

Introduction to Text Mining

Online course learning objectives

This course gives a foundational understanding of the analysis of textual data sets from social media sites, digital archives, and digital surveys and interviews through the study of language and social interactions in digital environments. This course is perfect for social scientists who want to gain a conceptual overview of the text mining landscape to take first steps towards working on a text mining project or collaborating with computational colleagues.

This course will help learners to:

- Learn the foundations of Natural Language Processing (NLP).
- Learn how text mining tools have been used successfully by social scientists.
- Understand basic text processing techniques.
- Understand how to approach narrative analysis, thematic analysis, and metaphor analysis.
- Learn about key computer science methods for text mining, such as text classification and opinion mining.

Language: English

Time to complete: 10 hours

Instructors: Rada Mihalcea and Gabe Ignatow

Online course full syllabus

MODULE ONE: FOUNDATIONS

In this module you will learn about the sources of internet-based textual data, and the key ethical and philosophical decisions that must be taken before acquiring data from human subjects.

Lesson one: Acquiring and analyzing texts

- Differentiate between text mining and text analysis methodologies.
- Compare major theoretical and methodological approaches to both text mining and text analysis.
- Identify a variety of different data sources used to compile text mining data sets.
- Assess the advantages and limitations of using social media to acquire data.
- Analyze examples of social research using data sets drawn from different sources.

Introduction to Text Mining

Lesson two: Ethical and philosophical issues

- Identify appropriate ethical guidelines for your research project.
- Decide whether your research needs approval from an institutional review board (IRB).
- Define major philosophy of social science concepts that are relevant to the practice of text mining.
- Recognize the interdependence of philosophical assumptions and decisions about methodology.

MODULE TWO: RESEARCH DESIGN AND BASIC TOOLS

In this module you will learn how to design a social science study centered on internet-based textual data and gain a fundamental understanding of scraping textual data from websites.

Lesson three: Research designs for text mining

- Compare various research designs for qualitative, quantitative, and mixed methods research.
- Determine where the most critical decisions are made in text mining research.
- Reference influential text mining studies that employ a variety of research designs.

Lesson four: Web scraping and crawling

- Define the main techniques for web crawling.
- Explore available software packages for automatically collecting textual data from webpages.
- Compare web crawling and web scraping techniques.
- Compare tools and supporting material available for web crawling and scraping techniques.

MODULE THREE: TEXT MINING FUNDAMENTALS

In this module you will learn fundamental Natural Language Processing (NLP) procedures and gain familiarity with online resources available for using NLP for social research.

Lesson five: Lexical resources

- Understand the different types of lexical resources.
- Describe the representation and content of several lexical resources.
- Discuss the role played by different lexical resources in text mining.
- Identify publicly available lexical resources and software packages used to access them.

Lesson six: Basic text processing

- Define basic text processing steps, such as tokenization, stop word removal, stemming, and lemmatization.
- Explain text statistics and laws that govern the distribution of words in text.
- Explore the basics of language models and evaluate their applications.
- Discuss the main goals of more advanced text processing steps.

Introduction to Text Mining

Lesson seven: Supervised learning

- Recognize the task of supervised learning and its large range of applications.
- Define feature representation and weighting in supervised learning, and classify specific learning algorithms.
- Evaluate supervised learning methods.
- Explore available software packages for supervised learning.

MODULE FOUR: METHODS FROM THE HUMANITIES AND SOCIAL SCIENCES

In this module you will learn the main approaches to text mining and text analysis that have been applied successfully in social science research.

Lesson eight: Narrative analysis

- Explore the foundational role of narrative in human thought and communication.
- Summarize theories of narrative from multiple disciplines.
- Examine what has been learned about narrative from social science research.
- Discover tools for conducting narrative analysis.

Lesson nine: Thematic analysis

- Explain how human communication is structured by themes.
- Describe thematic analysis techniques from the social sciences and humanities.
- Evaluate various approaches to thematic analysis.

Lesson ten: Metaphor analysis

- Learn the basic concepts of cognitive metaphor theory (CMT).
- Theorize why social scientists analyze metaphorical language and summarize different approaches.
- Recognize the benefits and drawbacks of different strategies for analyzing metaphor.

MODULE FIVE: COMPUTER SCIENCE METHODS

In this module you will learn about several advanced text mining methods.

Lesson eleven: Text classification

- Explore the foundational role of narrative in human thought and communication.
- Summarize theories of narrative from multiple disciplines.
- Examine what has been learned about narrative from social science research.
- Discover tools for conducting narrative analysis.

Introduction to Text Mining

Lesson twelve: Opinion mining

- Define the task of opinion mining and explore its applications.
- Compare the different lexical resources and textual collections typically used for opinion mining.
- Summarize the main approaches to automatic opinion mining.
- Locate online opinion mining resources and tools.

Lesson thirteen: Information extraction

- Define the task of information extraction (IE) and its applications.
- Explain entity and relation extraction.
- Familiarize yourself with more advanced topics, such as web IE and template filling.
- Learn about existing software and data sets for IE.

Lesson fourteen: Topic models

- Describe the theoretical foundations of topic modelling methods.
- Understand the statistical methods used in topic models.
- Discover how social scientists have used topic models in their research.