CSC 483/583: Text Retrieval and Web Search

Gould-Simpson 906, Mon/Wed 12:30 – 1:45

Description of Course

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.

Course Prerequisites or Co-requisites

The students taking this course must know how to program, and have a decent understanding of data structures such as hash maps and trees. Ideally, the students should have taken a calculus course. We will, however, cover the necessary math background in class.

- Prerequisites: **CSC 345**
- Recommended: **Math 129 (Calc II)**

Instructor and Contact Information

Instructor: Mihai Surdeanu
Email: msurdeanu@email.arizona.edu
Web: http://surdeanu.info/mihai
Office: Gould-Simpson 746
Office hours: Mon/Wed 11 – noon

Teaching assistant: Enrique Noriega
Email: enoriega@email.arizona.edu
Office: Gould-Simpson 931
Office hours: Tue/Thu 11 – 12:30

Course Format and Teaching Methods

The course will be delivered using in-person lectures. No lab sections will be offered but the instructor encourages additional discussion on the topics introduced in the lecture materials. These discussions will be managed on a Piazza site controlled by the instructor.

The Piazza site is available here: https://piazza.com/arizona/spring2017/csc483583/home

Course Objectives and Expected Learning Outcomes

At the conclusion of this course students should: (a) understand multiple crawling, indexing, and retrieval methodologies (essentially what makes an information retrieval system), (b) have the capability to use this knowledge to code an information retrieval system (potentially using some low-level components, such as machine learning algorithms, from existing libraries); and (c) use existing information retrieval technology to build higher-level applications, such as question answering (e.g., IBM's Watson).

Graduate students are expected to have an in-depth understanding of these techniques. For example, graduate students are expected to know how to code the underlying machine learning framework necessary for text retrieval, such as algorithms for language models, classification, clustering, and “learning to rank” algorithms by understanding the underlying framework of the corresponding machine learning algorithms, such as reranking Perceptron or structured Support Vector Machines.

Absence and Class Participation Policy

UA’s policy concerning Class Attendance, Participation, and Administrative Drops is available at http://catalog.arizona.edu/policy/class-attendance-participation-and-administrative-drop

The UA policy regarding absences for any sincerely held religious belief, observance or practice will be accommodated where reasonable: http://policy.arizona.edu/human-resources/religious-accommodation-policy.

Absences preapproved by the UA Dean of Students (or dean’s designee) will be honored. See https://deanofstudents.arizona.edu/absences

Participating in the course and attending lectures and other course events are vital to the learning process. As such, attendance is required at all lectures and discussion section meetings. Students who miss class due to illness or emergency are required to bring documentation from their health-care provider or other relevant, professional third parties. Failure to submit third-party documentation will result in unexcused absences.

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.

Required Texts or Readings

This course follows the following textbook:

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

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

Required or Special Materials

No special tools or supplies needed.

Assignments and Examinations: Schedule/Due Dates

Grading will be based on four written assignments, two exams (midterm and final), a programming project, and overall in-class participation.

As a rule, work will not be accepted late except in case of documented emergency or illness. You may petition the professor in writing for an exception if you feel you have a compelling reason for turning work in late.

The due dates are as follows:

Task	Deadline
HW 1	January 29
HW 2	February 19
Midterm	March 1
HW 3	March 12
Spring recess	March 13 - 17
HW 4	April 9
Project	May 7

Final Examination and Project

The students will have multiple choices for a final project:

- Reconstructing (parts of) Watson, IBM's Question Answering system for the Jeopardy trivia game.
- Learning how to rank answers to science exam questions, based on the Allen Institute for Artificial Intelligence (AI2) Kaggle challenge: https://www.kaggle.com/c/the-allen-ai-science-challenge
- The Fake News Challenge (FNC): http://www.fakenewschallenge.org. This latter option is pending on the instructor's arranging a protocol with the task organizers.

This course will have a comprehensive written final examination. Information on the final exam regulations and schedule:

https://www.registrar.arizona.edu/courses/final-examination-regulations-and-information
http://www.registrar.arizona.edu/schedules/finals.htm

Grading Scale and Policies

The grading scheme is as follows:

Component	Weight
Written assignments	300 pts
Midterm exam	200 pts
Final exam	275 pts
Programming project	200 pts
In-class participation	25 pts
Total	**1000 pts**

Grade	Point Range
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 graduate and undergraduate students, the instructor will require graduate students to implement more complex algorithms for the programming project, which might require additional reading of research articles. The instructor will provide the additional reading material and will guide the research process. Similarly, assignments and 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).

Requests for incomplete (I) or withdrawal (W) must be made in accordance with University policies, which are available at http://catalog.arizona.edu/policy/grades-and-grading-system#incomplete and http://catalog.arizona.edu/policy/grades-and-grading-system#Withdrawal, respectively.

Scheduled Topics/Activities

The course will cover the topics listed below ("IIR x" indicates the corresponding chapter in the Introduction to Information Retrieval textbook):

Week	Topics
1	Introduction. Boolean retrieval (IIR 1)
2	The term vocabulary and postings lists (IIR 2). Dictionaries and tolerant retrieval (IIR 3)
3	Index construction (IIR 4). Index compression (IIR 5)
4	Scoring, term weighting, and the vector space model (IIR 6)
5	Computing scores in a complete search system (IIR 7). Evaluation in information retrieval (IIR 8)
6	Relevance feedback and query expansion (IIR 9)
7	Probabilistic information retrieval (IIR 11)
8	Language models for information retrieval (IIR 12 and lecture notes)
9	Text classification and Naive Bayes (IIR 13)
10	Vector space classification (IIR 14 and lecture notes)
11	Flat clustering (IIR 16). Hierarchical clustering (IIR 17)
12	Matrix decompositions and latent semantic indexing (IIR 18)
13	Web search basics (IIR 19)
14	Web crawling and indexes (IIR 20)
15	Link Analysis (IIR 21)

Department of Computer Science Code of Conduct

The Department of Computer Science is committed to providing and maintaining a supportive educational environment for all. We strive to be welcoming and inclusive, respect privacy and confidentiality, behave respectfully and courteously, and practice intellectual honesty. Disruptive behaviors (such as physical or emotional harassment, dismissive attitudes, and abuse of department resources) will not be tolerated. The complete Code of Conduct is available on our department web site. We expect that you will adhere to this code, as well as the UA Student Code of Conduct, while you are a member of this class.

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.).

Inclusive Excellence is a fundamental part of the University of Arizona's strategic plan and culture. As part of this initiative, the institution embraces and practices diversity and inclusiveness. These values are expected, respected and welcomed in this course.

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.

Threatening Behavior Policy

The UA Threatening Behavior by Students Policy prohibits threats of physical harm to any member of the University community, including to oneself. See http://policy.arizona.edu/education-and-student-affairs/threatening-behavior-students.

Elective Name and Pronoun Usage

This course supports elective gender pronoun use and self-identification; rosters indicating such choices will be updated throughout the semester, upon student request. As the course includes group work and in-class discussion, it is vitally important for us to create an educational environment of inclusion and mutual respect.

Accessibility and Accommodations

Our goal in this classroom is that learning experiences be as accessible as possible. If you anticipate or experience physical or academic barriers based on disability, please let me know immediately so that we can discuss options. You are also welcome to contact the Disability Resource Center (520-621-3268) to establish reasonable accommodations. For additional information on the Disability Resource Center and reasonable accommodations, please visit http://drc.arizona.edu.
If you have reasonable accommodations, please plan to meet with me by appointment or during office hours to discuss accommodations and how my course requirements and activities may impact your ability to fully participate.
Please be aware that the accessible table and chairs in this room should remain available for students who find that standard classroom seating is not usable.

Code of Academic Integrity

Students are encouraged to share intellectual views and discuss freely the principles and applications of course materials. However, graded work/exercises must be the product of independent effort unless otherwise instructed. Students are expected to adhere to the UA Code of Academic Integrity as described in the UA General Catalog. See http://deanofstudents.arizona.edu/academic-integrity/students/academic-integrity.

The University Libraries have some excellent tips for avoiding plagiarism, available at http://www.library.arizona.edu/help/tutorials/plagiarism/index.html.

Selling class notes and/or other course materials to other students or to a third party for resale is not permitted without the instructor's express written consent. Violations to this and other course rules are subject to the Code of Academic Integrity and may result in course sanctions. Additionally, students who use D2L or UA e-mail to sell or buy these copyrighted materials are subject to Code of Conduct Violations for misuse of student e-mail addresses. This conduct may also constitute copyright infringement.

UA Nondiscrimination and Anti-harassment Policy

The University is committed to creating and maintaining an environment free of discrimination; see http://policy.arizona.edu/human-resources/nondiscrimination-and-anti-harassment-policy

Our classroom is a place where everyone is encouraged to express well-formed opinions and their reasons for those opinions. We also want to create a tolerant and open environment where such opinions can be expressed without resorting to bullying or discrimination of others.

Additional Resources for Students

UA Academic policies and procedures are available at http://catalog.arizona.edu/policies

Student Assistance and Advocacy information is available at
http://deanofstudents.arizona.edu/student-assistance/students/student-assistance
Office of Diversity information is available at http://diversity.arizona.edu/

Campus Health information may be found here: http://www.health.arizona.edu/counseling-and-psych-services

OASIS Sexual Assault and Trauma Services
http://oasis.health.arizona.edu/hpps_oasis_program.htm

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.