Course:

Big Data Processing and Tools

Part I. Data Science Analytics Foundation

Syllabus and Procedures

Description:

Big Data Processing is using both advanced data analytics applications and software and requires high performance Big Data platforms and tools capable of processing big amount of data to extract values from data to achieve specific organisational goals. Developing new data analysis applications and effectively using existing applications for processing organisational data requires knowledge of the core data analytics techniques and experience with the modern data analytics tools.

This course provides introduction into Data Analytics basics, platforms and tools that should provide a strong foundation for understand the whole spectrum of the related technologies such as Machine Learning, Data Mining, Exploratory, Predictive and Prescriptive Analytics. The course will include introduction into the most widely used technologies such as Classification and Cluster analysis. This will be supported by detailed discussion of the data preparation.

Students will apply the taught concepts to a set of hands on labs and projects that will be both individual and performed by small groups to facilitate learning.

Aims:

The aim of this course is to create a consistent understanding of the whole Big Data and Data Analytics domains that should rely on the strong foundation of the general statistical data analysis and important data analytics and machine learning algorithms and techniques. This should provide sufficient initial knowledge for independent work and further (self-)study.

Learning outcomes

By the end of this course, students will be able to do following:

- Understand the concept of Big Data Processing by using main Data Analytics techniques, algorithms, and tools.
- Understand basic statistical concepts used in data analysis and imbedded in Data Mining and Machine Learning algorithms
- Understand importance of data preparation and the whole data lifecycle
- Use such data analytics tools as RapidMiner, R-Studio and Python Jupyter Notebook and develop simple applications
- Have hands on experience with one of such tools as a results of a project completion

Text Book:

Selected books and papers are provided as supplemental learning materials in the shared course folder.

Other References:

To be provided during the course

Recommended Prior Knowledge:

Familiarity with at least one of programming languages.

Course Structure and works:

Lectures

Lab, practical workshops

Projects: individual and by student groups

Syllabus:

[by lecture days and topics]

Day 1 – 17 May, 2017

Lecture 1: Introduction and course overview: Data Science and Big Data technologies, Data Science competence and skills, Research Methods in Data Science, Machine Learning overview
Lecture 2: Data description and Statistical Data Analysis
Lecture 3: Data preparation: data entry, data cleaning, data pre-processing

Lab/Workshop: Hands on labs with RapidMiner (Lab 1 Regression analysis and data explorations)

Day 2 – 18 May, 2017

Lecture 4.1: Classification: methods and algorithms Part 1
Lecture 4.2: Classification: methods and algorithms Part 2

Lab/Workshop: Hands on labs with RapidMiner (Lab 2 and Lab 3)

Day 3 – 19 May, 2017
Lecture 5: Cluster analysis basics
Lecture 6: Performance of data analytics algorithms and tools

Lab/Workshop: Hands on labs with RapidMiner (Lab 4 with individual and group tasks)