Text Mining with RapidMiner

Course Overview

Text Mining with RapidMiner is a one-day expert seminar regarding data science techniques for manipulating and using unstructured text data from documents, emails, and other textual sources. It focuses on the necessary preprocessing steps and successful methods for automatic text machine learning such as Naive Bayes, k-Nearest Neighbors (k-NN), and Support Vector Machines (SVM) as well as LDA topic modeling.

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, prepare documents through pre-processing, apply different statistical text-processing methods, and perform machine learning techniques on text data. Practical exercises during the course prepare students to take the knowledge gained and apply it to their own text mining challenges. Examples include: predictive modeling using text data and metadata, LDA topic modeling, and sentiment analysis of text documents from news, product reviews, or other similar documents. The class exercises and labs are hands-on, so students will internalize the topics covered, which will provide a jumpstart to the real world application of these techniques.

Prerequisites & Target Audience

This class is aimed at Analysts and Data Scientists. It assumes a basic knowledge of computer programming principles and higher mathematics (through calculus). It also requires either the successful completion of the basic-level training courses (RapidMiner & Data Science: Foundations and RapidMiner & Data Science: Advanced) or successful completion of the RapidMiner Analyst Certification exam (or functional equivalence in terms of knowledge of RapidMiner and basic data science).

Course Outline

- Introduction to Text Mining
- Loading of Documents (from flat files & datasets)
- Text Preprocessing & Handling Unstructured Text Data
    - Tokenizing
    - Filtering of Tokens
    - Stemming & n-grams
- Text Processing & Word Vectors
    - Term & Document Frequencies
    - TF-IDF
    - Basic Text Visualizations
- Modeling with Textual Data
    - Support Vector Machines
    - Naive Bayes
    - k-NN
    - LDA topic modeling