CS 276 / LING 286: Information Retrieval and Web Search

Course schedule

Lectures take place on Tuesdays and Thursdays from 4:30 to 5:50pm in NVIDIA Auditorium (https://campus-map.stanford.edu/?srch=NVIDIA+Auditorium), located in the basement of the Jen-Hsun Huang Engineering Center on campus. Lecture videos are recorded by SCPD and available to all enrolled students here (https://mvideox.stanford.edu/Course/904).

Students are also expected to become familiar with the course material presented in a series of video lectures that are hosted on Coursera. Access the **Coursera videos** here (https://www.coursera.org/learn/cs276/home) (be sure to log in with your Stanford email account to get access).

Week	Date	Event	Description & materials	Readings & other resources
Week 1	Tues. 4/4	Lecture (Pandu)	**Introduction to the course** Coursera content: • Videos: "Boolean Retrieval" • Slides: PPT (handouts/lecture1-intro.ppt) | PDF/6 (handouts/lecture1-intro-handout-6-per.pdf) | PDF/1 (handouts/lecture1-intro-handout-1-per.pdf)	• IIR chapter 1 (http://nlp.stanford.edu/IR-book/pdf/01bool.pdf) • MG section 3.2 • MIR section 8.2 • Shakespeare plays (http://www.rhymezone.com/shakespeare/)
	Thurs. 4/6	In-class lab (Chris)	**Merge algorithm for proximity queries using a positional index** (starter code (https://github.com/manning/MergeAlgorithms)) You will need the following Google doc during the in-class lab session: Postings list intersection class session (https://docs.google.com/document/d/15xnoTUm3kyeIlDqPTs4H2w_CYf4uaua5MeukFWx1rW8/edit) Coursera content: • Videos: "Term Vocabulary and Postings Lists" • Slides: PPT (handouts/lecture2-dictionary.ppt) | PDF/6 (handouts/lecture2-dictionary-handout-6-per.pdf) | PDF/1 (handouts/lecture2-dictionary-handout-1-per.pdf) • Videos: "Index Construction" • Slides: PPT (handouts/lecture4-indexconstruction.ppt) | PDF/6 (handouts/lecture4-indexconstruction-handout-6-per.pdf) | PDF/1 (handouts/lecture4-indexconstruction-handout-1-per.pdf)	• IIR chapter 2 (http://nlp.stanford.edu/IR-book/pdf/02voc.pdf) • IIR chapter 4 (http://nlp.stanford.edu/IR-book/pdf/04const.pdf) • MG sections 3.6, 4.3 • MIR section 7.2 • Porter's stemmer (MIR) (http://www.sims.berkeley.edu/~hearst/irbook/porter.html) • Porter stemming algorithm (Official) (http://www.tartarus.org/~martin/PorterStemmer/) • A skip list cookbook (Pugh 1990) (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.17.524) • Fast phrase querying with combined indexes (Williams, Zobel, Bahle 2004) (http://portal.acm.org/citation.cfm?id=1028102) • Efficient phrase querying with an auxiliary index (Bahle, Williams, Zobel 2002) (http://portal.acm.org/citation.cfm?id=564415)
	Thurs. 4/6	PA1 released	**Programming assignment #1 (pa/pa1.pdf) released**	
Week 2	Tues. 4/11	Lecture (Pandu)	**Spelling correction** In-class lecture notes: • PDF/6 (handouts/spell_correction-6up.pdf) | PDF/1 (handouts/spell_correction.pdf) Coursera content: • Videos: "Dictionaries and Tolerant Retrieval" • Slides: PPT (handouts/lecture3-tolerant-retrieval.ppt) | PDF/6 (handouts/lecture3-tolerant-retrieval-handout-6-per.pdf) | PDF/1 (handouts/lecture3-tolerant-retrieval-handout-1-per.pdf)	• IIR chapter 3 (http://nlp.stanford.edu/IR-book/pdf/03dict.pdf) • MG section 4.2 • How to write a spelling corrector (Peter Norvig) (http://norvig.com/spell-correct.html) • Techniques for automatically correcting words in text (Kukich 1992) (http://portal.acm.org/citation.cfm?id=146380) • Finding approximate matches in large lexicons (Zobel and Dart 1995) (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.9400) • Efficient Generation and Ranking of Spelling Error Corrections (Tillenius) (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.49.1392)
	Thurs. 4/13	In-class lab (Chris)	**Algorithms for postings list compression** (starter code (https://github.com/manning/CompressionAlgorithms)) (Google doc with class results (https://docs.google.com/spreadsheets/d/1znmKYjuj4hIehTrotqvkT_HLHTH6WSEEZ0ActjX1xo0/edit#gid=0)) Coursera content: • Videos: "Index Compression" • Slides: PPT (handouts/lecture5-compression.ppt) | PDF/6 (handouts/lecture5-compression-handout-6-per.pdf) | PDF/1 (handouts/lecture5-compression-handout-1-per.pdf)	• IIR chapter 5 (http://nlp.stanford.edu/IR-book/pdf/05comp.pdf) • MG sections 3.3-3.4 • Compression of inverted indexes for fast query evaluation (Scholer et al. 2002) (http://portal.acm.org/citation.cfm?id=564416) • Inverted index compression using word-aligned binary codes (Anh and Moffat 2005) (http://www.springerlink.com/content/j66851228120170t/) • Inverted index compression and query processing with optimized document ordering (Yan et al. 2009) (http://engineering.nyu.edu/~suel/papers/comp.pdf)
	Thurs. 4/13	PS1 released	**Problem set #1 (ps/ps1.pdf) released**	[PS1 solution] (restricted/ps1-solution.pdf)

	Thurs. 4/13	Query quiz released	Query quiz (https://goo.gl/forms/vpUekPg6isLlpP8H3) released	
Week 3	Tues. 4/18	Lecture (Chris)	**Probabilistic IR: the binary independence model** In-class lecture notes: • PPT (handouts/lecture11-probir.pptx)	PDF/6 (handouts/lecture11-probir-6up.pdf)	PDF/1 (handouts/lecture11-probir.pdf) Coursera content: • Videos: "Vector Space Model" • Slides: PPT (handouts/lecture6-tfidf.ppt)	PDF/6 (handouts/lecture6-tfidf-handout-6-per.pdf)	PDF/1 (handouts/lecture6-tfidf-handout-1-per.pdf)	• IIR chapter 6 (http://nlp.stanford.edu/IR-book/pdf/06vect.pdf) • IIR chapter 11 (http://nlp.stanford.edu/IR-book/pdf/11prob.pdf)
	Thurs. 4/20	Lecture (Chris)	**Computing scores and BM25F** In-class lecture notes: • PPT (handouts/lecture12-bm25etc.pptx)	PDF/6 (handouts/lecture12-bm25etc-6up.pdf)	PDF/1 (handouts/lecture12-bm25etc.pdf) Coursera content: • Videos: "Computing Scores" • Slides: PPT (handouts/lecture7-vectorspace.pptx)	PDF/6 (handouts/lecture7-vectorspace-6per.pdf)	PDF/1 (handouts/lecture7-vectorspace-1per.pdf)	• IIR chapter 7 (http://nlp.stanford.edu/IR-book/pdf/07system.pdf) • IIR chapter 11 (http://nlp.stanford.edu/IR-book/pdf/11prob.pdf)
	Thurs. 4/20	PA1 due	**Programming assignment #1 due**	
	Thurs. 4/20	PA2 released	**Programming assignment #2 (pa/pa2.pdf) released**	
Week 4	Tues. 4/25	Lecture (Pandu)	**Evaluation methods & NDCG** In-class lecture notes: • PPT (handouts/lecture8-evaluation_2014.ppt)	PDF/6 (handouts/lecture8-evaluation_2014-six-per-page.pdf)	PDF/1 (handouts/lecture8-evaluation_2014-one-per-page.pdf) Coursera content: • Videos: "Result Summaries" • Slides: PPT (handouts/lecture8-evaluation.ppt)	PDF/6 (handouts/lecture8-evaluation-handout-6-per.pdf)	PDF/1 (handouts/lecture8-evaluation-handout-1-per.pdf)	• IIR chapter 8 (http://nlp.stanford.edu/IR-book/pdf/08eval.pdf) • MG section 4.5 • MIR chapter 3
	Thurs. 4/27	Lecture (Pandu)	**Systems issues in efficient retrieval and scoring** In-class lecture notes: • PPT (handouts/lecture10-efficient-scoring.ppt)	PDF/6 (handouts/lecture10-efficient-scoring-six-per-page.pdf)	PDF/1 (handouts/lecture10-efficient-scoring-one-per-page.pdf)	• IIR chapter 6 (http://nlp.stanford.edu/IR-book/pdf/06vect.pdf) • IIR chapter 7 (http://nlp.stanford.edu/IR-book/pdf/07system.pdf) • Efficient Query Evaluation using a Two-Level Retrieval Process (Broder et al. 2003) (http://cis.poly.edu/westlab/papers/cntdstrb/p426-broder.pdf)
	Thurs. 4/27	PS1 due	**Problem set #1 due**	
	Thurs. 4/27	Ranking quiz released	Ranking quiz (https://web.stanford.edu/class/cs276/cgi-bin/queryrank/queryrank.php) released	
Week 5	Tues. 5/2	Guest lecture	**"Amazon Product Search: Search Relevance and Query Understanding"** *Guest lecture by Ravi Jammalamadaka and Erick Cantu-Paz (Engineering Managers, Amazon Search)* **NOTE: attendance required** for on-campus students	

	Thurs. 5/4	Lecture (Chris)	**Text classification (Naive Bayes, kNN, decision boundaries)** In-class lecture notes: • PPT (handouts/lecture12-textcat.pptx) | PDF/6 (handouts/lecture12-textcat-6up.pdf) | PDF/1 (handouts/lecture12-textcat.pdf) Coursera content: • Videos: "Naive Bayes"	• IIR chapter 13 (http://nlp.stanford.edu/IR-book/pdf/13bayes.pdf) • IIR chapter 14 (http://nlp.stanford.edu/IR-book/pdf/14vcat.pdf) • Reuters-21578 (http://www.daviddlewis.com/resources/testcollections/reuters21578/) • Machine learning in automated text categorization (Sebastiani 2002) (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.17.6513) • A re-examination of text categorization methods (Yang et al. 1999) (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.9519) • A Comparison of event models for naive Bayes text classification (McCallum et al. 1998) (http://citeseer.ist.psu.edu/mccallum98comparison.html) • Tackling the poor assumptions of Naive Bayes classifier (Rennie et al. 2003) (http://www.stanford.edu/class/cs276/handouts/rennie.icml03.pdf) • Machine learning in automated text categorization (Sebastiani 2002) (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.17.6513) • A re-examination of text categorization methods (Yang et al. 1999) (http://citeseer.ist.psu.edu/yang99reexamination.html) • Evaluating and optimizing autonomous text classification systems (Lewis 1995) (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.33.867) • Tom Mitchell. *Machine Learning*. McGraw-Hill, 1997. • Trevor Hastie, Robert Tibshirani, Jerome Friedman. *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York, 2001. • Open Calais • Weka
	Thurs. 5/4	PA2 due	**Programming assignment #2 due**	
	Thurs. 5/4	PA3 released	**Programming assignment #3 (pa/pa3.pdf) released**	
Week 6	Tues. 5/9	Lecture (Chris)	**Text classification (Support vector machines)** In-class lecture notes: • PPT (handouts/lecture13-SVMs.ppt) | PDF/6 (handouts/lecture13-SVMs-6up.pdf) | PDF/1 (handouts/lecture13-SVMs.pdf)	• IIR chapter 15 (http://nlp.stanford.edu/IR-book/pdf/15svm.pdf) • Reuters-21578 (http://www.daviddlewis.com/resources/testcollections/reuters21578/) • A tutorial on support vector machines for pattern recognition (Burges 1998) (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.1083) • Using SVMs for text categorization (Dumais 1998) (http://ezproxy.stanford.edu:2072/xpls/abs_all.jsp?isnumber=15361&arnumber=708428&count=14&index=5) • Inductive learning algorithms and representations for text categorization (Dumais et al. 1998) (http://portal.acm.org/citation.cfm?id=288651) • A Re-examination of text categorization methods (Yang et al. 1999) (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.30.5594) • Text categorization based on regularized linear classification methods (Zhang et al. 2001) (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.5553) • A loss function analysis for classification methods in text categorization (Li et al. 2003) (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.8.2881) • Trevor Hastie, Robert Tibshirani, Jerome Friedman. *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag,New York, 2001. • Thorsten Joachims. *Learning to Classify Text using Support Vector Machines*. Kluwer, 2002.
	Thurs. 5/11	Lecture (Chris)	**Learning to rank** In-class lecture notes: • PPT (handouts/lecture14-learning-ranking.pptx) | PDF/6 (handouts/lecture14-learning-ranking-6up.pdf) | PDF/1 (handouts/lecture14-learning-ranking.pdf)	• IIR sections 6.1.2-6.1.3 (http://nlp.stanford.edu/IR-book/pdf/06vect.pdf) • IIR section 15.4 (http://nlp.stanford.edu/IR-book/pdf/15svm.pdf) • LETOR benchmark datasets (http://research.microsoft.com/users/LETOR/) • Discriminative models for information retrieval (Nallapati 2004) (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.2.9065) • Adapting ranking SVM to document retrieval (Cao et al. 2006) (http://portal.acm.org/citation.cfm?id=1148205) • A support vector method for optimizing average precision (Yue et al. 2007) (http://portal.acm.org/citation.cfm?id=1277790)

	Thurs. 5/11	PS2 released	Problem set #2 (ps/ps2.pdf) released	[PS2 solution] (restricted/ps2-solution.pdf)
Week 7	Tues. 5/16	Lecture (Chris)	Distributed word representations for IR In-class lecture notes: • PPT (handouts/lecture20-distributed-representations.pptx) | PDF/6 (handouts/lecture20-distributed-representations-6up.pdf) | PDF/1 (handouts/lecture20-distributed-representations.pdf)	• Distributed Representations of Words and Phrases and their Compositionality (Mikolov et al., 2013) (http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf) • GloVe: Global Vectors for Word Representation (Pennington et al., 2014) (http://nlp.stanford.edu/pubs/glove.pdf)
	Thurs. 5/18	Guest lecture	Guest lecture by Jeff Dean (Google Senior Fellow) NOTE: attendance required for on-campus students	
	Thurs. 5/18	PA3 due	Programming assignment #3 due	
	Thurs. 5/18	PA4 released	Programming assignment #4 (pa/pa4.pdf) released	
Week 8	Tues. 5/23	Lecture (Pandu)	Link analysis In-class lecture notes: • PPT (handouts/lecture17-linkanalysis.ppt) | PDF/6 (handouts/lecture17-linkanalysis-six-per-page.pdf) | PDF/1 (handouts/lecture17-linkanalysis-one-per-page.pdf)	• IIR chapter 21 (http://nlp.stanford.edu/IR-book/pdf/21link.pdf) • Ranking the web frontier (Eiron et al. 2004) (http://portal.acm.org/citation.cfm?id=988714) • The WebGraph framework I: Compression techniques (Boldi et al. 2004) (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.6417) • Extrapolation methods for accelerating PageRank computations (Kamvar et al. 2003) (http://portal.acm.org/citation.cfm?id=775190) • Searching the workplace web (Fagin et al. 2003 (http://portal.acm.org/citation.cfm?id=775152.775204)
	Thurs. 5/25	In-class lab (Pandu)	MapReduce with Java (Hadoop setup instructions (https://docs.google.com/document/d/1i1Be9GDgS11cOD2JtlOUjTQT6Zenmr6UKB60D-mSaXs/edit)) (starter code (examples.tar.gz)) In-class lecture notes: • PDF/1 (handouts/mapreduce-pact06-keynote.pdf)	• IIR section 4.4 (http://nlp.stanford.edu/IR-book/pdf/04const.pdf) • Jeff Dean's MapReduce Slides (http://research.google.com/archive/mapreduce-osdi04-slides/index.html)
	Thurs. 5/25	PS2 due	Problem set #2 due	
Week 9	Tues. 5/30	Lecture (Pandu)	Crawling and near-duplicate pages In-class lecture notes: • PPT (handouts/lecture18-crawling.ppt) | PDF/6 (handouts/lecture18-crawling-6-per-page.pdf) | PDF/1 (handouts/lecture18-crawling-1-per-page.pdf)	• IIR chapter 19 (http://nlp.stanford.edu/IR-book/pdf/19web.pdf) • IIR chapter 20 (http://nlp.stanford.edu/IR-book/pdf/20crawl.pdf) • Mercator: A scalable, extensible web crawler (Heydon et al. 1999) (https://webarchive.jira.com/wiki/download/attachments/5441/1999-Mercator.pdf) • A standard for robot exclusion (http://www.robotstxt.org/orig.html)
	Thurs. 6/1	Lecture (Pandu)	Personalization In-class lecture notes: • PPT (handouts/personalization-lecture.ppt) | PDF/6 (handouts/personalization-lecture-6-per-page.pdf) | PDF/1 (handouts/personalization-lecture-1-per-page.pdf)	• J. Teevan, S. Dumais, E. Horvitz. Potential for personalization. 2010 (http://research.microsoft.com/pubs/154553/tochi10.pdf) • J. Pitkow et al. Personalized search. 2002 (http://www.cond.org/p50-pitkow.pdf) • J. Teevan, S. Dumais, E. Horvitz. Personalizing search via automated analysis of interests and activities. 2005 (http://research.microsoft.com/en-us/um/people/sdumais/SIGIR2005-PersonalizedSearch.pdf) • P. Bennett et al. Inferring and using location metadata to personalize Web search. 2011 (http://research.microsoft.com/pubs/150729/BennettSIGIR2011.pdf) • T. Haveliwala. Topic-sensitive pagerank. 2002. (http://www-cs-students.stanford.edu/~taherh/papers/topic-sensitive-pagerank.pdf) • G. Jeh and J. Widom. Scaling personalized Web search. 2003 (http://infolab.stanford.edu/~glenj/spws.pdf) • M. Curtiss et al. Unicorn: A system for searching the social graph. 2013 (http://www.vldb.org/pvldb/vol6/p1150-curtiss.pdf)
	Thurs. 6/1	PA4 due	Programming assignment #4 due	

Week 10	Tues. 6/6	Lecture (Chris)	**Question answering** In-class lecture notes: • PPT (handouts/lecture19-Web-QA.pptx) | PDF/6 (handouts/lecture19-Web-QA-6up.pdf) | PDF/1 (handouts/lecture19-Web-QA.pdf)	
Exam week	Mon. 6/12	Final exam	**Alternate final exam (7:00-10:00pm)**	
	Wed. 6/14	Final exam	**Final exam (3:30-6:30pm)**	Practice final (2016) (restricted/final_2016.pdf) Practice final solutions (restricted/final_2016_sol.pdf)