ISTA 456/556:
Text Retrieval and Web Search

Mihai Surdeanu

Last Revised August 19, 2014

1 Course Information

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.

For the class project, students can use any language that the instructor knows as well, e.g., Python, Java, Scala, Clojure, C/C++, or Perl. We will also review several natural language processing and information retrieval toolkits that might be useful for the implementation of the final project, such as the Natural Language Toolkit (NLTK)[1], Stanford's CoreNLP,[2], and Lucene, which is a widely used indexing and search toolkit.[3]

[1] `http://nltk.org/`
[2] `http://nlp.stanford.edu/software/corenlp.shtml`
[3] `http://lucene.apache.org/`

Credits

3 units

Prerequisites

ISTA 350 (Programming for Informatics Applications) and Math 215 (Linear Algebra) or equivalent.

Locations and Times

Monday/Wednesday 11:00AM – 12:15PM, in Social Sciences, Room 411

Readings

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.

Instructor Information

Mihai Surdeanu
Email: `msurdeanu@email.arizona.edu`
Office: Gould-Simpson 811
Office Hours: only by request

2 Schedule

Tentative Schedule of Lecture Topics

- Boolean retrieval (IIR 1)
- The term vocabulary and postings lists (IIR 2)
- Dictionaries and tolerant retrieval (IIR 3
- Index construction (IIR 4)
- 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)
- Text classification and Naive Bayes (IIR 13)
- Vector space classification (IIR 14)
- Support vector machines and machine learning on documents (IIR 15) (possibly skipped)
- 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:

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

3 Important Dates

What	When
Midterm	October 15
Project due	December 7 before 11:59PM
Project presentations	December 10
Final	Between December 12 – 18 (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.

4 Grading

Grades are based on 4–6 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:

Component	**Weight**
Written assignments	300 pts
Midterm Exam	200 pts
Final Exam	200 pts
Programming Project	200 pts
Final Presentation	50 pts
In-class Participation	50 pts
Total	1000 pts

Grade	**Point Range**
A	900 – 1000
B	800 – 899
C	700 – 799
D	600 – 699
E	0 – 599

Programming Project

Students will have two choices for a final project:

1. One of the default projects proposed by the instructor. This semester, the instructor offers two alternatives: (a) we will reconstruct (parts of) Watson, IBM's Question Answering system for the Jeopardy trivia game, or (b) we will improve the eLEAS (Levels of Emotional Awareness Scale) test (`http://www.eleastest.net/`).
2. Come up with your own idea. This idea must fall into one of these categories: (a) it implements a complex IR model from scratch, or (b) it uses an existing IR system (e.g., Lucene) to implement a real-world application.

Grade Disputes

Disputes about grades on a particular assignment or project will be entertained for two weeks from the day the assignment or project is returned, or 1 day after the final exam, whichever is sooner. These will be resolved by re-grading the entire work. Note that this can result in a lower grade in the event that new mistakes are discovered.

No negotiations about individual students' letter grades will be entertained once final grades are assigned, except as permitted by the policy stated above.

Collaboration Policy

Students are encouraged to work together, both in class / office hours and otherwise, to understand problems and general approaches for solutions. However, **written assignments, project implementations and the associated documentation for the project must be completed individually**. **Copying another person's work (even if it comes from a website) is not permitted and will be treated as a case of academic dishonesty**.

Late Policy

Projects are due electronically via D2L by the stated deadline. Permission for an extension must be granted by the instructor *in advance* of the deadline in order to receive full credit for a late submission. The first request by a given student is likely to be granted; the probability decreases with each subsequent request. No project will be accepted once solutions are posted online.

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, state-of-the-art 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. Because of this, projects will be graded separately for undergraduate and graduate students, as described in the project's description. Furthermore, the written assignments will have additional questions for graduate students.

5 University Policies

Missed Classes (Absence)

Accommodation of Religious Observance and Practice: `http://deanofstudents.arizona.edu/religiousobservanceandpractice`

All holidays or special events observed by organized religions will be honored for those students who show affiliation with such religions. Absences pre-approved by the UA Dean of Students office will be honored. No matter the reason for missing class, the student is always responsible for the missed material.

With the exception of the above, attendance is mandatory. Students who miss more than $1/3$ **of classes will be dropped.**

Classroom Behavior

Students are expected to behave respectfully toward each other and to the instructor and TAs. Disrespectful behavior includes the use of cell phones or other electronic devices in the classroom during class hours. Please do not play computer games, check your email, surf the web, text your friends, read the paper, chatter at length with fellow students, etc. If you don't want to listen to the lecture and participate in classroom discussions, please leave the lecture hall.

Asking Questions: During class, feel free to interrupt with questions whenever they occur to you. The instructor may ask you to hold off on your question for a few moments.

Answering Questions: We frequently ask questions of the class during lectures to judge the level of understanding (and to break up the monotony). Some students really like answering questions, sometimes to the point of discouraging anyone else from answering. If you are an eager answerer, pace yourself; let someone else answer an easy one once in a while, and save the hard ones for yourself.

Note that the in-class participation credit (5/100) will be assigned based on both the questions asked and the questions answered in class.

The Arizona Board of Regents Student Code of Conduct is here: `http://deanofstudents.arizona.edu/studentcodeofconduct`

ABOR Policy 5-308, prohibits threats of physical harm to any member of the University community, including to oneself. See: http://policy.web.arizona.edu/threatening-behavior-students.

Special Needs and Accommodations

Students who need special accommodation or services should contact the

Disability Resources Center
1224 East Lowell Street, Tucson, AZ 85721
(520) 621-3268
FAX (520) 621-9423
email: uadrc@email.arizona.edu
web: http://drc.arizona.edu/.

You must register and request that the DRC send official notification of your accommodations needs as soon as possible. Please plan to meet with the instructor by appointment or during office hours to discuss accommodations and how the course requirements and activities may impact your ability to fully participate. The need for accommodations must be documented by the appropriate office.

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 must be the product of independent effort unless otherwise instructed. Students are expected to adhere to the UA Code of Academic Integrity as described here: http://deanofstudents.arizona.edu/codeofacademicintegrity.

Confidentiality of Student Records

See http://www.registrar.arizona.edu/ferpa/default.htm

On Dropping Classes

If you find yourself thinking about dropping this (or any other) class, first make sure that that's what you really want to do. Chatting with the instructor or your academic advisor may help. If you drop within the first four weeks of the semester, there will be no notation on your transcript; it will be as though you'd never enrolled. During the fifth through the eighth weeks, a drop will be recorded on your transcript. You will receive a "WP" (withdrawn passing) only if you were passing the class at the time of your drop. After the eighth week, dropping becomes a challenge, because you need to explain to the instructor and to the dean why you were unable to drop the class during the first half of the semester.

Subject to Change Statement

The instructors reserve the right to change with advance notice where appropriate the content of the course. This right does not apply to posted grading and absence policies or University Policies.