Course Overview
Text Mining with RapidMiner is a one day course and is an introduction into knowledge discovery using unstructured data like text documents. It focuses on the necessary preprocessing steps and the most successful methods for automatic text classification including: Naive Bayes, Support Vector Machines (SVM), and text clustering.

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

Practical exercises during the course prepare students to take the knowledge gained and apply to their own text mining challenges. Examples include: e-mail spam detection, automatic e-mail routing, adaptive personal news filtering, sentiment analysis of text documents like news, web pages, blogs, e-mail, or PDF documents. Since the class labs are hands-on and performed on the participants’ personal laptops, students will take actual classwork home with them, which will provide a jumpstart to the real world.

Target Audience
Analysts, Developers, and Administrators

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
- Apply different statistical text-processing methods
- Perform text classification and text clustering
- Work on recent tasks like sentiment analysis or opinion mining

**Course Outline**

- **Loading of Texts**
  - Loading from Flat Files
  - Loading from Data Sets
  - Loading from Databases
  - Loading from Web Sources (e.g. Twitter)

- **Concepts**
  - 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**
  - Methods for High Dimensional Data
  - Support Vector Machines
  - Text Classification
  - Text Clustering

- **Web Mining**
  - Crawling the Web
  - Extracting Information from Web Sites
  - Transforming Web Sites to Documents
  - Information Extraction using XPath or Regular Expressions