Representing and Accessing [Textual] Digital Information (COMS/INFO 630), Spring 2006
1/24/06: **Course Description and Policies**

> All satisfied with their seats? O.K. No talking, no smoking, no knitting, no newspaper reading, no sleeping, and for God's sake take notes.
>
> – Nabokov, *Lectures on Literature*

Instructor Prof. Lillian Lee. Office hours: Wednesdays 10:30-12 in Upson 4152, or by appointment; for contact info and updates, see http://www.cs.cornell.edu/home/llee .

Lecture time and place TR 2:55-4:10. While the first few lectures will be in Bard 140, we anticipate moving to Upson 315 (with advance notice, of course).

Course homepage http://www.cs.cornell.edu/courses/cs630/2006sp ; (some) handouts, references, and resources will be posted.

Prerequisites Elementary computer science background (e.g., what a graph is), elementary knowledge of probability (e.g., what conditional probability is), elementary knowledge of linear algebra (e.g., how to compute an inner product), mathematical maturity.

Course work The approximate proportion of the course grade is indicated in brackets.

- Lecture guides [60%]. It is often said that the best way to learn something is to teach it. To that end, for each lecture and/or topic, one student or student group will be responsible for writing up, within a week, a hardcopy document consisting of: (a) roughly textbook-quality "scribe notes" (edited transcriptions) for the lecture(s); and (b) an original, non-trivial "finger-exercise" problem testing one or more basic concepts, together with a worked solution to that problem. For full credit, the guide should also include a "deeper" question and solution (partial solutions may be acceptable if the question is sufficiently "research-y"). Extra credit may be awarded for extra effort, such as including brief summaries of recent related results or posing additional interesting problems.

  The lecture guides will be posted and/or distributed to the class (with attribution) after any further revisions requested by the instructor are completed.

  More details will be forthcoming (for example, some are dependent on student enrollment), but the current intent is that each group will prepare a lecture guide roughly every two weeks.

- Exams: A midterm (in class, Thursday March 16) [15%] and final (7pm-9:30pm, Wednesday May 17) [15%]. These are intended to test basic knowledge of the course material. The worked problems given in the lecture guides should provide helpful preparation, and it is intended that the exam questions should be close in spirit to those problems (this is contingent on the quality of such problems.)

- Participation [10%]. Participation can occur outside of the regular class period (e.g., technical conversation in office hours or via email), but lecture attendance will figure into this component.

Recommended reference texts Baeza-Yates and Ribeiro-Neto, *Modern Information Retrieval*, 1999; Jelinek, *Statistical Methods for Speech Recognition*, 1997; Jurafsky and Martin, *Speech and Language Processing*, 2000; Manning and Schütze, *Foundations of Statistical Natural Language Processing*, 1999.

Tentative syllabus Likely representative source material is indicated.

• Basics of "classic" information retrieval. Foundations for term weighting. Singhal, Buckley, Mitra, "Pivoted document length normalization", SIGIR 1996; Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF", *Journal of Documentation*, 2004; Fang, Tao, Zhai, "A formal study of information retrieval heuristics", SIGIR 2004.

• Latent semantic indexing. Generative and alternative analyses. Papadimitriou, Raghavan, Tamaki, Vempala, "Latent Semantic Indexing: A Probabilistic Analysis", *Journal of Computer and Systems Sciences*, 2000; Ando, Lee, "Iterative Residual Rescaling: An analysis and generalization of LSI", SIGIR 2001; Bast, Majumdar, "Why Spectral Retrieval Works", SIGIR 2005.

• (Statistical) source models for IR and related tasks. The language-modeling approach to IR. Basics of information theory. Smoothing. Probabilistic Latent Semantic Indexing/Analysis (PLSI/PLSA). Latent Dirichlet Allocation (LDA). Ponte, Croft, "A language modeling approach to information retrieval", SIGIR 1998; Lafferty, Zhai, "Probabilistic relevance models based on document and query generation", and Sparck Jones, Robertson, Hiemstra, Zaragoza, "Language modeling and relevance", both in *Language Modeling for Information Retrieval*, 2003; Lafferty, Zhai, "Document language models, query models, and risk minimization for information retrieval", SIGIR 2001; Hofmann, "Unsupervised learning by probabilistic latent semantic analysis", *Machine Learning Journal*, 2001; Blei, Ng, Jordan, "Latent Dirichlet allocation", *Journal of Machine Learning Research*, 2003.

• Clustering. Interactions with language modeling. Co-clustering. Kurland, Lee, "Corpus structure, language models, and ad hoc information retrieval", SIGIR 2004; Liu, Croft, "Cluster-based retrieval using language models", SIGIR 2004; Pereira, Tishby, Lee, "Distributional clustering of English words", ACL, 1993; Tishby, Pereira, Bialek, "The Information Bottleneck Method", Allerton Conference on Communication, Control and Computing, 1999; Dhillon, "Co-Clustering documents and words using bipartite spectral graph partitioning", KDD 2001.

• Text categorization: topic, sentiment, author. Joachims, "A statistical learning model of text classification with Support Vector Machines", SIGIR 2001; Pang, Lee, Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques", EMNLP 2002; Mosteller, Wallace, *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*, 1984.

• Sequence models: hidden Markov models (HMMs) and conditional random fields (CRFs). Grenager, Klein, Manning, "Unsupervised learning of field segmentation models for information extraction", ACL 2005; Barzilay, Lee, "Catching the drift: Probabilistic content models, with applications to generation and summarization", HLT-NAACL 2004; Abney, Light, "Hiding a semantic class hierarchy in a Markov model", ACL workshop on Unsupervised Learning in Natural Language Procesing, 1999; McCallum, Freitag, Pereira, "Maximum entropy Markov models for information extraction and segmentation", ICML 2000; Lafferty, McCallum, Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", ICML 2001.

• Grammars: feature-based context-free grammars; tree adjoining grammars (TAGs); synchronous grammars; inversion transduction grammars. Melamed, "Multitext Grammars and Synchronous Parsers", HLT/NAACL 2003; Shieber, "Synchronous grammars as tree transducers", TAG+ 2004; Wu, "Stochastic inversion transduction grammars and bilingual parsing of parallel corpora", *Computational Linguistics*, 1997; Zhang, Gildea, "Stochastic lexicalized inversion transduction grammar for alignment", ACL 2005.

• Transformations: statistical machine translation; sentence compression; paraphrasing. Brown et al, "The mathematics of statistical machine translation: parameter estimation", *Computational Linguistics*, 1993; Knight, Marcu, "Statistics-based summarization — Step one: sentence compression", AAAI 2000; Barzilay, Lee, "Learning to Paraphrase: An Unsupervised Approach using Multiple-Sequence Alignment", HLT/NAACL 2003; Pang, Knight, Marcu, "Syntax-based alignment of multiple translations: Extracting paraphrases and generating new Sentences", HLT/NAACL 2003.

• Object lesson: the first Loebner Prize. Shieber, "Lessons from a restricted Turing test", *Communications of the ACM*, 1994.