Syllabus

Help Center

Text Mining and Analytics

On this page:

Course Description Course Goals and Objectives Textbook Course Outline

Multiple Ways to Complete This Course Elements of This Course Discussion Forums

Getting and Giving Help

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 amount 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).

Recommended Reading

Please note: There are no required textbooks or required readings for this course. The following are recommendations only.

- Manning, Chris D., and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1999. (Chapters 1, 5, 6, 8, 14, & 16)
- Manning, Chris D., Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2007. (Chapters 13-18)
- Aggarwal, Charu, and ChengXiang Zhai, eds. *Mining Text Data*. New York: Springer, 2012. (Chapters 1, 4, 6, 12, 13)
- Zhai, ChengXiang, and Sean Massung. *Text Data Analysis and Management: A Practical Introduction to Text Mining and Information Retrieval*. San Francisco: Morgan & Claypool Publishers, forthcoming.

Course Outline

The course consists of 4 weekly modules, each of which will be released to you shortly before the module begins.

Module	Key Concepts	Recommended Readings
Week 1	• Overview of text mining and analytics • NLP and text representation • Paradigmatic and syntagmatic word relations • Mining word associations	• Manning, Chris and Hinrich Schütze. *Foundations of Statistical Natural Language Processing.* Cambridge: MIT Press, 1999. (Chapter 5) • Zhai, ChengXiang. "Exploiting Context to Identify Lexical Atoms: A Statistical View of Linguistic Context."*Proceedings of the International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT-97)*, Rio de Janeiro, Brazil, 4-6 Feb. 1997. (pages 119-129) • Jiang, Shan, and ChengXiang Zhai. "Random Walks on Adjacency Graphs for Mining Lexical Relations from Big Text Data." *Big Data Conference*. 27-30 Oct. 2014, Washington DC. Washington DC: IEEE International Conference on Big Data, 2014. 549-554.
	• Overview of topic mining and analysis • One topic per document: document clustering for topic mining • Multiple topics	• Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schuetze. *Introduction to Information Retrieval*. Cambridge: Cambridge UP, 2007. (Chapters 16 and 17) • Aggarwal, Charu, and ChengXiang Zhai, eds. *Mining Text Data*. New York: Springer, 2012.

Week 2	per document: statistical topic models for topic mining • Probabilistic Latent Semantic Analysis • EM algorithm	(Chapter 4) • Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. *Automatic Labeling of Multinomial Topic Models*. Proceedings of ACM KDD, 2007. (pages 490-499)
Week 3	• Incorporating prior into a topic model • Text categorization • Sentiment analysis	• Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schuetze. *Introduction to Information Retrieval*. Cambridge: Cambridge UP, 2007. (Chapters 13 and 14) • Aggarwal, Charu, and ChengXiang Zhai, eds. *Mining Text Data*. New York: Springer, 2012. (Chapter 6) • Liu, Bing. *Sentiment Analysis and Opinion Mining.* Morgan & Claypool Publishers, 2012.
Week 4	• Joint analysis of text and non-textual data • Advanced topic models • Towards a general text analysis engine • Summary	• Blei, D. “Probabilistic Topic Models.” *Communications of the ACM.* 55.4 (2012): 77–84. • Zhai, ChengXiang. *Statistical Language Models for Information Retrieval.* Morgan & Claypool Publishers, 2008. (Chapter 7) • Kim, Hyun Duk, Malú Castellanos, Meichun Hsu, ChengXiang Zhai, Thomas A. Rietz, and Daniel Diermeier. "Mining Causal Topics in Text Data: Iterative Topic Modeling with Time Series Feedback." *Proceedings of CIKM,* 2015. (pages 885-890) • Smith, Noah. *Text-Driven Forecasting.* 31 May 2015.

Multiple Ways to Complete This Course (And Multiple Potential Benefits)

I am continually looking to improve this course, and we 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.

This course offers a free, no-risk Signature Track Trial. To qualify for a Verified Certificate, simply start verifying your coursework at the beginning of the course (with no upfront charges), and pay the $49 Signature Track registration fee anytime before the last week of the course. You can delay payment until you’re confident you’ll pass. Coursera Financial Aid is available to offset the registration cost for students with demonstrated economic need. If you have questions about this trial, please see the help topics here.

If you choose not to verify your work, you can still participate in the complete course. While your final grade will be recorded on your Course Records page, **this course will not offer a Statement of Accomplishment.** You will however, still receive any badges you earn, as described below.

I recognize that this is no ordinary course. You may have different perspectives and different goals for this course than some of your peers or than I could have anticipated. Therefore, I want to empower you to customize this course to meet your needs. To this end, we have designed multiple "badges" you can earn through participation in this course. You need to only earn the Course Achievement Badge to pass the course and apply towards a Verified Certificate from Coursera marking successful completion, though I encourage you to earn as many badges as possible as each badge provides unique benefits to you, as described in detail on the How to Pass the Class page and summarized below.

	What It's Called	How It's Earned	What You Get
	Course Achievement Badge	Total score of 70% or higher for all quizzes combined	• Qualify for a Verified Certificate • Course Achievement Badge
	Course Mastery Badge	Total score of 90% or higher for all quizzes combined	• Qualify for a Verified Certificate • Course Mastery Badge • Inclusion in the Quiz Mastery Hall of Fame
	Programming Achievement Badge	70% or higher average score on programming assignments	• Programming Achievement Badge
	Programming Mastery Badge	90% or higher average score on programming assignments	• Programming Mastery Badge
	Text Mining Competition Leader Badge	Top 30% on Programming Competition Leaderboard	• Text Mining Competition Leader Badge

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.

- **Quizzes.** Each week will include one practice quiz and one for-credit quiz. You will be allowed **one** attempt at the for-credit quiz. Thus, we have created a practice quiz you can take before taking the for-credit quiz. You may take the practice quiz as many times as you would like. Your for-credit quiz scores will be used when calculating your final score in the class. There is no time limit on how long you take to complete the for-credit quiz. All quizzes must be completed by the end of the course.
- **Programming Assignments.** An optional programming assignment is offered (with corresponding badges) to provide an opportunity for you to gain hands-on experience with implementing some text mining algorithms and experimenting with them on a real world data set. It also contains a text mining competition task that allows you to explore any of your own ideas for improving the performance of a text-based prediction task. We will establish a leaderboard for this prediction task and grant Text Mining Competition Leader Badges to the top 30% submissions on the leaderboard. Note that we have made the programming assignment optional, rather than required, not because it is not important, but because we suspect that some of you may not necessarily have the needed computing resources to complete the programming assignment and hope to accommodate as many people as possible to pass this course. Indeed, working on the programming assignment is the best way of digesting and applying the knowledge that you will learn in this course. We thus strongly encourage you to work on the programming assignment if you have the resources and extra time.

Information About Lectures

The lectures in this course contain the most important information you need to know. You can access these lectures via the **All Videos** link in the main menu or via the weekly overview pages (preferred). 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. Within the player that appears, you can click the **CC** button to activate closed captions. English captions are available for all videos. In some cases, the captions have been translated by your peers into other languages and made available to you. Learn more about translating captions into other languages.
- 🖼The Lecture Notes or Lecture Slides provide you with a reference of the key points raised in the lecture.
- ≡The Transcript provides you with the text of the speaker's words. It is provided in English only.
- ⬇The Download link allows you to download a copy of the file in MP4 format (which most video player software can handle). This option may be useful if you are on a slower Internet connection or prefer to view the videos when not connected to the Internet. Each file is automatically numbered in the order it appears in the course and includes the duration (in hours-minutes-seconds format) in the file name as well.
- ☰If you choose to download the video, you may optionally wish to download the closed-caption SRT file to accompany it. Consult your video player's documentation on how to load the SRT file with your video. SRT files are only available in English.
- 💬Most videos 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 an important 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.

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. You can access the Help Center by clicking on the **Help Center** link at the top right of any course page. If you can not 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 TAs will monitor this forum and respond to issues.

Note: Due to the large number of students enrolled in this course, the instructor is not able to answer emails sent directly to his account. Rather, all questions should reported as described above.

Created Fri 13 Jan 2012 9:33 PM PST

Last Modified Fri 5 Jun 2015 11:30 AM PDT