CS 410: Text Information Systems

Course Description

The growth of “big data” created unprecedented opportunities to leverage computational and statistical approaches, which turn raw data into actionable knowledge that can support various application tasks. This is especially true for the optimization of decision making in virtually all application domains, such as health and medicine, security and safety, learning and education, scientific discovery, and business intelligence. This course covers general computational techniques for building intelligent text information systems to help users manage and make use of large amounts of text data in all kinds of applications.

Text data include all data in the form of natural language text (e.g., English text or Chinese text), including all web pages, social media data such as tweets, news, scientific literature, emails, government documents, and many other kinds of enterprise data. Text data play an essential role in our lives. Since we communicate using natural languages, we produce and consume a large amount of text data every day covering all kinds of topics. The explosive growth of text data makes it impossible for people to consume all the relevant text data in a timely manner.

The two main techniques to assist people in consuming, digesting, and making use of the text data are:

1. Text retrieval, which helps identify the most relevant text data to a particular problem from a large collection of text documents, thus avoiding processing a large number of non-relevant documents
2. Text mining, which helps users further analyze and digest the found relevant text data and extract actionable knowledge for finishing a task

This course covers both text retrieval and text mining, so as to provide you with the opportunity to see the complete spectrum of techniques used in building an intelligent text information system. Building on two MOOCs covering the same topic and including a course project, this course enables you to learn the basic concepts, principles, and general techniques in text retrieval and mining, as well as gain hands-on experience with using software tools to develop interesting text data applications.

Course Goals and Objectives

Upon successful completion of this course, you will be able to:

- Explain all the basic concepts in text retrieval and text mining.
- Explain the main ideas behind the major models and algorithms for text retrieval and text mining.

- Explain how the major models and algorithms for text retrieval and text mining work.
- Explain how to implement some of the commonly used algorithms for text retrieval and text mining.
- Explain how to evaluate applications of text retrieval and text mining.

Textbook

There is not a required textbook for this course, but there are several optional readings suggested in each week's overview page. All readings listed in the weekly overview pages are optional and are primarily from the following textbook:

Zhai, C. & Massung, S. (2016). *Text data management and analysis: A practical introduction to information retrieval and text mining.* ACM Book Series. Morgan & Claypool Publishers.

Course Outline

Week	Dates	Topics
Week 1	August 28 - September 1	Part of Speech tagging, syntactic analysis, semantic analysis, ambiguity, “bag of words” representation, push, pull, querying, browsing, probability ranking principle, relevance, vector space model, dot product, bit vector representation
Week 2	September 4 - 10	Term frequency (TF), document frequency (DF), and inverse document frequency (IDF), TF transformation, pivoted length normalization, BM25, inverted index and postings, binary coding, unary coding, gamma-coding, d-gap, Zipf’s law
Week 3	September 11 - 17	Cranfield evaluation methodology, precision and recall, average precision, mean average precision (MAP), geometric mean average precision (gMAP), reciprocal rank, mean reciprocal rank, F-measure , Normalized Discounted Cumulative Gain (nDCG), statistical significance test
Week 4	September 18 - 24	p(R=1|q,d), query likelihood, p(q|d), statistical and unigram language models, maximum likelihood estimate, background, collection, and document

		language models, smoothing of unigram language models, relation between query likelihood and TF-IDF weighting, linear interpolation (i.e., Jelinek-Mercer) smoothing, Dirichlet Prior smoothing
Week 5	September 15 - October 1	Relevance feedback, pseudo-relevance feedback, implicit feedback, Rocchio feedback, Kullback-Leiber divergence (KL-divergence) retrieval function, mixture language model, scalability and efficiency, spams, crawler, focused crawling, and incremental crawling, Google File System (GFS), MapReduce, link analysis and anchor text, PageRank
Week 6	October 2 - 8	Content-based filtering, collaborative filtering, Beta-Gamma threshold learning, linear utility , user profile, exploration-exploitation tradeoff, memory-based collaborative filtering, cold start
Week 7	October 9 - 15	Text representation (especially bag-of-words representation), context of a word, context similarity, paradigmatic relation, syntagmatic relation, **Exam 1**
Week 8	October 16 - 22	Entropy, conditional entropy, mutual information, topics, coverage of topic , language model, generative model, unigram language model, word distribution, background language model, parameters of a probabilistic model, likelihood, Bayes rule, maximum likelihood estimation, prior and posterior distributions, Bayesian estimation & inference, maximum a posteriori (MAP) estimate, prior model, posterior mode
Week 9	October 23 - 29	Mixture model, component model, constraints on probabilities, Probabilistic Latent Semantic Analysis (PLSA), Expectation-Maximization (EM) algorithm, E-step and M-step, hidden variables, hill climbing, local maximum, Latent Dirichlet Allocation (LDA)
Week 10	October 30 - November 5	Clustering, document clustering, term clustering, clustering bias, perspective of similarity, Hierarchical Agglomerative Clustering, k-Means, direction evaluation

		(of clustering), indirect evaluation (of clustering), text categorization, topic categorization, sentiment categorization, email routing , spam filtering, naïve Bayes classifier, smoothing
Week 11	November 6 - 12	Generative classifier vs. discriminative classifier, training data, logistic regression, K-Nearest Neighbor classifier, classification accuracy, precision, recall, F measure, macro-averaging, micro-averaging, opinion holder, opinion target, sentiment, opinion representation, sentiment classification, features, n-grams, frequent patterns, overfitting
Week 12	November 13 - 19	Text-based prediction, the “data mining loop”, context (of text data), contextual text mining, contextual probabilistic latent semantic analysis (CPLSA), views of a topic, coverage of topics, spatiotemporal trends of topics, event impact analysis, network-regularized topic modeling, NetPLSA, causal topics, iterative topic modeling with time series supervision
Week 13	November 20 - 26	*Thanksgiving Break,* **Exam 2**
Week 14	November 27 - December 3	No new content - work on your final project! Presentation due end of Week 14
Week 15	December 4 - 10	No new content - work on your final project! Peer Reviews due end of Week 15
Week 16	December 11 - 14	Final Project Report Due Friday of Week 16

Elements of This Course

- **Lecture videos.** Each week your instructor will teach you the concepts you need to know through a collection of short video lectures. You may either stream these videos for playback within the

browser by clicking on their titles, or you can download each video for later offline playback by clicking the download icon. **The videos usually total 1.5 to 2 hours each week**, but you generally need to spend at least the same amount of time digesting the content in the videos. The actual amount of time needed to digest the content will naturally vary according to your background.

- **Quizzes.** Most weeks will include one for-credit quiz. You will have two attempts for each quiz, with your highest score used toward your final grade. Your top 10 quiz scores will be used to calculate your final grade (i.e., we will drop the two lowest quiz scores).
- **Exams**. This course will have two 2-hour exams. The exams are intended to test your understanding of the material you learn in the course and will contain questions similar to those seen in the weekly quizzes.
- **Programming Assignments.** The programming assignments for this course provide an opportunity for you to practice your programming skills and apply what you've learned in the course. Set aside about 2 hours each week to work on the programming assignment if you plan to finish it; you may need to budget more time for this if you are not familiar with C++ programming.
- **Final Course Project.** There will also be one culminating project due at the end of course. The first part will be individual and posted on Piazza, and the second part will require you to work in groups of at most three. The first part of the project will focus on simulation, while the second part will use methods from the course to perform data analysis and generate a written report.

Grading

Your final grade will be calculated based on the activities listed in the table below.

Activity	**Percent of Final Grade**
Quizzes	25%
Programming Assignments	25%
Course Project	20%
Exam 1	15%
Exam 2	15%

Letter Grade	Percent Needed	Letter Grade	Percent Needed	Letter Grade	Percent Needed
A+	95	**B+**	80	**C**	60
A	90	**B**	75	**D**	55
A-	85	**B-**	70	**F**	<55

Additional Course Policies

Student Code and Policies

A student at the University of Illinois at the Urbana-Champaign campus is a member of a University community of which all members have at least the rights and responsibilities common to all citizens, free from institutional censorship; affiliation with the University as a student does not diminish the rights or responsibilities held by a student or any other community member as a citizen of larger communities of the state, the nation, and the world. See the University of Illinois Student Code for more information.

The CS department also maintains a policies handbook for graduate student. For more information, see the Graduate Student Handbook.

Additionally, all Coursera learners are required to follow an Honor Code and a Code of Conduct. Please review both of these items before commencing your studies.

Academic Integrity

All students are expected to abide by the campus regulations on academic integrity found in the Student Code of Conduct. These standards will be enforced and infractions of these rules will not be tolerated in this course. Sharing, copying, or providing any part of a homework solution or code is an infraction of the University's rules on academic integrity. We will be actively looking for violations of this policy in homework and project submissions. Any violation will be punished as severely as possible with sanctions and penalties typically ranging from a failing grade on this assignment up to a failing grade in the course, including a letter of the offending infraction kept in the student's permanent university record.

Again, a good rule of thumb: *Keep every typed word and piece of code your own*. If you think you are operating in a gray area, you probably are. If you would like clarification on specifics, please contact the course staff.

Disability Accommodations

Students with learning, physical, or other disabilities requiring assistance should contact the instructor as soon as possible. If you're unsure if this applies to you or think it may, please contact the instructor and Disability Resources and Educational Services (DRES) as soon as possible. You can contact DRES at 1207 S. Oak Street, Champaign, via phone at (217) 333-1970, or via email at disability@illinois.edu.

Late Policy

Late homework and homework by email will not be accepted by the TA or the instructors without prior instructor approval.

Course Deadlines

Quizzes

Assignment	Release Date	Hard Deadline
Quiz 1	First day of class	End of Week 1
Quiz 2	First day of class	End of Week 2
Quiz 3	First day of class	End of Week 3
Quiz 4	First day of class	End of Week 4
Quiz 5	First day of class	End of Week 5
Quiz 6	First day of class	End of Week 6

Quiz 7	First day of class	End of Week 7
Quiz 8	First day of class	End of Week 8
Quiz 9	First day of class	End of Week 9
Quiz 10	First day of class	End of Week 10
Quiz 11	First day of class	End of Week 11
Quiz 12	First day of class	End of Week 12

Programming Assignments

Assignment	Release Date	Hard Deadline
Programming Assignment 1	First day of class	End of Week 3
Programming Assignment 2	First day of class	End of Week 6
Programming Assignment 3	First day of class	End of Week 9

Course Project

Assignment	Release Date	Hard Deadline
Course Project Proposal	First day of class	End of Week 9

Course Project Presentation Peer Review Submission	First day of class	End of Week 14
Course Project Presentation Peer Review	First day of class	End of Week 15
Course Project Presentation and Report submission	First day of class	End of Week 16

Exams

Exam Name	Exam Start Date	Exam End Date
Exam 1	Mon., Oct 9, 12:15am CT	Sat., Oct 14, 10:30pm CT
Exam 2	Mon., Nov 20, 12:15am CT	Sat., Nov 25, 10:30pm CT

Late Policy

- Unless otherwise specified, all assignments are **due at 11:59 p.m. US Central Time on the due date**. (Time Zone Converter)
- The hard deadline is the deadline after which you will receive 0 points on assignments regardless how well you did on the assignment. No late submission will be accepted except under extremely rare non-academic circumstances (which usually require approval from the Dean's office).

Academic Calendar

- The Graduate College at the University of Illinois maintains a Graduate College Calendar. The calendar includes important dates such as final exam dates, course registration and cancellation, and holidays.
- There is also a campus wide calendar available.
- The CS Department also sends reminders about upcoming deadlines. You will also receive the Graduate College newsletter in your Exchange email account.