INLS 509-001: Information Retrieval

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

Fall 2018

Course Information

Time:	Tuesday & Thursday, 9:30-10:45am
Room:	Manning Hall 208
Instructor:	Yue “Ray” Wang
Office:	Manning Hall 7B (Garden Level)
Office hours:	Tuesday & Thursday 1-2:30pm, or by appointment via email
Email:	wangyue AT email DOT unc DOT edu

Along with the explosive growth of online textual information (e.g., Web pages, social media, news articles, email, 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 a relational database, 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, and applications of text information retrieval.*

This course is designed for graduate students and advanced undergraduate students of 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, and statistics. 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, 2009. [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 inside UNC campus network]

- 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 come).

Grade breakdown:

- Class participation: 10%
- Homework assignments: 30%
- Midterm exam: 15%

- Final exam: 15%
- Literature review: 30%
  - Proposal: 5%
  - Final presentation: 10%
  - Literature review paper: 15%

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 retrieval model design and analysis;
- Text classification concepts and applications.

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 marked 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. 21: **Introduction to Information Retrieval**
   - A general overview of information retrieval; course structure and administration.
2. Thursday, Aug. 23: **Introduction to IR; Mathematical Basics**
   - Basic concepts in linear algebra, probabilities, and statistics that will be discussed.
   - Reading:
     - Vannevar Bush. *As We May Think*, The Atlantic Monthly, 1945. [Available online]
     - [ZM] Chapter 2, Background
3. Tuesday, Aug. 28: **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

4. Thursday, Aug. 30: **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]

5. Tuesday, Sep. 4: **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 out**

6. Thursday, Sep. 6: **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

7. Tuesday, Sep. 11: **Retrieval Models: Vector Space Models I**
   - Motivation behind vector space models; TFIDF term weighting.
   - Reading: [CMS] 7.1.2 The Vector Space Model; [ZM] 6.3 Vector Space Retrieval Models

8. Thursday, Sep. 13: **(No class due to Hurricane Florence)**

9. Tuesday, Sep. 18: **Retrieval Models: Vector Space Models II**
   - Vector space models continued; 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]

10. Thursday, Sep. 20: **Probabilistic Models; Query Likelihood Models I**
    - Probabilistic models of language and relevance; query likelihood models.
    - Reading: [ZM] 6.4 Probabilistic Retrieval Models
    - **HW1 due**

11. Tuesday, Sep. 25: **Probabilistic Models; Query Likelihood Models II**
    - Continue the topic and reading from previous lecture.

12. Thursday, Sep. 27: **Virtual IR Lab registration; Query Expansion and Relevance Feedback I**
    - Query reformulation concepts

- Reading: [CMS] 6.1 Information Needs and Queries; 6.2 Query Transformation and Refinement
- **HW2 out**

13. Tuesday, Oct. 2: **Query Expansion and Relevance Feedback II**
    - Query reformulation concepts(continued); relevance feedback & pseudo-relevance feedback
    - Reading: [ZM] Chapter 7, Feedback
14. Thursday, Oct. 4: **Virtual IR Lab & Batch Evaluation**
    - Walking through the usage of Virtual IR Lab by examples; preparation for HW2
15. Tuesday, Oct. 8: **Evaluation Overview**
    - Overview of different evaluation methodologies
16. Thursday, Oct. 11: **Midterm Review**
17. Tuesday, Oct. 16: **Midterm Exam**
18. Thursday, Oct. 18: **Fall Break (no class)**
    - **HW2 due**
19. Tuesday, Oct. 23: **Batch Evaluation**
    - TREC-style evaluation
    - Reading: Reading: [CMS] 8.2, The Evaluation Corpus
20. Thursday, Oct. 25: **Revisiting Relevance; User-Centered Relevance & User Studies**
    - Revisiting the notion of relevance, behavior-based evaluation
    - The design of interactive IR studies
    - Reading:
      - Tefko Saracevic. *Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part III: Behavior and Effects of Relevance*, JASIST'07. [Available online].
      - Diane Kelly, *Methods for Evaluating Interactive Information Retrieval Systems with Users*, Chapter 10: *"Measures"*. Foundations and Trends in Information Retrieval, 2009. [Available online].
      - Anastasios Tombros, Ian Ruthven, Joemon Jose. *How Users Assess Web Pages for Information Seeking*, JASIST'05. [Available online]
21. Tuesday, Oct. 30: **Web Search Engines: Web Models and Link Analysis**
    - Models of the Web; Web crawling and indexing; static ranking and link analysis
    - Reading:
      - [ZM] 10.1 Web Crawling; 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]

22. Thursday, Nov. 1: **Search Log Analysis and Mining**
    - Position bias, log-based evaluation, and behavioral log mining.
    - Reading:
        - Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Geri Gay. *Accurately interpreting clickthrough data as implicit feedback*, SIGIR'05. [Available online]
23. Tuesday, Nov. 6: **(Class canceled due to water system breakdown)**
24. Thursday, Nov. 8: **Spam Filtering and Text Classification**
    - Deduplication and hashing; text classification basic concepts, k-nearest neighbors
    - Reading: [ZM] Chapter 15, Text Categorization
25. Tuesday, Nov. 13: **Text Classification II**
    - Naive Bayes models; active learning in text classification
    - Reading: Continue the reading from previous lecture.
    - **HW3 out**
26. Thursday, Nov. 15: **Adaptive Information Filtering**
    - Content-based Recommendation
    - Reading: [ZM] 11.1 Content-based Recommendation
27. Tuesday, Nov. 20: **Collaborative Filtering and Recommender Systems**
    - User and item matrices; collaborative filtering; applications of recommender systems
    - Reading:
        - [ZM] 11.2 Collaborative Filtering
        - Dietmar Jannach, Paul Resnick, Alexander Tuzhilin, Markus Zanker. *Recommender Systems: Beyond Matrix Completion*. Communications of ACM'16. [Available online]
28. Thursday, Nov. 22: **Thanksgiving (no class)**
29. Tuesday, Nov. 27: **IR Frontiers and Course Summary**
    - We will survey the frontiers of information retrieval technologies and applications, including learning to rank, aggregated search, social search, mobile search, professional search, knowledge graph ("things, not strings"), deep learning, question answering, conversational AI, and beyond.
    - **HW3 due** (Nov. 28)
30. Thursday, Nov. 29: **Student presentations I**
31. Tuesday, Dec. 4: **Student presentations II**
32. **Final exam**
    - Date & time: Tuesday, Dec. 11, 8-11 am
    - Room: Manning 208

Course policies

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., travel, 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.

Laptops and cellphones:

Usage of laptop computers, cellphones, and other electronic devices is **discouraged** during class. While laptops and tablets are convenient for note-taking, they are also a source of distraction. Please keep in mind that participation is 10% of your grade. If your laptop is open and your mind is elsewhere, it will show. As an etiquette, please mute your phone before class starts.

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 11:59pm 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 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 .

Recording:

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