CSC 483/583:
Text Retrieval and Web Search

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.

Locations and Times

Monday/Wednesday 5:00PM – 6:15PM, in Gould-Simpson, Room 906

Instructor Information

Instructor: Mihai Surdeanu
Email: `msurdeanu@email.arizona.edu`
Office: Gould-Simpson 746
Office Hours: by request

TA: Enrique Noriega
Email: `enoriega@email.arizona.edu`

Office: Gould-Simpson 934
Office Hours: Wednesday, 2 – 3PM

Course web page: `http://surdeanu.info/mihai/teaching/csc483-fall15/`

Piazza: `https://piazza.com/class/idhoajqezmu1qi`

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, and clustering.

Topics

- Boolean retrieval (IIR 1)
- The term vocabulary and postings lists (IIR 2)
- Dictionaries and tolerant retrieval (IIR 3)
- Index compression (IIR 5)
- Scoring, term weighting, and the vector space model (IIR 6)
- Computing scores in a complete search system (IIR 7)
- Evaluation in information retrieval (IIR 8)
- Tutorial of useful tools: Lucene, NLTK, CoreNLP (lecture notes)
- Relevance feedback and query expansion (IIR 9)
- Probabilistic information retrieval (IIR 11)

- Language models for information retrieval (IIR 12 and lecture notes)
- Text classification and Naive Bayes (IIR 13)
- Vector space classification (IIR 14)
- Flat clustering (IIR 16)
- Hierarchical clustering (IIR 17)
- Matrix decompositions and latent semantic indexing (IIR 18)
- Web search basics (IIR 19)
- Web crawling and indexes (IIR 20)
- Link Analysis (IIR 21)

If time permits:

- Introduction to computational advertising (lecture notes)
- Evolution of the Google IR System (lecture notes)
- Building Watson: An Overview of the DeepQA Project (lecture notes)

Course Format

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.

Required and Recommended Texts

This course follows the following textbook:

- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 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

All materials made available by the instructor or The University of Arizona remain the property of the copyright holders. They are provided for the use of students in this course for the duration of the course, except as noted.

Prerequisites/Recommended Knowledge

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)

Grading Policy

University policy regarding grades and grading systems is available at: `http://catalog.arizona.edu/2015-16/policies/grade.htm`

Grading

Grades are based on 4 written assignments, two exams (midterm and final), a programming project, a presentation of the project, and overall in-class participation. The grading scheme is as follows:

Requests for incompletes (I) and withdrawal (W) must be made in accordance with university policies, which are available at: `http://catalog.arizona.edu/2015-16/policies/grade.htm`.

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

Component	Weight
Written assignments	300 pts
Midterm Exam	200 pts
Final Exam	250 pts
Programming Project	200 pts
Final Presentation	25 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

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.

Programming Project
Students will have two choices for a final project:

1. The default project proposed by the instructor. This semester, we will reconstruct (parts of) Watson, IBM's Question Answering system for the Jeopardy trivia game.
2. Come up with your own idea. Your idea must use an existing IR system (e.g., Lucene) or IR technology to implement a real-world application.

Honors Credit

Students wishing to contract this course for Honors Credit should email me to set up an appointment to discuss the terms of the contract and to sign the Honors Course Contract Request Form. The form is available at `http://www.honors.arizona.edu/future-students/honors-credit-across-campus`.

Late Work Policy

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.

Attendance Policy

The UA's policy concerning Class Attendance and Administrative Drops is available at: http://catalog.arizona.edu/2015-16/policies/classatten.htm.

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 healthcare provider or other relevant, professional third parties. Failure to submit third-party documentation will result in unexcused absences.

The UA policy regarding absences on and accommodation of religious holidays is available at http://deanofstudents.arizona.edu/policies-and-codes/accommodation-religiou

From https://deanofstudents.arizona.edu/faqs:

> A Dean's Excuse provides excused absences for university-sponsored events/activities for academic, non-academic, and recognized student organizations. If a student must miss a class or classes for a university-sponsored event, the faculty or staff responsible for that event request a UA Official Activity Excused Absence Request Form from the Dean of Students Office.
>
> The Dean of Students Office does not have oversight of academic departments or faculty members and does not grant individual excused absences. Each faculty member manages his or her classroom in the manner in which they see fit and are the only ones who may determine what constitutes an excused absence. Therefore, we are unable to excuse absences for students, grant extensions, require that professors allow students to make-up missed work, or ensure students may miss class and submit late work without penalty, etc.
>
> The best thing to do is for you to communicate directly with your professor regarding your absence. Your professor is the only person who can excuse your absence, and determine if alternatives or make-up work is an option. Your professor may also request documentation of your situation. If your professor will not excuse your absence or grant make-up work the Dean of Students Office is not able to require them to do so.

Assignment/Testing Schedule/Due Dates

What	When
Midterm	October 14
Project due	December 8 before 11:59PM
Project presentations	December 9
Final	Between December 11 – 17 (TBA)

Written assignments will be due approximately every three weeks, as announced by the instructor. All assignments are due in the D2L dropbox by 11:59 P.M. on the indicated day.

Code of Conduct

The Arizona Board of Regents' Student Code of Conduct, ABOR Policy 5-308, prohibits threats of physical harm to any member of the University community, including to one's self: `http://azregents.asu.edu/rrc/Policy%20Manual/5-308-Student%20Code%20of%20Conduct.pdf`

All students, instructors, teaching assistants and section leaders in this course are expected to treat each other respectfully at all times. To foster a positive learning environment, students may not text, chat, make phone calls, play games, read the newspaper, or surf the web during lecture and discussion. 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. Students who continue to disrupt the class will be asked to leave the lecture and may be reported to the Dean of Students.

Classroom Electronics

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

Accessibility and Accommodations

It is the University's goal 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 Disability Resources (520–621–3268) to establish reasonable accommodations. For additional information on Disability Resources 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.

Student 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: `http://deanofstudents.arizona.edu/policies-and-codes/code-academic-integrity`. In particular:

- Students may not discuss individual homework with anybody other than the instructors and teaching assistants.
- Students may not share individual homework solutions with anybody.
- Students may post questions to Piazza, but should refrain from posting solutions or partial solutions.
- Students may not share test cases with anybody.
- Students may share class notes with anybody.
- Students may not seek individual homework help from anybody other than the instructors, teaching assistants, or departmental tutors.

If permitted, the use of open source or third party materials in student submissions must be clearly identified and credited. Assignment and project submissions must

be substantially the work of the student who submits the work. Copyrights, legal, and regulatory restrictions must be respected.

Students who violate the Code should expect a penalty that is **greater than the value of the work in question up to and including failing the course**. A record of the incident **will** be sent to the Dean of Students office. If you have been involved in other Code violations, the Dean of Students may impose additional sanctions.

Please read the CS department policy, at `http://www.cs.arizona.edu/policies/collaboration.html`, and the UA's Code of Academic Integrity, at `http://deanofstudents.arizona.edu/codeofacademicintegrity`, for further details on what constitutes cheating, the penalties that may result, and the procedures involved.

Additional Resources for Students

UA Non-discrimination and Anti-harassment policy: `http://policy.arizona.edu/human-resources/nondiscrimination-and-anti-harassment-policy`.

Student Assistance and Advocacy information is available at: `http://deanofstudents.arizona.edu/student-assistance/students/student-assistance`.

Confidentiality of Student Records

`http://www.registrar.arizona.edu/ferpa/ferpa-compliance`

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.