Quandatics

Big Data, IoT, Smart Analytics

www.quandatics.com

rapidminer

# Text, Web, and Social Media Mining

## Course Overview

This Text, Web, and Social Media Mining course is a 2 days course and is an introduction into knowledge discovery using unstructured data like text documents, web and social media contents. It focuses on the necessary preprocessing steps and the most successful methods for automatic text classification including: Naive Bayes, Support Vector Machines (SVM), and clustering. Hands-on exercises will be carried out using RapidMiner Studio, hence there will be an introduction section on the first day to help participants familiarize with the environment.

Upon completion of this course, participants will have a solid understanding of typical text mining workflows and be able to identify techniques for processing unstructured data, apply different statistical text-processing methods, and perform content classification & clustering.

Practical exercises during the course prepare students to take the knowledge gained and apply to their own text mining challenges. Examples include: adaptive personal news filtering, patent clustering, 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.

## 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 popular tasks like sentiment analysis or opinion mining

# Quandatics

Big Data, IoT, Smart Analytics

www.quandatics.com

## Course Outline

- **Introduction to RapidMiner Studio**

- **Loading of Texts**
  - Loading from Flat Files
  - Loading from Data Sets
  - Loading from Databases
  - Loading from Web Sources (e.g. Web Pages, 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**
  - Fetching data from Twitter
  - Crawling the Web
  - Extracting Information from Web Sites
  - Transforming Web Sites to Documents

For any enquiries, please contact:
**Quandatics**
e: contact@quandatics.com
m: +6 016 223 9422
( Tess Tan)