Course Description

Recent years have seen a **dramatic growth** of natural language **text data**, including web pages, news articles, scientific literature, emails, enterprise documents, and social media such as blog articles, forum posts, product reviews, and tweets. This has led to an increasing demand for powerful software tools to help people analyze and manage vast amounts of text data effectively and efficiently. Unlike data generated by a computer system or sensors, text data are usually generated directly by humans and are accompanied by semantically rich content. As such, text data are **especially valuable for discovering knowledge about people's opinions and preferences**, in addition to many other kinds of knowledge that we encode in text. However, in contrast to structured data, which conform to well-defined schemas, and are thus relatively easy for computers to handle, text has less explicit structure, thus requiring computer processing toward understanding of the content encoded in text. The current technology of natural language processing has not yet reached a point to enable a computer to precisely understand natural language text, but a wide range of statistical and heuristic approaches to mining and analysis of text data have been developed over the past few decades. They are usually **very robust** and can be applied to analyze and manage text data in **any natural language** and about **any topic**.

This course provides an introduction to some of these approaches with an **emphasis on approaches that do not require (much) manual effort**, including those for mining word associations, mining and analyzing topics in text, clustering and categorizing text data, opinion mining and sentiment analysis, and joint analysis of text and non-textual data. You will learn the **most useful basic concepts, principles, and techniques** in text mining and analytics that can be applied to build **a wide range of text mining and analytics application systems**.

Course Goals and Objectives

By the end the course, you will be able to do the following:

- Explain many basic concepts and multiple major algorithms in text mining and analytics.
- Explain how statistical language models, particularly topic models, can be applied to arbitrary text data to discover and analyze topics in text.
- Implement some text mining and analytics algorithms, run text mining experiments, and experiment with ideas on a real text mining task to improve the mining results (if you complete the programming assignment).

Course Outline

The course consists of **6 weekly modules**. Please note: There are no required readings for this course. All readings listed below are optional.

Week 1

Key Concepts:

- Part of speech tagging

- Syntactic analysis
- Semantic analysis
- Ambiguity
- Text representation, especially bag-of-words representation
- Context of a word; context similarity
- Paradigmatic relation
- Syntagmatic relation

Recommended Readings:

- C. Zhai and S. Massung, *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining.*ACM and Morgan & Claypool Publishers, 2016. Chapters 1-4, Chapter 13.
- Chris Manning and Hinrich Schütze, *Foundations of Statistical Natural Language Processing.* MIT Press. Cambridge, MA: May 1999. (Chapter 5 on collocations)
- Chengxiang Zhai, *Exploiting context to identify lexical atoms: A statistical view of linguistic context*. Proceedings of the International and Interdisciplinary Conference on Modelling and Using Context (CONTEXT-97), Rio de Janeiro, Brazil, Feb. 4-6, 1997. pp. 119-129.
- Shan Jiang and ChengXiang Zhai, *Random walks on adjacency graphs for mining lexical relations from big text data*. Proceedings of IEEE BigData Conference 2014, pp. 549-554.

Week 2

Key Concepts:

- Entropy
- Conditional entropy
- Mutual information
- Topic and 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

Recommended Readings:

- C. Zhai and S. Massung, *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining.*ACM and Morgan & Claypool Publishers, 2016. Chapters 13, 17

Week 3

Key Concepts:

- 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)

Recommended Readings:

- C. Zhai and S. Massung, *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. ACM and Morgan & Claypool Publishers, 2016. Chapter 17.
- Blei, D. 2012. *Probabilistic Topic Models*. Communications of the ACM 55 (4): 77–84. doi: 10.1145/2133806.2133826.
- Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. *Automatic Labeling of Multinomial Topic Models*. Proceedings of ACM KDD 2007, pp. 490-499, DOI=10.1145/1281192.1281246.
- Yue Lu, Qiaozhu Mei, and Chengxiang Zhai. 2011. *Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA*. Information Retrieval, 14, 2 (April 2011), 178-203. doi: 10.1007/s10791-010-9141-9.

Week 4

Key Concepts:

- Clustering, document clustering, and term clustering
- Clustering bias
- Perspective of similarity
- Mixture model, likelihood, and maximum likelihood estimation
- EM algorithm, E-step, M-step, underflow, normalization (to avoid underflow)
- Hierarchical Agglomerative Clustering, and k-Means
- Direction evaluation (of clustering), indirect evaluation (of clustering)
- Text categorization, topic categorization, sentiment categorization, email routing

- Spam filtering
- Naïve Bayes classifier
- Smoothing

Recommended Readings:

- C. Zhai and S. Massung, *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. ACM and Morgan & Claypool Publishers, 2016. Chapters 14 & 15.
- Manning, Chris D., Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2007. (Chapters 13-16)
- Yang, Yiming. *An Evaluation of Statistical Approaches to Text Categorization*. Inf. Retr. 1, 1-2 (May 1999), 69-90. doi: 10.1023/A:1009982220290

Week 5

Key Concepts:

- Generative classifier vs. discriminative classifier
- Training data
- Logistic regression
- K-Nearest Neighbor classifier
- Support Vector Machine (SVM), margin, and linear separator
- Classification accuracy, precision, recall, F measure, macro-averaging, and micro-averaging
- Opinion holder, opinion target, sentiment, opinion representation
- Sentiment classification
- Features, n-grams, frequent patterns, and overfitting
- Ordinal logistic regression
- Rating prediction

Recommended Readings:

- C. Zhai and S. Massung, *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining.*ACM and Morgan & Claypool Publishers, 2016. Chapters 15 & 18
- Yang, Yiming. An Evaluation of Statistical Approaches to Text Categorization. Inf. Retr. 1, 1-2 (May 1999), 69-90. doi: 10.1023/A:1009982220290
- Bing Liu, *Sentiment analysis and opinion mining.* Morgan & Claypool Publishers, 2012.
- Bo Pang and Lillian Lee, *Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieva*l 2(1-2), pp. 1–135, 2008.

Week 6

Key Concepts:

- Aspect rating and aspect weight

- Latent aspect rating analysis (LARA)
- Latent rating regression model
- Generative model
- Rating prediction
- Normal/Gaussian distribution
- Prior vs. posterior probability
- Text-based prediction
- The “data mining loop”
- Context (of text data) and contextual text mining
- Contextual probabilistic latent semantic analysis (CPLSA): views of a topic and coverage of topics
- Spatiotemporal trends of topics
- Event impact analysis
- Network-regularized topic modeling
- NetPLSA
- Causal topics
- Iterative topic modeling with time series supervision

Recommended Readings:

- C. Zhai and S. Massung, *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining.*ACM and Morgan & Claypool Publishers, 2016. Chapters 18 & 19
- Hongning Wang, Yue Lu, and ChengXiang Zhai, *Latent aspect rating analysis on review text data: a rating regression approach*. In Proceedings of ACM KDD 2010, pp. 783-792, 2010. doi: 10.1145/1835804.1835903
- Hongning Wang, Yue Lu, and ChengXiang Zhai. 2011. *Latent aspect rating analysis without aspect keyword supervision*. In Proceedings of ACM KDD 2011, pp. 618-626. doi: 10.1145/2020408.2020505
- ChengXiang Zhai, Atulya Velivelli, and Bei Yu. *A cross-collection mixture model for comparative text mining*. In Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2004). ACM, New York, NY, USA, 743-748. doi: 10.1145/1014052.1014150
- Qiaozhu Mei, Contextual Text Mining, Ph.D. Thesis, University of Illinois at Urbana-Champaign, 2009.
- Hyun Duk Kim, Malu Castellanos, Meichun Hsu, ChengXiang Zhai, Thomas Rietz, and Daniel Diermeier. *Mining causal topics in text data: Iterative topic modeling with time series feedback*. In Proceedings of the 22nd ACM international conference on information & knowledge management (CIKM 2013). ACM, New York, NY, USA, 885-890. doi: 10.1145/2505515.2505612
- Noah Smith, *Text-Driven Forecasting*. Retrieved May 31, 2015 from http://www.cs.cmu.edu/~nasmith/papers/smith.whitepaper10.pdf

Elements of This Course

The course is comprised of the following elements:

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 each week usually total 1.5 to 2 hours, but you generally need to spend at least the same amount of time digesting the content in the video. The actual amount of time needed to digest the content would naturally vary according to your background.

Quizzes. Each week will include one for-credit quiz. Your cumulative score will be used when calculating your final score in the class. There is no time limit on how long you take to complete each quiz. The deadline for all quizzes is the last day of the course.

The weekly quiz should take less than 1 hour to finish, assuming you have mastered the materials to be tested in the quiz, and you should make sure that you have mastered the materials before you attempt to work on the quizzes. You should use the practice quizzes to help you understand and master the materials.

Programming assignment. The programming assignment for this course is optional, but it provides an opportunity for you to practice your programming skills and apply what you've learned in the course. Plan about 2 hours each week to work on it if you plan to finish it; you may need to budget more time for this if you are not familiar with C++ programming.

Information about Lectures

The lectures in this course contain the most important information you need to know. You can access these lectures in each week's lesson section. The following resources accompany each video:

- The play button will open the video up in your browser window and stream the lecture to you. The duration of the video (in hours-minutes-seconds format) is also listed.
- English subtitles are available for all videos.All video lectures have a discussion forum dedicated to them. This is a great place to discuss any questions you have about the content of the video or to share your ideas and responses to the video.

Discussion Forums

The discussion forums are a key element of this course. Be sure to read more **about the discussion forums** and how you can make the most of them in this class.

How to Pass the Course

I am continually looking to improve this course and may encounter some issues requiring us to make changes sooner rather than later. As such, this syllabus is subject to change. I appreciate your input and ask that you have patience as we make adjustments to this course. I also recognize that this is no ordinary course. You may have different perspectives and different goals for this course than some of your peers, or that I could have anticipated. Therefore, I want to empower you to customize this course to meet your needs.

To qualify for a Course Certificate, simply start verifying your coursework at the beginning of the course, get an **70% or higher** on all quizzes and assignments combined, and pay the fee. Coursera Financial Aid is available to offset the registration cost for learners with demonstrated economic needs. If you have questions about Course Certificates, please see the help topics here.

Also note that this course is in the Data Mining Specialization offered by the University of Illinois at Urbana-Champaign. By earning a Course Certificate in this course, you are on your way toward earning a Specialization Certificate in Data Mining. You may also choose to pre-pay for the entire Specialization, at a discount. See more information about Specialization payments here.

If you choose not to pay the fee, you can still audit the course. You will still be able to view all videos, submit practice quizzes, and view required assessments. Auditing does not include the option to submit required assessments. As such, you will not be able to earn a grade or a Course Certificate.

Getting and Giving Help

You can get/give help via the following means:

- Use the **Learner Help Center** to find information regarding specific technical problems. For example, technical problems would include error messages, difficulty submitting assignments, or problems with video playback. If you cannot find an answer in the documentation, you can also report your problem to the Coursera staff by clicking on the **Contact Us!** link available on each topic's page within the Learner Help Center.
- Use the **Content Issues** forum to report errors in lecture video content, assignment questions and answers, assignment grading, text and links on course pages, or the content of other course materials. University of Illinois staff and Community Mentors will monitor this forum and respond to issues.

Note: Due to the large number of learners enrolled in this course, I am not able to answer emails sent directly to my account. Rather, all questions should be reported as described above.