Data Quality and Data Wrangling

Course Description

The goal of data science can be summarized as the extraction of insights (hence, value) from data. It is self-evident that this objective cannot be successfully achieved based on unreliable and untrustworthy data. This course aims at establishing the notion of data quality and the pertinent methods for data quality management. Furthermore, techniques for acquiring data as well as formatting and tidying data in order to make it suitable for subsequent analytical treatment are covered.

Contents

1. Data Quality
   1. Introduction to data quality
   2. Data quality dimensions and issue types
2. Data Quality Management
   1. Data governance and stewardship
   2. Activities and processes
3. Data Acquisition
   1. Web scraping
   2. Data APIs
4. Working with Common Data Formats
   1. Text-based formats (CSV, XML, JSON)
   2. Binary formats (HDF 5, Parquet, Arrow)
5. Tidy Data
   1. Structuring
   2. Cleansing
   3. Enrichment