INLS 509-001 Information Retrieval

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

Fall 2023
(Last update: September 6, 2023)

Course Information

Time:	Mon & Wed 9:05 am – 10:20 am
Room:	Manning Hall 208
Instruction mode:	In person, on campus
Instructor:	Yue "Ray" Wang
Office:	Manning Hall 7B (the "Garden Level")
Office hours:	Wed 11:00 am – 12:00 pm or 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 and probabilistic 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%
  - Final presentation: 15%
  - 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. Monday, Aug. 21: **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. Wednesday, Aug. 23: **Introduction to IR; Mathematical Basics**
   - Basic concepts in linear algebra, probabilities, and statistics that will be discussed.
   - Reading:

  - [ZM] Chapter 2, Background

3. Monday, Aug. 28: **Mathematical Basics (continued)**
   - Continue the topic and reading from previous lecture.
4. Wednesday, Aug. 30: **Class canceled (in memory of Prof. Zijie Yan)**
5. Monday, Sep. 4: **Labor Day (no class)**
6. Wednesday, Sep. 6: **Mathematical Basics (continued)**
   - Continue the topic and reading from previous lecture.
7. Monday, Sep. 11: **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
8. Wednesday, Sep. 13: **Text Processing and Analysis I (continued)**
   - Continue the topic and reading from previous lecture.
   - **HW1 out**
9. Monday, Sep. 18: **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]
10. Wednesday, Sep. 20: **Text Processing and Analysis II (continued)**
    - Continue the topic and reading from previous lecture.
11. Monday, Sep. 25: **Well-Being Day (no class)**
12. Wednesday, Sep. 27: **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)
13. Monday, Oct. 2: **Text Retrieval Systems I (continued)**
    - Continue the topic and reading from previous lecture.
    - **HW1 due**
14. Wednesday, Oct. 4: **Solr Lab Session I**
    - Apache Solr installation and usage walkthrough; preparation for HW2.
    - **HW2 out**

15. Monday, Oct. 9: **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
    - Oct. 10: **Literature review proposal due**
16. Wednesday, Oct. 11: **Text Retrieval Systems II (continued)**
    - Continue the topic and reading from previous lecture.
17. Monday, Oct. 16: **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
    - Oct. 17: **Take-Home Midterm**
18. Wednesday, Oct. 18: **Solr Lab Session II**
    - Test collection construction and evaluation.
19. Monday, Oct. 23: **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]
20. Wednesday, Oct. 25: **Retrieval Models: Query Likelihood Models I**
    - Probabilistic models of language and relevance; query likelihood models.
    - Reading: [ZM] 6.4 Probabilistic Retrieval Models
21. Monday, Oct. 30: **Retrieval Models: Query Likelihood Models II**
    - Continue the topic and reading from previous lecture.
22. Wednesday, Nov. 1: **Query Expansion and Relevance Feedback**
    - **Solr Lab: Test Collection Showcase**
    - 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
23. Monday, Nov. 6: 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]
- **HW2 due**
- **HW3 out**

24. Wednesday, Nov. 8: **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. Monday, Nov. 13: **Search Log Analysis and Mining**
    - Continue the topic and reading from previous lecture.
26. Wednesday, Nov. 15: **Adaptive Information Filtering**
    - Content-based recommender systems
    - Reading: [ZM] 11.1 Content-based Recommendation
27. Monday, Nov. 20: **Collaborative Filtering and Recommender Systems**
    - User and item matrices; collaborative filtering; applications of recommender systems
    - Reading:
      - [ZM] 11.2 Collaborative Filtering
    - **HW3 due**
28. Wednesday, Nov. 22: **Thanksgiving Break (no class)**
29. Monday, Nov. 27: **Collaborative Filtering and Recommender Systems II**
    - Reflections and critiques on recommender systems
    - Reading:
      - Dietmar Jannach, Paul Resnick, Alexander Tuzhilin, Markus Zanker. *Recommender Systems: Beyond Matrix Completion*. Communications of ACM'16. [Available online]
30. Wednesday, Nov. 29: **Student Presentations I**
31. Monday, Dec. 4: **Student Presentations II**
32. Wednesday, Dec. 6: **Student Presentations III**
    - **Literature review paper due on Friday, Dec. 8**. Early submission is strongly recommended.
33. Monday, Dec. 12: **Take-Home Final Exam**
    - Open-book, open-notes, open-slides
    - Will be released in Canvas as an Assignment.

Course policies

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 your suggestions!

Mask Use (In-Person Instruction Modes)

Masks continue to be encouraged yet optional in all University buildings, including classrooms. If you choose to wear a mask, we recommend choosing one that is comfortable and fits well. There are many reasons why a person may decide to continue to wear a mask, and let us respect that choice. Conversely, let us respect the choice of students who choose not to wear a mask in classrooms where it is now optional.

Attendance:

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 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;
2. Disability/religious observance/pregnancy, as required by law and approved by Accessibility Resources and Service and/or 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).

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.

Late Policy:

The student should submit her/his homework solution to the Sakai site by 11:55pm 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.

Honor Code:

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.

Students with Disabilities:

The University of North Carolina at Chapel Hill facilitates the implementation of reasonable accommodations, including resources and services, for students with disabilities, chronic medical conditions, a temporary disability or pregnancy complications resulting in difficulties with accessing learning opportunities. All accommodations are coordinated through the Accessibility Resources and Service Office. See the ARS Website for contact information.

Relevant policy documents as they relation to registration and accommodations determinations and the student registration form are available on the ARS website under the About ARS tab .

Counseling and Psychological Services (CAPS):

CAPS is strongly committed to addressing the mental health needs of a diverse student body through timely access to consultation and connection to clinically appropriate services, whether for short or long-term needs. Go to their website: https://caps.unc.edu/ or visit their facilities on the third floor of the Campus Health Services building for a walk-in evaluation to learn more.

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 at https://eoc.unc.edu/report-an-incident/. Please contact the University's Title IX Coordinator (Elizabeth Hall, cehall@email.unc.edu), Report and Response Coordinators in the Equal Opportunity and Compliance Office (reportandresponse@unc.edu), Counseling and Psychological Services (confidential), or the Gender Violence Services Coordinators (gvsc@unc.edu; confidential) to discuss your specific needs. Additional resources are available at safe.unc.edu.

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 the context of information retrieval systems!