ISTA 455/555:
Applied Natural Language Processing

Mihai Surdeanu

Last Revised July 26, 2013

1 Course Information

Most of web data today consists of unstructured text. This course will cover the fundamental knowledge necessary to organize such texts, search them a meaningful way, and extract relevant information from them. This course will teach natural language processing through the design and development of end-to-end natural language understanding applications, including sentiment analysis (e.g., is this review positive or negative?), information extraction (e.g., extracting named entities and their relations from text), and question answering (retrieving exact answers to natural language questions such as "What is the capital of France?" from large document collections). We will use several natural language processing (NLP) and machine learning (ML) toolkits, such as NLTK, scikit-learn, and Stanford's CoreNLP. The main programming language used in the course will be Python (version 2), but code written in other languages, such as Java, Scala, Clojure, or C/C++, will be accepted as well. Languages tailored specifically for Microsoft platforms (e.g., C#, F#) will not be accepted.

Course objectives: Students will be exposed to a range of important natural language processing and information retrieval problems, such as sentiment analysis, named entity recognition, relation extraction, and question answering. Students will learn how to address these problems using a variety of techniques, ranging from lexicon-based approaches to supervised or lightly supervised ML models. Lastly, students will construct complex applications that combine multiple algorithms into systems that solve real-world problems.

Credits: 4 units.

Prerequisites

Two programming courses at the level of ISTA 130 or higher.

Highly recommended: A NLP course such as LING 439/539 or LING 338, and a ML course such as ISTA 421/521. If you did not take any of the recommended courses, you will need the instructor's permission before taking this class.

Locations and Times

Lectures: Monday/Wednesday 8:30 - 9:45 in Social Sciences, Room 118

Lab: Wednesday 10:00 – 11:50 in McClelland Park, Room 102

Readings

This course does not follow any particular book, but it is recommended that students read at least one book on NLP and one on ML. The recommended textbooks for this course are:

- Daniel Jurafsky and James H. Martin. 2008. *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Second Edition*. Pearson/Prentice-Hall. `http://www.cs.colorado.edu/~martin/slp.html` (in the bookstore)
- Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. 6th printing with corrections, 2003. The MIT Press. `http://nlp.stanford.edu/fsnlp/` (available for free electronically through UA library)
- 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/`

- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media. Available for free at `http://nltk.org/book/`
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer. Available for free at `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`
- Christopher M. Bishop. 2009. *Pattern Recognition and Machine Learning*. Springer.
- *Applied machine learning in Python with scikit-learn*. Available here `http://scikit-learn.github.io/scikit-learn-tutorial/`

Instructor Information

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

2 Grading

This course will not have written assignments or exams. Instead, the students will design and develop four systems from scratch: (a) a "getting your feet wet" project; (b) a sentiment analysis system; (c) an information extraction system, and (d) a culminating project that focuses on question answering (or an alternative topic). The grading will be based on the successful completion of these systems. Each of the four projects will be organized along several clearly defined milestones, which will be graded individually.

The overall course grade will consist of these four programming projects, two in-class presentations, and in-class participation. The first presentation will be on one of projects two or three, and will be up to 10 minutes per student. The second presentation will describe the final project and will be up to 20 minutes per student. The grading scheme is as follows:

Component	**Weight**
Project 1	10 pts
Project 2	20 pts
Project 3	20 pts
Project 4	30 pts
Presentation 1	5 pts
Presentation 2	10 pts
In-class Participation	5 pts
Total	100 pts

Grade	**Point Range**
A	90 – 100
B	80 – 89
C	70 – 79
D	60 – 69
E	0 – 59

Grade Disputes

Disputes about grades on a particular project will be entertained for two weeks from the day the project is due, or 1 day before grades are due, whichever is sooner. These will be resolved by re-grading the entire project. 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 / lab / office hours and otherwise, to understand problems and general approaches for solutions. However, **project implementations and the associated documentation for each 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 lab 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.

455 vs. 555

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 assigned projects. This 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 each project's description.

3 Schedule

Tentative Schedule of Lecture Topics and Presentations

Topic	Approximate Dates	Topics
Introduction	Aug 26, 28, Sep 4	Syllabus, NLP and ML review
Sentiment Analysis	Sep 9, 11, 16, 18	Lexicon-based methods; supervised classification; distant supervision; using sentiment analysis for event forecasting
Information Extraction	Sep 23, 25, 30, Oct 2, 7, 9	IE as text segmentation; relation extraction: distant supervision, rule-based, supervised, unsupervised; event extraction: rule-based, supervised, unsupervised
Question Answering (1)	Oct 14, 16	Factoid QA
Presentations	**Oct 21**	**Presentations of project 2 or 3**
Question Answering (2)	Oct 23, 28, 30, Nov 4	Factoid QA; non-factoid QA; web-based QA
Information Retrieval	Nov 6, 13, 18, 20, 25, 27	Boolean retrieval; vector space models; relevance feedback; query expansion; link analysis
Topic Modeling	Dec 2, 4	Latent semantic analysis; latent Dirichlet allocation
Presentations	**Dec 9, 11**	**Presentations of final project**

Project Deadlines

All assignments are due in the D2L dropbox by 11:59 P.M. on the indicated day.

Assignment	Due Date
Project 1	September 8
Project 2	September 22
Project 3	October 13
Project 4	November 24

4 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.