INLS 509-001 Information Retrieval

School of Information and Library Science
University of North Carolina at Chapel Hill

Fall 2024
(Last update: October 23, 2024)

Course Information

Time:	Tue & Thu 11:00 am – 12:15 pm
Room:	Manning Hall 001
Instruction mode:	In person, on campus
Instructor:	Yue "Ray" Wang
Office:	Manning Hall 7B (the "Garden Level")
Office hours:	By appointment
	If you choose to discuss over Zoom, Please send me an email beforehand.
Email:	wangyue AT unc DOT edu

Along with the explosive growth of online textual information (e.g., Web pages, social media, news articles, emails, and scientific literature), it is increasingly important to develop tools to help users access, manage, and use the huge amount of information. Web search engines, such as Google and Bing, are good examples of such tools, and they are now an essential part of everyone's life. In this course, you will learn the underlying technologies of these and other powerful tools for connecting people with information, for accessing and mining unstructured information, especially text. You will be able to learn the basic principles and algorithms for information retrieval as well as obtain hands-on experience in designing your own search engines and improving their performance.

Unlike *structured data*, which is typically managed with relational database systems, textual information is unstructured and poses special challenges due to the difficulty in precisely understanding natural language and users' information needs. In this course, we will introduce a variety of techniques for accessing and mining text information and methods for evaluating these techniques. Topics to be covered include, among others, *text processing, inverted index, retrieval models (e.g., vector space models and language models), IR evaluation, Web search engines, information filtering/recommender systems, and applications of text information retrieval.*

This course is designed for graduate and advanced undergraduate students in the School of Information and Library Science. The course is lecture based.

Prerequisites: There are no prerequisites for this course. It would be a plus if you have some basic familiarity with linear algebra, probability, statistics, and programming. We will cover the mathematical essentials in the class, and will emphasize on concepts and rather than technical details.

Learning Objectives

Throughout the course, students will gain understanding and appreciation of the fundamental concepts and a broad range of topics in the field of information retrieval. In particular, students will:

- Understand how search engines work;
- Understand the limits of existing search technology;
- Understand the distinctive nature of text data;
- Learn about text similarity measures;
- Learn about classical relevance retrieval models;
- Learn to evaluate information retrieval systems;
- Appreciate the role of feedback in information retrieval;
- Appreciate the complexity of relevance in different search scenarios;
- Learn about modern Web search engine technologies;
- Learn about how text classification works and its applications;
- Understand the underlying mechanisms of recommender systems;
- Learn about the state of the art in IR research and applications.

References

- ***Required***: [CMS] W. Bruce Croft, Donald Metzler, and Trevor Strohman. *Search Engines - Information Retrieval in Practice*, Cambridge University Press, 2015. [Available online]
- ***Required***: [ZM] ChengXiang Zhai and Sean Massung. *Text Data Management: A Practical Introduction to Information Retrieval and Text Mining*. ACM and Morgan & Claypool Publishers, 2016. [Free access with UNC ONYEN login].
- Additional resource: Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. *Introduction to Information Retrieval*, Cambridge University Press, 2008. [Available online]
- Papers and chapters from other books will be assigned for reading. They will be available online.

Coursework

There will be a small number of homework assignments, a midterm exam, a final exam. 2-4 students will work together to explore their interests in a semester-long literature review project (details to be announced).

Grade breakdown:

- Class participation: 10%
- Homework assignments: 30%
- Midterm exam: 15%
- Final exam: 15%
- Literature review: 30%
  - Proposal: 5%
  - Progress update: 5%
  - Final presentation: 10%
  - Literature review paper: 10%

Undergraduate grading scale: A 95-100%, A- 90-94%, B+ 87-89%, B 84-86%, B- 80-83%, C+ 77-79%, C 74-76%, C- 70-73%, D+ 67-69%, D 64-66%, D- 60-63%, F 0-59%

Graduate grading scale: H 95-100%, P 80-94%, L 60-79%, and F 0-59%. All assignments, exams, and the literature review will be graded on a curve.

Sample Assignments

- Text data processing and understanding;
- Search evaluation metric calculation;
- Basic search system design and evaluation;
- Web graph link analysis.

Tentative Schedule

The following schedule is subject to change. The purpose of reading materials for each class (if any) is to provide a preview and reference for that class. Students are expected to read the materials before coming to the class. Certain content in class may not be found in the reading materials. As noted in the *References*: [CMS] refers to the textbook by Croft, Metzler, and Strohman: *Search Engines - Information Retrieval in Practice*; [ZM] refers to the textbook by Zhai and Massung: *Text Data Management: A Practical Introduction to Information Retrieval and Text Mining*.

1. Tuesday, Aug. 20: **Introduction to Information Retrieval**
   - A general overview of information retrieval; course structure and administration.
   - Reading:
     - Vannevar Bush. *As We May Think*, The Atlantic Monthly, 1945. [Available online]
2. Thursday, Aug. 22: **Introduction to IR; Mathematical Basics**
   - Basic concepts in linear algebra, probabilities, and statistics that will be discussed.

- Reading:
  - [ZM] Chapter 2, Background

3. Tuesday, Aug. 27: **Mathematical Basics (continued)**
   - **HW0 out**
   - Continue the topic and reading from previous lecture.
4. Thursday, Aug. 29: **Mathematical Basics (continued)**
   - Continue the topic and reading from previous lecture.
5. Tuesday, Sep. 3: **Well-Being Day (no class)**
6. Thursday, Sep. 5: **Text Processing and Analysis I**
   - Document representation, term selection, statistical properties of text.
   - Reading: [CMS] Chapter 4, Processing Text: 4.1 From Words to Terms, 4.2 Text Statistics
   - **HW0 due**
   - **HW1 out**
7. Tuesday, Sep. 10: **Text Processing and Analysis I (continued)**
   - Continue the topic and reading from previous lecture.
8. Thursday, Sep. 12: **Text Processing and Analysis II**
   - Natural language processing and its applications in information retrieval.
   - Reading:
     - [ZM] Chapter 3, Text Data Understanding
     - Kenneth Church, Patrick Hanks. *Word Association Norms, Mutual Information, and Lexicography*. Computational Linguistics 1990. [Available online]
9. Tuesday, Sep. 17: **Text Processing and Analysis II (continued)**
   - Continue the topic and reading from previous lecture.
10. Thursday, Sep. 19: **Text Retrieval Systems I**
    - Text retrieval system architecture; document selection vs. document ranking; evaluation metrics.
    - Reading: [ZM] Chapter 5, Text Data Access; Chapter 9, Search Engine Evaluation (9.2, 9.3)
    - **HW1 due**
11. Tuesday, Sep. 24: **Text Retrieval Systems I (continued)**
    - Continue the topic and reading from previous lecture.
12. Thursday, Sep. 26: **Solr Lab Session I**
    - Apache Solr installation and usage walkthrough; preparation for HW2.
    - Sep. 30: **Literature review proposal due**

13. Tuesday, Oct. 1: **Text Retrieval Systems II**
    - Inverted index; boolean queries and boolean retrieval.
    - Reading: [CMS] 5.3 Inverted Indexes; 7.1 Overview of Retrieval Models; 7.1.1 Boolean Retrieval
    - **HW2 out**
14. Thursday, Oct. 3: **Text Retrieval Systems II (continued)**
    - Continue the topic and reading from previous lecture.
15. Tuesday, Oct. 8: **Retrieval Models: Vector Space Models I**
    - Motivation behind vector space models; basic TFIDF term weighting.
    - Reading: [CMS] 7.1.2 The Vector Space Model; [ZM] 6.3 Vector Space Retrieval Models
16. Thursday, Oct. 10: **Solr Lab Session II**
    - Test collection construction and evaluation.
17. Tuesday, Oct. 15: **Retrieval Models: Vector Space Models II**
    - Vector space models; axiomatic approach to retrieval model design.
    - Reading on axiomatic approach:
      - Fang Hui, Tao Tao, Chengxiang Zhai. *A Formal Study of Information Retrieval Heuristics*, SIGIR 2004. [Available online]
    - Oct. 15-16: **Take-Home Midterm**
18. Thursday, Oct. 17: **Fall Break (no class)**
19. Tuesday, Oct. 22: **Group Activity Day**
    - **Literature review update talks**
    - Solr Lab: in-class peer evaluation
20. Thursday, Oct. 24: **Retrieval Models: Query Likelihood Models I**
    - Probabilistic models of language and relevance; query likelihood models.
    - Reading: [ZM] 6.4 Probabilistic Retrieval Models
    - **HW2 due**
21. Tuesday, Oct. 29: **Retrieval Models: Query Likelihood Models II**
    - Continue the topic and reading from previous lecture.
    - **HW3 out**
22. Thursday, Oct. 31: **Retrieval Models: Query Likelihood Models III**
    - Continue the topic and reading from previous lecture.
23. Tuesday, Nov. 5: **Web Search Engines: Web Models and Link Analysis**

- Models of the Web; Web crawling and indexing; static ranking and link analysis
- Reading:
  - [ZM] 10.3, Link Analysis
  - (Optional) Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd. *The PageRank Citation Ranking: Bringing Order to the Web*, Technical Report, Stanford InfoLab, 1998. [Available online]

24. Thursday, Nov. 7: **Web Models and Link Analysis (continued); Search Log Analysis and Mining**

    - Position bias, log-based evaluation, behavioral log mining; privacy and ethics in search log analysis.
    - Reading:
      - Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Geri Gay. *Accurately interpreting clickthrough data as implicit feedback*, SIGIR'05. [Available online]

25. Tuesday, Nov. 12: **Search Log Analysis and Mining**

    - Continue the topic and reading from previous lecture.
    - **HW3 due**

26. Thursday, Nov. 14: **Query Expansion and Relevance Feedback**

    - Query reformulation concepts; relevance feedback & pseudo-relevance feedback
    - Reading:
      - [CMS] 6.1 Information Needs and Queries; 6.2 Query Transformation and Refinement
      - [ZM] Chapter 7, Feedback

27. Tuesday, Nov. 19: **Student Presentations I**

28. Thursday, Nov. 21: **Information Filtering and Recommender Systems**

    - Content-based recommender systems;
    - User and item matrices; collaborative filtering; applications of recommender systems
    - Reading: [ZM] 11.1 Content-based Recommendation; 11.2 Collaborative Filtering
    - User and item matrices; collaborative filtering; applications of recommender systems

29. Tuesday, Nov. 26: **Student Presentations II**

30. Thursday, Nov. 28: **Thanksgiving Break (no class)**

31. Tuesday, Dec. 3: **Student Presentations III**

    - **Literature review paper due on Friday, Dec 6**. Early submission is strongly recommended.

32. Monday, Dec 9: **Take-Home Final Exam**

    - Open-book, open-notes, open-slides
    - Will be released in Canvas as an Assignment.

Course policies

Late Submission Policy:

The student should submit her/his homework solution to the Canvas site by 11:59 pm EST of the announced due date. Each late day will result in 10% reduction of the homework grade. If a homework is late for more than 5 calendar days, the grade of that homework will be zero. In case there is an emergency before the submission deadline, please inform the instructor as early as possible.

Collaboration:

You are encouraged to learn from each other. However, all the work you hand in must be your own. This means that you cannot look at another student's answer and copy or re-word it as your own. Your work is a part of you; do not let someone else represent you. If someone helps you with a homework assignment, please give them credit by writing their name on the top of your homework. This will not hurt you (provided your answer is your own), but it will help them.

If you are the student giving help, don't give away the answer. Rather, try to help the student arrive at the answer themselves. If you are the student asking for help, don't ask for the answer. Rather, ask about the material. Your own answer must come from your own intuition. You must fully understand what you write and be able to explain your answer to the instructor.

Class Participation:

As a student, you are expected to attend every class throughout the semester. In each class, you are expected to ask questions, express opinions, and actively participate in discussions. Sharing your view with your peers is an important part of your education. It will sharpen your understanding of the material and help you build confidence in the area of study. Class participation will be 10% of your final grade.

During the semester, missing one or two classes due to legitimate reasons (e.g., sickness) is fine. However, if you expect to miss more than twice during the semester, please notify the instructor prior to the missing class. Your attendance factors into your participation grade. If you have to miss a class, make sure to get lecture notes from one of your peers. In-class discussions are excellent source of exam questions.

University's Attendance Policy: As stated in the University's Class Attendance Policy, no right or privilege exists that permits a student to be absent from any class meetings, except for these University Approved Absences:

1. Authorized University activities: University Approved Absence Office (UAAO) website provides information and FAQs for students and FAQs for faculty related to University Approved Absences;
2. Disability/religious observance/pregnancy, as required by law and approved by the Equal Opportunity and Compliance Office (EOC);
3. Significant health condition and/or personal/family emergency as approved by the Office of the Dean of Students, Gender Violence Service Coordinators, and/or the Equal Opportunity and Compliance Office (EOC).

Syllabus Changes

The instructor reserves the right to make changes to the syllabus including project due dates and test dates. These changes will be announced as early as possible.

Acceptable Use Policy

By attending the University of North Carolina at Chapel Hill, you agree to abide by the University of North Carolina at Chapel Hill policies related to the acceptable use of IT systems and services. The Acceptable Use Policy (AUP) sets the expectation that you will use the University's technology resources responsibly, consistent with the University's mission. In the context of a class, it's quite likely you will participate in online activities that could include personal information about you or your peers, and the AUP addresses your obligations to protect the privacy of class participants. In addition, the AUP addresses matters of others' intellectual property, including copyright. These are only a couple of typical examples, so you should consult the full Information Technology Acceptable Use Policy, which covers topics related to using digital resources, such as privacy, confidentiality and intellectual property.

Additionally, consult the Safe Computing at UNC website for information about data security policies, updates, and tips on keeping your identity, information, and devices safe.

Code of Conduct

The University of North Carolina at Chapel Hill has a student-led honor system (the UNC Honor Code). We are all responsible for upholding the ideals of honor and academic integrity. The student-led honor system is responsible for adjudicating any suspected violations of the Honor Code and all suspected instances of academic dishonesty will be reported to the honor system. Information, including your responsibilities as a student is outlined in the Instrument of Student Judicial Governance. Your full participation and observance of the Honor Code is expected.

All written submissions must be your own, original work. Original work for narrative questions is not mere paraphrasing of someone else's completed answer: you must not share written answers with each other at all. At most, you should be working from notes you took while participating in a study session. Largely duplicate copies of the same assignment will receive an equal division of the total point score from the one piece of work.

You may incorporate selected excerpts, statements or phrases from publications by other authors, but they must be clearly marked as quotations and must be attributed. If you build on the ideas of prior authors, you must cite their work. You may obtain copy editing assistance, and you may discuss your ideas with others, but all substantive writing and ideas must be your own, or be explicitly attributed to another.

Equal Opportunity and Compliance - Accommodations

Equal Opportunity and Compliance Accommodations Team (Accommodations – UNC Equal Opportunity and Compliance) receives requests for accommodations for disability, pregnancy and related conditions, and sincerely held religious beliefs and practices through the University's Policy on Accommodations. EOC Accommodations team determines eligibility and reasonable accommodations consistent with state and federal laws.

Counseling and Psychological Services

UNC-Chapel Hill is strongly committed to addressing the mental health needs of a diverse student body. The Heels Care Network website is a place to access the many mental health resources at Carolina. CAPS

is the primary mental health provider for students, offering timely access to consultation and connection to clinically appropriate services. Go to the CAPS website or visit their facilities on the third floor of the Campus Health building for an initial evaluation to learn more. Students can also call CAPS 24/7 at 919-966-3658 for immediate assistance.

Title IX Resources

Any student who is impacted by discrimination, harassment, interpersonal (relationship) violence, sexual violence, sexual exploitation, or stalking is encouraged to seek resources on campus or in the community. Reports can be made online to the EOC or by contacting the University's Title IX Coordinator (titleixcoordinator@unc.edu), Elizabeth Hall, or the Report and Response Coordinators (reportandresponse@unc.edu) in the Equal Opportunity and Compliance Office. Please note that I am designated as a Responsible Employee, which means I must report to the EOC any information I receive about the forms of misconduct listed in this paragraph. If you'd like to speak with a confidential resource, those include Counseling and Psychological Services and the Gender Violence Services Coordinators (gvsc@unc.edu). Additional resources are available at safe.unc.edu.

Policy on Non-Discrimination

The University is committed to providing an inclusive and welcoming environment for all members of our community and to ensuring that educational and employment decisions are based on individuals' abilities and qualifications. Consistent with this principle and applicable laws, the University's Policy Statement on Non-Discrimination offers access to its educational programs and activities as well as employment terms and conditions without respect to race, color, gender, national origin, age, religion, genetic information, disability, veteran's status, sexual orientation, gender identity or gender expression. Such a policy ensures that only relevant factors are considered, and that equitable and consistent standards of conduct and performance are applied.

If you are experiencing harassment or discrimination, you can seek assistance and file a report through the Report and Response Coordinators (email reportandresponse@unc.edu or see additional contact info at safe.unc.edu) or the Equal Opportunity and Compliance Office. Please note that I am designated as a Responsible Employee, which means that I must report to the EOC any information I receive about harassment or discrimination. If you'd like to speak with a confidential resource, those include Counseling and Psychological Services and the University's Ombuds Office.

Diversity, Equity, Inclusion Statement

I value the perspectives of individuals from all backgrounds reflecting the diversity of our students. I broadly define diversity to include race, gender identity, national origin, ethnicity, religion, social class, age, sexual orientation, political background, and physical and learning ability. I strive to make this classroom an inclusive space for all students. Please let me know if there is anything I can do to improve. I appreciate any suggestions.

Learning Center

Want to get the most out of this course or others this semester? Visit UNC's Learning Center to make an appointment or register for an event. Their free, popular programs will help you optimize your academic performance. Try academic coaching, peer tutoring, STEM support, ADHD/LD services, workshops and study camps, or review tips and tools available on the website.

Writing Center

For free feedback on any course writing projects, check out UNC's Writing Center. Writing Center coaches can assist with any writing project, including multimedia projects and application essays, at any stage of the writing process. You don't even need a draft to come visit. To schedule a 45-minute appointment, review quick tips, or request written feedback online, visit UNC's Writing Center online.

Recording:

Please attend in-person class sessions. Please do not record the lectures in audio or video form, or share the recording on the Internet without explicit permission of the instructor.

Usage of ChatGPT and Generative AI:

ChatGPT and other generative Artificial Intelligence (AI) technologies can produce text, images, and other media. These tools can assist with brainstorming, finding information, and even reading and creating materials; however, they must be used appropriately and ethically, and you must understand their limitations. Regardless of your use of any AI tools, ***you*** are ultimately responsible for the final product of your work.

Generative AI is extremely useful; however, it has many limitations. The limitations include, but not limited to: (1) AI-generated outputs may be inaccurate or entirely fabricated even if they appear reliable or factual; (2) the sourcing and ownership of the outputs are unclear, raising ethical and intellectual property concerns; (3) the outputs are based on existing data (often scraped from online sources) and may reflect biased views and values inherited from those data.

If you decide to use generative AI in your coursework, I urge you to use it responsibly. You should critically evaluate AI-generated outputs and consider potential biases, limitations, and ethical implications when using these outputs. If you don't know whether a statement about *any item* in the output is true, then your responsibility is to research it. If you cannot verify it as factual, you should delete it. You hold full responsibility for AI-generated content as if you had produced the materials yourself. You should clearly document and report the usage of AI-generated outputs in detail if you submit material that contains them, is based on them, or is derived from them. You should use generative AI to help you think, not think for you.

Pertaining to the subject of this course, I strongly encourage you to research the mechanisms, applications, and implications of generative AI in current and future information retrieval systems.