CSC 4/583 Text Retrieval & Web Search

GS 906, Tuesday and Thursday 12:30PM - 1:45PM

Course Description

Most of the web data today consists of unstructured text. Of course, the fact that this data exists is irrelevant, unless it is made available such that users can quickly find information that is relevant for their needs. This course will cover the fundamental knowledge necessary to build these systems, such as web crawling, index construction and compression, Boolean, vector-based, and probabilistic retrieval models, text classification and clustering, link analysis algorithms such as PageRank, and computational advertising. The students will also complete one programming project, in which they will construct one complex application that combines multiple algorithms into a system that solves real-world problems.

Instructor and Contact Information

Instructor: Eduardo Blanco, GS 739, eduardoblanco@arizona.edu

Office hours: Tuesdays 2-4PM or by appointment

Teaching Assistant: harshita Narnoli, harshitanarnoli@arizona.edu

Office hours: Mondays 1-3PM or by appointment; reach out to her or the instructor for location

We will use the d2l platform to share the quiz materials and grades. We will use piazza for discussions (including lectures and homeworks).

Course Format and Teaching Methods

Lecture only. As a rule, lectures will be recorded and made available online to students. Students are expected to attend lectures in person. Recordings may not always be complete, and the quality of recordings may not always be ideal.

Course Objectives

Students will:

1. Learn multiple crawling and indexing algorithms.
2. Understand several search/retrieval methodologies.
3. Have the capacity to apply these algorithms at scale, i.e., on a web-scale search engine.

Expected Learning Outcomes

1. Students will be able to identify instances of information retrieval problems in the real world, e.g., web search or music retrieval, and examine information retrieval techniques, e.g., search methods, language models, text classification, link analysis, which apply to the problem at hand.
2. Students will be able to design and implement information retrieval systems that solve the above problems.
3. Students will be able to analyze the behavior of their implemented systems on the task addressed, and appraise the performance on real-world data.

The graduate students taking this course will have an additional set of learning outcomes, as follows:

1. Graduate students will be able to design and implement information retrieval methods driven by word embedding methods such as word2vec and transformer networks.
2. Graduate students will be able to analyze the behavior of these ML-driven methods on real-world information retrieval tasks, and compare these methods against traditional IR methods covered in the course.

Makeup Policy for Students Who Register Late

Students who register after the first class meeting may make up the missed assignments. The deadline for doing so will be extended by the numbers of days between the start of the semester and the student's registration.

Course Communications

Please use the email addresses above to contact the instructor or the TA. All course materials will be posted in D2L. Please use the Piazza site above to ask clarification questions about the material. As a general rule, you can expect an answer within 24 hours.

Required Texts or Readings

This course follows the following textbooks:

- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. 2008. Introduction to Information Retrieval. Cambridge University Press. Available for free at http://nlp.stanford.edu/IR-book (IIR)
- Mihai Surdeanu and Marco A. Valenzuela-Escárcega.2023. Deep Learning for Natural Language Processing: A Gentle Introduction. Available for free at https://clulab.org/gentlenlp.html (DLNLP)

Additional research articles covered in class will be distributed by the instructor.

Assignments and Examinations: Schedule/Due Dates

Grading will be based on four assignments (including written and programming portions), two exams (midterm and final exam), and a project.

Late work will not be accepted unless there is an excused absence as defined by the university. Reach out to the instructor if you think you qualify.

Here is the planned schedule for the class. The schedule may be adjusted based on instructor discretion.

Date	Topic	Reading	HW out	HW due	Quiz
1/16/2025	Logistics, Intro, boolean retrieval	IIR 01	HW1a		
1/21/2025	Boolean retrieval	IIR 01			
1/23/2025	The term vocabulary & postings lists	IIR 02		HW1a	
1/28/2025	The term vocabulary & postings lists	IIR 02	HW1b		
1/30/2025	Dictionaries and tolerant retrieval	IIR 03			Q1
2/4/2025	Dictionaries and tolerant retrieval	IIR 03			

2/6/2025	Index construction, Index compression	IIR 04, IIR 05			
2/11/2025	Index compression	IIR 05		HW1b	
2/13/2025	Scoring, term weighting & the vector space model	IIR 06	HW#2		
2/18/2025	Scoring, term weighting & the vector space model	IIR 06			Q2
2/20/2025	Computing scores in a complete search system	IIR 07			
2/25/2025	Computing scores in a complete search system	IIR 07			
2/27/2025	Evaluation in information retrieval	IIR 08		HW#2	Q3
3/4/2025	Evaluation in information retrieval	IIR 08			
3/6/2025	Midterm Exam				
3/11/2025	Spring Break				
3/13/2025	Spring Break				
3/18/2025	Projects intro, discussion		HW#3		
3/20/2025	Probabilistic information retrieval	IIR 11			
3/25/2025	Probabilistic information retrieval	IIR 11			Q4
3/27/2025	Language models for information retrieval	IIR 12			
4/1/2025	Language models for information retrieval	IIR 12		HW#3	
4/3/2025	Link analysis	IIR 21	HW#4		
4/8/2025	Prompt engineering	notes, papers			
4/10/2025	The perceptron	DLNLP 02			
4/15/2025	Logistic Regression	DLNLP 03			Q5
4/17/2025	Feed-forward neural networks	DLNLP 05		HW#4	
4/22/2025	Best practices in deep learning	DLNLP 06			
4/24/2025	Distributional hypothesis and representation learning	DLNLP 08			
4/29/2025	Contextualized embeddings and transformers	DLNLP 12			
5/1/2025	Encoder-Decoder methods	DLNLP 14			
5/6/2025	Dense passage retreival	notes, papers			
5/14/2025	Final Exam, 1-3pm				

Final Examination

The final exam will be on May 14, 2025, 1-3pm, as determined by the Registrar’s office: https://registrar.arizona.edu/faculty-staff-resources/room-class-scheduling/schedule-classes/final-exams

Grading Scale and Policies

The grading scheme is as follows:

Component	**Weight**
Homeworks	350
Quizzes	100
Project	200
Midterm exam	150
Final Exam	200
Total	**1000**

Point grades will be mapped to letter grades as follows:

Grade	**Points**
A	900-1000
B	800-899
C	700-799
D	600-699
E	0-599

Undergraduate vs. Graduate Requirements:

This course will be co-convened. To differentiate between undergraduate and graduate students, the instructor will require graduate students to implement more complex algorithms in homework assignments and the project. This may require additional reading of research articles. The instructor will provide the additional reading material and will guide the research process. Similarly, the midterm and final exams will have additional questions for graduate students. The overall grading scheme will be the same between graduate and undergraduate students (see the two tables above).

Incomplete (I) or Withdrawal (W):

Requests for incomplete (I) or withdrawal (W) must be made in accordance with University policies, which are available at https://catalog.arizona.edu/policy/courses-credit/grading/grading-system.

Honors Credit

Students wishing to contract this course for Honors Credit should e-mail the instructor to set up an appointment to discuss the terms of the contact and to sign the Honors Course Contract

Request Form. The form is available at http://www.honors.arizona.edu/honors-contracts. The expectations will be similar to the work required for graduate students, including a project that makes substantial research contributions.

Scheduled Topic and Activities

Please see the schedule under "Assignments and Examinations: Schedule / Due Dates". Note that the schedule may be adjusted based on instructor discretion.

Classroom Behavior Policy

To foster a positive learning environment, students and instructors have a shared responsibility. We want a safe, welcoming, and inclusive environment where all of us feel comfortable with each other and where we can challenge ourselves to succeed. To that end, our focus is on the tasks at hand and not on extraneous activities (e.g., texting, chatting, reading a newspaper, making phone calls, web surfing, etc.).

Students are asked to refrain from disruptive conversations with people sitting around them during lecture. Students observed engaging in disruptive activity will be asked to cease this behavior. Those who continue to disrupt the class will be asked to leave lecture or discussion and may be reported to the Dean of Students.

Some learning styles are best served by using personal electronics, such as laptops and iPads. These devices can be distracting to other learners. Therefore, students who prefer to use electronic devices for note-taking during lecture should use one side of the classroom.

Safety on Campus and in the Classroom

For a list of emergency procedures for all types of incidents, please visit the website of the Critical Incident Response Team (CIRT): https://cirt.arizona.edu/case-emergency/overview

Also watch the video available at

https://arizona.sabacloud.com/Saba/Web_spf/NA7P1PRD161/app/me/ledetail;spf-url=common%2Flearningeventdetail%2Fcrtfy000000000003841

University-wide Policies link

Links to the following UA policies are provided here: https://catalog.arizona.edu/syllabus-policies

- Absence and Class Participation Policies
- Threatening Behavior Policy
- Accessibility and Accommodations Policy
- Code of Academic Integrity
- Nondiscrimination and Anti-Harassment Policy

Department-wide Syllabus Policies and Resources link

Links to the following departmental syllabus policies and resources are provided here, https://www.cs.arizona.edu/cs-course-syllabus-policies :

- Department Code of Conduct
- Class Recordings
- Illnesses and Emergencies
- Obtaining Help
- Preferred Names and Pronouns
- Confidentiality of Student Records

- Additional Resources
- Land Acknowledgement Statement

Subject to Change Statement

Information contained in the course syllabus, other than the grade and absence policy, may be subject to change with advance notice, as deemed appropriate by the instructor.