General information

COMPSCI 446 Search Engines
Fall 2023

Credit Hours: 3 credits
Prerequisites: CMPSCI 240, CMPSCI 383, or equivalent.

Instructor

- Ali Montazer
- montazer@cs.umass.edu

Class meetings:

- Tuesdays and Thursday 4:00-5:15pm
- Morrill Science Center (II) rm 131

Teaching staff:

- See Moodle for the list of staff

Office Hours:

- See Moodle for office hours

Other platforms used:

- Piazza for class discussion. Rather than emailing questions to the teaching staff, I encourage you to post your questions on Piazza – though please do not post personal or private information to the open forums!
- Gradescope for most graded activities.

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 Java, Python, or some other closely related language. This course is programming intensive.

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 be able 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 programming problems, such as computing PageRank, implementing a basic indexing and retrieval system, 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.

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. [cup] The authors of this text maintain a web site with information about the book, including a couple of on-line versions of the text.

List of Topics

The following topics will be covered in this course, in roughly the order listed, though the actual order and decisions to omit some topics will be determined in part by student interest and class discussion.

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

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
    - Searching within Communities

- Displaying the Results
  - Result Pages and Snippets
  - Advertising and Search
  - Clustering the Results
  - Translation
- Filtering and recommending

11. Deep Learning for IR

Grading Criteria

Your final grade in this class will be based upon in-class activities, homework, projects, a mid-term exam, and a final exam. The relative contributions of the parts are:

- Knowledge check-ins 15%
    - Homework-based knowledge check-ins (KH1 to KH7)
- Programming assignments (P0, P1 to P3, and PX): 45%
- Midterm exam (MX): 20%
- Final exam (FX): 20%

Knowledge check-ins are lower-stakes on-line that help you and the instructor have a sense of whether you are absorbing the material in the textbooks, the projects, and the lectures. There are two broad types of knowledge check-ins:

- KH* are homework assignments that become available at some point and must be completed by a specified deadline. They are designed so that they should take you no more than an hour, though that will vary depending on your own preparation, the instructor's ability to correctly calibrate the difficulty of the questions, and so on. Each is worth 2% of your overall class grade. There are seven, so they total 14% of your overall class grade.

Programming assignments are substantial coding projects that touch on aspects of search engine technology. At a high level:

- P0 is an infrastructure setup, ensuring that you can submit programming assignments. It is worth no points, but failure to complete it may result in losing points on future assignments if P0 would have caught them.
- P1 explores text processing and text statistics (15%)
- P2 develops evaluation tools for search engines (10%)
- P3 asks you to build an index for a search engine and run queries using several different models (20%)
- PX is an extra credit programming assignment, involving the possibility of implementing PageRank, some extra evaluation tools, alternative retrieval models, and/or some text processing work. The possible choices will be announced when PX becomes available (see the course schedule). PX will be due late in the semester.

Unless announced otherwise, extra credit will accrue to the major category and will not affect other categories.

Examinations: Both the midterm and final exams will have two components. One part will be taken in a classroom setting – probably in the evening for the midterm and during the normal final exam time for the final. The remainder of each exam is offered in a several-day day window on either Gradescope or Moodle; you choose the time block in which to take the exams. The schedule shows the tentative expected dates for those

Make-up or rescheduling policies: All assignments are due as indicated on Moodle (and in Gradescope). All dates are posted on Moodle at the start of the semester; they may shift but will only be moved to a later date if that happens. Late assignments will only be accepted as described with the assignment or in accordance with University policy as described in the academic regulations (largely starting with Section VII), at the sole discretion of the instructor. Please notify the instructor as soon as you know that there is an issue with a due date or a concern with the timing of an exam.

Points-to-letter-grade. The following table reflects the intended score-to-grade conversion for this course. Curving will happen if needed.

Highest Percentage	**Lowest Percentage**	**Letter**
100.00%	93.00%	A
92.99%	90.00%	A-
89.99%	87.00%	B+
86.99%	83.00%	B
82.99%	80.00%	B-
79.99%	77.00%	C+
76.99%	73.00%	C
72.99%	70.00%	C-
69.99%	67.00%	D+
66.99%	65.00%	D
64.99%	0.00%	F

Expectations and Requirements: No additional requirements.

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 the class Moodle.

Class Number and Date	**Topic/session details**	**Read to prepare**	**Quick reminders / deadlines**
1. Tue Sep 5	Introduction	Ch. 1	
2. Thu Sep 7	Processing Text	Ch. 4	
3. Tue Sep 12	Processing Text	Ch. 4	*KH1 due 4:00am*
4. Thu Sep 14	Processing Text: PageRank	§4.5.2	
Fri, Sep 15			*P0 due 11pm*
5. Tue Sep 19	Processing Text		*KH2 due 4:00pm*
6. Thu Sep 21	Collecting Documents	Ch. 3	
7. Tue Sep 26	Collecting Documents		*KH3 due 4:00pm*
8. Thu Sep 28	~~PageRank~~: Collecting Documents		
Mon Oct 2			
9. Tue Oct 3	Evaluation	Ch. 8	*KH4 due 4:00pm*
10. Thu Oct 5	Evaluation		*PX can be submitted now*
Oct 9			*P1 due 11pm*

Oct 10	***No class***		*Monday Schedule*
11. Thu Oct 12	Indexing	Ch. 5	
12. Tue Oct 17	Indexing		
13. Thu Oct 19	Indexing & Review		*In-person midterm, 7-9pm*
14. Tue Oct 24	Indexing		
15 Thu Oct 26	Indexing		*KH5 due 4:00pm*
16 Tue Oct 31	Retrieval models	Ch. 7	
17 Thu Nov 2	Retrieval Models		
18 Tue Nov 7	Retrieval Models		
19 Thu Nov 9	Retrieval Models		
Fri Nov 10			*P2 due 11:59pm*
20 Tue Nov 14	Retrieval Models		*KH6 due 4:00pm*
21 Thu Nov 16	Retrieval Models		
22 Tue Nov 21	Queries/interfaces	Ch. 6	
23 Tue Nov 28	Queries/interfaces		

24 Thu Nov 30	Other topics, Deep Learning	Ch. 9 & 10; Review Ch. 7.6	
25 Tue Dec 5	Deep Learning and Wrap Up		
26 Thu Dec 7	Last class		*KH7 due 4:00pm*
Fri Dec 8			*Last day of other classes*
Fri Dec 8			*P3 due 11:59pm*
Fri Dec 8			*Last day to submit PX*
Sat Dec 9			*Reading day*
Fri Dec 15, 3:30-5:30pm			*In-person final exam*

Communication Policy

The official means of communication for this class will be in-class announcements and posts by the professor to Piazza and Moodle. 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.

For any material you hand in, 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 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 exception of implicitly because of the in-class knowledge check-in exercises), 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 knowledge check-ins 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, please let me known within the first two weeks of the semester so that we may make appropriate arrangements. For further information, please visit Disability Services (https://www.umass.edu/disability/). Because of their in-class, low-stakes qualities, it is not possible to provide additional time for in-class knowledge check-ins.

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

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 let me know what name and pronouns I should use for you if they are not on the roster. Given that there are over 100 students in the class, it will be hard to keep track of everyone, please provide gentle correction if I make a mistake.

Title IX Statement

UMass is committed to fostering a safe learning environment by responding promptly and effectively to complaints of all kinds of sexual misconduct. If you have been the victim of sexual violence, gender discrimination, or sexual harassment, the university can provide you with a variety of support resources and accommodations.

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

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/get-involved/success/academic-support
- Single Stop Resources - https://www.umass.edu/studentlife/single-stop