in the course. They will be graded on a 5-point scale of excellent (4-5), satisfactory (2-3), poor (1), fail/absent (0). Unless otherwise noted, these assignments cannot be made up for any reason and cannot be redone; they're just not worth enough for that to make sense.

The class will also include _13 at-home reviews of material_ and _3 programming projects_. All assignments are due as indicated on the assignment on Canvas and/or Gradescope.

The _programming projects (P1-3)_ allow 3 late days combined across them all that students can use at their discretion. Note that using only a few minutes of a late day counts as a full day. Otherwise, late assignments will only be accepted in accordance with University policy, at the sole discretion of the instructor. Accommodation will be granted when possible, following the guidance of the campus Disability Services office.

The _mid-term exam_ will be held 7-9pm in mid-October (tentatively on October 22), location TBD. The _final exam_ is offered during the University-scheduled time shown on SPIRE – i.e., December 12, 3:30-5:30pm in Thompson Hall 104 (that's the first day of final exams). Both exams are in-person. Watch for class announcements in case the time or room changes.

The (non-standard) points to letter grade table is as follows:

Score is	**Letter**	**Score is**	**Letter**
90 ≤ score	A	65 ≤ score < 70	C
87 ≤ score < 90	A-	60 ≤ score < 65	C-
84 ≤ score < 87	B+	55 ≤ score < 60	D+
77 ≤ score < 84	B	50 ≤ score < 55	D
74 ≤ score < 77	B-	score < 50	F
70 ≤ score < 74	C+		

Initial Course Schedule

This schedule may change based on student interest, opportunities, or other unusual circumstances. An up-to-date schedule will be maintained on Canvas.

Class Number and Date	**Topic/session**	**Read**	**Quick reminders / deadlines**
1. Tue Sep 3	Introduction	Ch. 1&2	
2. Thu Sep 5	Processing Text	Ch. 4	
Mon Sep 9			*Add/drop deadline*
3. Tue Sep 10	Processing Text		
4. Thu Sep 12	PageRank	§4.5.2	
5. Tue Sep 17	Collecting Docs	Ch. 3	
6. Thu Sep 19	Collecting Docs		
7. Tue Sep 24	Collecting Docs		
8. Thu Sep 26	Evaluation	Ch. 8	
Sun Sep 29			*P1 due before midnight*
9. Tue Oct 1	Evaluation		
10. Thu Oct 3	Indexing	Ch. 5	
11. Tue Oct 8	Indexing		
12. Thu Oct 10	Indexing		
Mon Oct 14	*Holiday*		*Holiday (Indigenous Peoples Day)*
Tue Oct 15	No class		Monday class schedule
13. Thu Oct 17	Indexing/review		
14. Tue Oct 22	No class (but possibly review)		EVENING MIDTERM EXAM (tentative)
15. Thu Oct 24	Indexing/Retrieval Models	Ch. 7	
Sun Oct 27			*P2 due before midnight*

16. Tue Oct 29	Retrieval Models		*Last day to drop with “DR”*
17. Thu Oct 31	Retrieval Models		
Tue Nov 5	No class		*Holiday (Election Day)*
18. Thu Nov 7	Retrieval Models		
Mon Nov 11			*Holiday (Veterans’ Day)*
19. Tue Nov 12	Retrieval Models		
20. Thu Nov 14	Retrieval Models		
21. Tue Nov 19	Retrieval Models		
22. Tue Nov 21	Queries/interfaces	Ch. 6	
23. Tue Nov 26	Queries/interfaces		*Break starts after last class*
Thu Nov 28	*No class; Thanksgiving break*		
24. Tue Dec 3	Other topics, deep learning, LLM	Ch. 9 & 10 & 7.6	
25. Thu Dec 5	LLM and wrap-up		
Sun Dec 8			*P3 due before midnight*
26. Tue Dec 10	Last class		FYI: Last day of all classes
Weds Dec 11			FYI: Reading day
Thu Dec 12			*In-person final exam 3:30-5:30, Thompson Hall 104.*

Communication Policy

The official means of communication for this class will be in-class announcements and posts by the professor to Piazza and/or Canvas. Communication should be via the Piazza forum for general questions of interest to the class. All other communication should be via email. In general, expect a response to email within 24 hours.

Incomplete Policy

An incomplete will be given only when documented, exceptional circumstances beyond your control have made it impossible to complete the assigned work before the end of the semester. It is your responsibility to contact the instructor regarding any such problems well before the end of the semester. Note that general rules of the University allow an incomplete only if most of the work has been completed satisfactorily before the end of the semester, so that the incomplete can be finished within the first four weeks of the immediately following semester. They further state that if a substantial amount of work remains undone then a retroactive drop should be obtained and the entire course repeated.

Auditing Policy

Official auditors will normally be expected to complete some amount of the course work to be sure that they are following the material (education by osmosis rarely works). Anyone enrolled for audit should contact the instructor early in the semester to discuss the requirements for receiving audit credit for this course. If the course is heavily enrolled, audits may not be possible.

Academic Honesty Policy

General principles. Since the integrity of the academic enterprise of any institution of higher education requires honesty in scholarship and research, academic honesty is required of all students at the University of Massachusetts Amherst. Academic dishonesty is prohibited in all programs of the University. Academic dishonesty includes but is not limited to: cheating, fabrication, plagiarism, and *facilitating* dishonesty. Appropriate sanctions may be imposed on any student who has committed an act of academic dishonesty. Instructors should take reasonable steps to address academic misconduct. Any person who has reason to believe that a student has committed academic dishonesty should bring such information to the attention of the appropriate course instructor as soon as possible. Instances of academic dishonesty not related to a specific course should be brought to the attention of the appropriate department Head or Chair. Since students are expected to be familiar with this policy and the commonly accepted standards of academic integrity, ignorance of such standards is not normally sufficient evidence of lack of intent (http://www.umass.edu/dean_students/codeofconduct/acadhonesty/).

Specific summary for this class. Your work must be your own. For anything other than exams, you are welcome to discuss general issues with other students, but the answer, the writing, and the final result that you hand in must be your own effort. Discussing or sharing answers to specific problems is considered dishonest. If you have questions about what is honest, please ask! One suggestion is never to write down anything while you're talking with someone about class work since that will require you to come up with the result again on your own later. You are strongly encouraged to cite your sources if you received extraordinary help from any person or text (including the Web), other than lecture content or the textbook.

The one exception to that policy for this class is the in-class group exercises (X*) where you are *expected* to make a single submission that was created collaboratively by your group.

For any material you hand in (whether individually or in a group), you must appropriately indicate when you are using work of others. If you use verbatim or only slightly altered text, you must clearly indicate (quotation marks, indented text, etc.) that you are quoting another source and what that source is. If you refer to work done by others, even if you do not quote it, you should include a reference to the original source. It does not matter if that work was published or not: if it is work other than your own, you are obligated to make it clear that you are using that person's work. Plagiarism will not be tolerated in this class. Plagiarism is a type of cheating and will be treated accordingly. The campus writing program provides more information about plagiarism.

If you use a service such as ChatGPT to help you, you must indicate that you did so in your answers and make it clear how you used the service and what you did after looking at it. You should be aware that ChatGPT produces things that read well but that it is often wrong. You may not use ChatGPT or the like to answer exam questions. You should also be aware that ChatGPT will not be usable on the exams, so it is in your interest to *understand* material, not just how to prompt an LLM to provide one.

You may (but probably won't) be using copyright-protected software as part of the class. Federal law and license agreements between the University and various software producers prohibit copying this software for any purpose. Such activity will be regarded as a form of cheating and will be dealt with as such.

The penalty for cheating in this class is (1) a zero on that assignment, (2) a reduction of one letter grade in the class, (3) a final course grade of "F," (4) referral to the Academic Dishonesty Committee, or some combination of the above.

Attendance Policy

Attendance is expected though not monitored (with the partial exception of implicitly gathering your attendance when the in-class exercises happen), with excused absences as provided by https://www.umass.edu/registrar/students/policies-and-practices/class-absence-policy. Please notify the instructor prior to the excused absence if it will affect the class that day. (And remember: in-class exercises cannot be made up, even for an excused absence.)

Accommodation Statement

The University of Massachusetts Amherst is committed to providing an equal educational opportunity for all students. If you have a documented physical, psychological, or learning disability on file with Disability Services (DS), you may be eligible for reasonable academic accommodations to help you succeed in this course. If you have a documented disability that requires an accommodation, I will be informed by Disability Services and reach out to you to make appropriate arrangements. For further information, please visit Disability Services (https://www.umass.edu/disability/)

Inclusivity Statement

We celebrate the diversity in our community and actively seek to include and listen to voices that are often silenced in the computing world. We welcome all individuals regardless of age, background, citizenship, disability, sex, education, ethnicity, family status, gender, gender identity, geographical origin, language, military experience, political views, race, religion, sexual orientation, socioeconomic status, and work experience. Even if you do not see yourself in that list, we welcome you.

Names & Pronouns

Everyone has the right to be addressed by the name and pronouns that they use for themselves. You can indicate your preferred/chosen first name and pronouns on SPIRE, and they appear on class rosters. I will strive to address you with your chosen name and pronouns. Please provide gentle correction in the event an incorrect name or pronoun is used.

Learning Support

There are a range of resources on campus, including:

- UMass Libraries: https://www.library.umass.edu/
- Writing Center - http://www.umass.edu/writingcenter
- Learning Resource Center - http://www.umass.edu/lrc
- Assistive Technology Center - https://www.umass.edu/it/assistive
- Disability Services - https://www.umass.edu/disability/
- Student Success - https://www.umass.edu/studentsuccess/
- Center for Counseling and Psychological Health (CCPH) http://www.umass.edu/counseling
- English as a Second Language (ESL) Program - http://www.umass.edu/esl
- CMASS Success Coach Program - https://www.umass.edu/cmass/your-cmass
- Single Stop Resources - https://www.umass.edu/studentlife/single-stop

Title IX Statement

In accordance with Title IX of the Education Amendments of 1972 that prohibits gender-based discrimination in educational settings that receive federal funds, the University of Massachusetts Amherst is committed to providing a safe learning environment for all students, free from all forms of discrimination, including sexual assault, sexual harassment, domestic violence, dating violence, stalking, and retaliation. This includes interactions in person or online through digital platforms and social media. Title IX also protects against discrimination on the basis of pregnancy, childbirth, false pregnancy, miscarriage, abortion, or related conditions, including recovery. There are resources here on campus to support you. A summary of the available Title IX resources (confidential and non-confidential) can be found at https://www.umass.edu/titleix/resources. You do not need to make a formal report to access them. If you need immediate support, you are not alone. Free and confidential support is available 24 hours a day / 7 days a week / every day of the year at the SASA Hotline 413-545-0800.

For purposes of Title IX reporting, I am a considered a “responsible employee” at UMass (https://www.umass.edu/titleix/about). That means that **if you tell me about a situation involving sexual assault, sexual harassment, domestic violence, dating violence, stalking, and retaliation, I must share that information with the Title IX Coordinator**. Making a report to the Title IX Coordinator is my legal obligation, meets the University's goal of providing members of our community with supportive resources they might need, and enables the University to obtain a more accurate picture of the extent of sexual violence in our community. It will be completely up to you to determine if and how you want to work with the Title IX Coordinator's office. You will not be in trouble for reporting to me that you have experienced any of these situations, and the law prohibits retaliation against anyone who participates in a Title IX process.

If you experience or witness sexual misconduct and wish to report the incident, please contact the UMass Amherst Equal Opportunity (EO) Office (413-545-3464 or by email at equalopportunity@admin.umass.edu) to request an intake meeting with EO staff. Members of the CICS community can also contact Erika Dawson Head, Executive Director of Diversity and Inclusion (erikahead@cics.umass.edu or 413-577-0338).