Web Mining with RapidMiner

Course Overview

Web Mining with RapidMiner is a one-day expert seminar regarding data science techniques for retrieving and using various types of data from the internet, including basic websites (extracting text from html pages), structured web tables, and other web data sources. It focuses on the necessary processing steps for extracting useful content from many different types of web data as well as typical machine learning techniques applied to such data.

After successfully completing this course, participants will have a solid understanding of how RapidMiner Studio supports web mining. Participants will be able to identify techniques for retrieving different types of data from the web, processing unstructured web text data, retrieving more structured web data from tables, via APIs, or from sources such as Twitter. Practical exercises during the course prepare students to take the knowledge gained and apply it to their own web mining challenges. Examples include: web scraping , sentiment analysis of text documents from web reviews or blogs, and use of third-party APIs and web services to enrich data science projects. 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). **It also assumes prior knowledge of text mining techniques.**

Course Outline

- Crawling the Web
- Extracting Information from Web Sites
- Transforming Web Data to Documents
    - HTML processing
    - Content extraction using string matching and Regular Expressions (regex)
- Retrieving Structured Web Data
    - Tables
    - Files & PDF documents
- Data ETL and Preprocessing for Web Sourced Data
- Enriching Data via Web Services
- Machine Learning Techniques for Web Mined Data
    - SVM
    - k-NN
    - Naive Bayes
- Twitter Operators
- Using Third Party Web Mining Extensions