Administrivia

Instructor:
Marti Hearst (hearst@sims)
212 South Hall, 510-642-8016
Office Hours: Mondays 3-4pm, Thursdays 11am-12pm

TA (informally):
Preslav Nakov (nakov@sims)
Lab / Office hours: Mondays 3-4pm

Class Meetings

Class meets on Monday and Wednesday from 10:30-12:00 in 202 South Hall. The format of the class will be primarily lecturing and in-class experimentation with NLP software.

Prerequisites

IS206, a CS background, or equivalent. This class will involve using various software tools and writing code to glue them together. We will be using the **Python** programming language.

Grading

Grading will be a mix of 50% homework assignments and 50% final project.

Late Policy

Late assignments will be penalized with a reduction in grade unless otherwise approved by the instructor.

Readings and Books

Unfortunately, there really is not appropriate textbook for this course. Instead we will be reading online readings and handouts.

Students will need to learn a bit of Python programming for this course, and so an introductory Python book is recommended.

See **resources** for suggested books.

Schedule

Note: Times and Topics Subject to Change

Activity: Install Python and NLTK-lite

	Date	Lecture Topics / Assignments	Readings (due on date shown)
1	Aug 28	Course Introduction	No reading.
	Aug 30	Introductions; Python Intro	Chapter 1 from Jurafsky and Martin Python Programming Fundamentals (NLTK-lite tutorial, Sections 2.1 - 2.4)
2	Sep 4	Holiday: no class	Do reading for Sept 6.
	Sep 6	Tokenization, Regular Expressions Assignment 1: Tokenization	Python Programming Fundamentals (NLTK-lite tutorial, Sections 2.5-2.8) Regular Expressions, NLTK-lite tutorial
3	Sep 11	Morphology and Stemming	Wikipedia entry on Morphology Words, Sections 3.1 - 3.3 and 3.5, NLTK-lite tutorial
	Sep 13	Computing with Ngrams	Words, Section 3.4 NLTK-lite tutorial
4	Sep 18	POS Tagging	Tagging, NLTK-lite Tutorial, Sections 4.0-4.4
	Sep 20	POS Tagging with n-grams Assignment 2 assigned	Tagging, NLTK-lite Tutorial, Sections 4.4-4.7
5	Sep 25	Shallow Parsing	Chunk Parsing, NLTK-lite Tutorial, Sections 5.0-5.3
	Sep 27	Shallow Parsing, cont. (code for lecture)	Chunk Parsing, NLTK-lite Tutorial, Sections 5.4-5.6 H.P. Luhn, The automatic creation of literature abstracts, *IBM Journal of R&D*, 2(2), 1958.
6	Oct 2	Summarization	H.P. Edmonson, New methods in automatic extracting, *JACM*, 16(2), 1969. J. Kupiec, J. Pedersen, F. Chen, A trainable document summarizer, *Proc. of SIGIR*, 1995.
	Oct 4	Summarization; Intro to Probability Theory Assignment 3 assigned	D. Marcu, Discourse trees are good indicators of importance in text, in *Advances in Automatic Text Summarization*, 1999. J. Goldstein, V. Mittal, J. Carbonell, M. Kantrowitz, Multi-Document Summarization by Sentence Extraction, ANLP/NAACL Workshop, 2000.
7	Oct 9	Probabilities, cont.; Author Identification	Can Pseudonymity Really Guarantee Privacy? by Rao and Rohatgi, in 9th USENIX Security Symposium, 2000
	Oct 11	Guest lecture: **Elizabeth Charnock** and **Steve Roberts** of Cataphora	
		Text Classification	Machine Learning in Automated Text Categorization,

8	Oct 16	**Intro** (Guest Lecture: Preslav Nakov)	Sebastiani, *ACM Computing Surveys* 34 (1), 2002. Sections 1-4. (Note: this reading is optional.)
	Oct 18	**Summarization experiment; Class project ideas**	**A comparative study on Feature Selection in Text Categorization**, Yang and Pedersen, *Proc. of ICML*, 1997. **Sebastiani Survey**, Section 5 (optional)
9	Oct 23	**Text Classification: Feature Selection** **Project Proposal assigned; due Oct 30**	**Sebastiani Survey**, Sections 6.3, 6.4, 6.8, 6.9, 6.10. (optional)
	Oct 25	Guest lecture: **Peter Jackson**, Chief Research Scientist and VP, Technology Thomson Legal & Regulatory	
10	Oct 30	**Text Classification: Using Weka** **Assignment 4 assigned; due Nov 13**	**Sebastiani Survey**, Sections 7.2, 7.3 (optional) **Weka Simple Experiments Documentation** **Weka Explorer Documentation**
	Nov 1	**Text Classification: Algorithms**	**Weka Advanced Experiments Documentation**
11	Nov 6	**Clustering, LSA**	(Optional) **An introduction to LSA** (found by Hannes)
	Nov 8	**More cluster examples; Blog Analysis**	**Predicting Movie Sales from Blogger Sentiment**, Mishne & Glance, AAAI-CAAW 2006. **Deriving Marketing Intelligence from Online Discussion**, Glance et al., KDD'05 (optional)
12	Nov 13	**Lexicon Acquisition**	**Extracting Product Features and Opinions from Reviews**, Popsecu & Etzioni, HLT/EMNLP 2005. **Towards a Robust Metric of Opinion**, Nigam & Hurst, AAAI-EAAT 2004.
	Nov 15	**Information Extraction**	
13	Nov 20	Guest lecture: Barbara Rosario, Intel Research**Finding Semantic Relations**	
	Nov 22	Guest lecture: Roger Magoulas, O'Reilly Media, **Text Mining at O'Reilly Media**	
14	Nov 27	**Discourse Processing** **Project Writeup and Presentation Schedule**	
	Nov 29	**Question Answering**	
15	Dec 4	Class Presentations	
	Dec 6	Class Presentations	

Assignments and Activities

For Dec 4, 6,and 13:
Project Writeup and Presentation Schedule

For Monday Nov 13:
Assignment 4: Text Categorization

For Mon Oct 30:
Final Project Proposal

For Wed, Oct 11 and Mon Oct 16:
Assignment 3.

For Wed, Sep 27:
Assignment 2.

For Wed, Sep 13:
Assignment 1.