

## Syllabus and Organization

### Syllabus

With the advent of the information age there is a growing need for automatic techniques for text processing. Most successful techniques today use statistical or machine learning components to achieve good performance. That is, a large corpus of annotated text is used to infer an effective performance element. The course will give a brief introduction to the area of natural language processing and proceed to review corpus-based techniques and applications. The course will combine the study of techniques with practical assignments applying them to real-world data. Techniques and tools draw on probability theory and statistics, information theory and algorithmic learning theory. Applications include tasks at various levels of complexity such as word-sense disambiguation, context-sensitive spelling correction, sentence parsing, document categorization and information retrieval.

#### Course Outline:

(4-5 weeks) Word-sense disambiguation and context-sensitive spelling exemplifying prediction tasks in Natural Language Processing. Foundations from probability and statistics as well as machine learning and application of the techniques to these tasks.

(6-7 weeks) Language modeling and processing. Background from Linguistics and information theory; part of speech tagging; n-gram models, hidden Markov models and probabilistic grammars and parsing.

(1-3 weeks) Additional applications and techniques from text categorization and information retrieval.

### Organization

**Location:** Lectures take place on Mondays and Wednesdays, 5:10-6:30, in Halligan H-106.

**Prerequisites:** COMP 160 or similar background and permission of instructor. Programming in C++ (from COMP11 and COMP15).

**Instructor:** Roni Khardon, [roni@eecs.tufts.edu](mailto:roni@eecs.tufts.edu), Tel: 627-5290, Office: H-230. Office hours Tue, Wed 3:30-4:30pm or by appointment.

**Teaching assistant:** Marta Arias, [marias@eecs.tufts.edu](mailto:marias@eecs.tufts.edu), Office hours TBA.

**Assignments:** There will be two kinds of assignments. There will be 4-5 Written assignments; these will include small exercises keeping track of material covered in class. There will be 4 computer based exercises. These will include either programming a particular technique and applying it to text data, or performing experiments with a given system, learning how to use it and studying the influence of various parameters on its performance.

**Collaboration:** You are allowed to discuss the problems with other students at a high level but must work out the details yourself and submit your own solutions. Obviously you should *not* copy work from any source under any circumstances.

**Exam:** There will be one exam in class during the last lecture slot: Mon, April 30th, 5:10-6:30.  
**Please make sure you can attend this exam.**

**Grading:** The course mark is based on written assignments (30%) computer based assignments (40%) and class exam (30%)

## Course Web Page

Assignments, notes and other pointers will be posted on: <http://www.eecs.tufts.edu/g/150TP/>

## Textbook and notes

- The required text is: *Foundations of Statistical Natural Language Processing*, by C. Manning and H. Schütze, MIT Press, 1999. The course does not follow the book in a linear order but does have a significant overlap with this book.
- Additional material (research papers) will be distributed.

There are several other books covering similar or related material that can be found on reserve in the library.

- James Allen, *Natural Language Understanding*, Addison-Wesley, 1995.
- Eugene Charniak, *Statistical Language Learning*, MIT Press, 1993.
- Daniel Jurafsky and James Martin, *Speech and Language Processing*, Prentice Hall, 2000.
- Tom Mitchell, *Machine Learning*, McGraw-Hill, 1997.