

# Text & Web Mining with RapidMiner

## Course Overview

Text & Web Mining with RapidMiner is a two-day introductory course into knowledge discovery using unstructured data like text documents and data sourced from the internet. It focuses on the necessary preprocessing steps and the most successful methods for automatic text machine learning including: Naive Bayes, Support Vector Machines (SVM), k-NN, and clustering.

After successfully completing this course, participants will have a solid understanding of how RapidMiner Studio supports text and web mining. Participants will be able to identify techniques for processing unstructured data, apply different statistical text-processing methods, perform text classification & clustering, source data from the web and prepare it for analysis, and will be ready to extend their knowledge to other advanced topics such as RapidMiner Server - Web Apps and Deployment, and Big Data Analytics with RapidMiner Radoop.

Practical exercises during the course prepare students to take the knowledge gained and apply it to their own text and web mining challenges. Examples include: document similarity clustering, sentiment analysis of text documents like news, web reviews, blogs, e-mail, or PDF documents, and predictive modeling using text data and metadata. The class exercises and labs are hands-on and performed on the participants' personal laptops, so students will internalize the topics covered, which will provide a jumpstart to the real-world application of these techniques.

## Target Audience

Advanced Analysts and Data Scientists

## Prerequisites

Basic knowledge of computer programs and mathematics  
RapidMiner Basics Part 1  
RapidMiner Basics Part 2

## Course Objectives

After the training, students will have the ability to:

- Identify techniques for processing unstructured data
- Transform textual data into a structured format and perform necessary pre-processing
- Apply different statistical text-processing methods

- Perform text classification and text clustering
- Work on popular tasks like sentiment analysis or opinion mining
- Source data from websites and process it for further analysis

## Course Outline

- **Loading of Texts**

- ◊ Loading from Flat Files
- ◊ Loading from Data Sets
- ◊ Loading from Databases
- ◊ Loading from Web Sources (e.g. URL crawling, Twitter)

- **Concepts**

- ◊ Text Processing
- ◊ Documents
- ◊ Tokens

- **Visualization**

- ◊ Visualizing Documents and Tokens
- ◊ High Dimensional Visualizations for Transformed Documents

- **Handling Unstructured Data**

- ◊ Preprocessing of Textual Data
- ◊ Tokenizing
- ◊ Stemming
- ◊ Filtering of Tokens
- ◊ Term Frequencies
- ◊ Document Frequencies
- ◊ TF-IDF

- **Advanced Modeling**

- ◊ Support Vector Machines
- ◊ Naive Bayes, k-NN
- ◊ Text Classification
- ◊ Text Clustering

- **Web Mining**

- ◊ Crawling the Web
- ◊ Extracting Information from Web Sites
- ◊ Transforming Web Sites to Documents