Study programme(s): Applied Mathematics – Data Science

Level: Master studies

Course title: Large scale data mining

Lecturer: Dušan Jakovetić, Miloš Radovanović, Vladimir Kurbalija

Status: obligatory

ECTS: 5

Requirements: Pattern recognition and machine learning, Graph theory

Learning objectives

- Introducing the methods for large-scale computational data analysis
- Learning programming skills and tools for storing, querying, and analysing large-scale data
- Ability to combine skills from areas such as data storage, distributed systems design, signal processing, statistical data analysis, machine learning, graph theory, etc. in order to create value from Big Data

Learning outcomes

- Experience in analysis and processing of massive data sets
- Ability to design and implement an analytical solution: choose appropriate storage, algorithms, provide result interpretation and visualisation
- Ability to work and solve problems in a variety of data intensive areas

Syllabus

- Data storage (Files, SQL, noSQL, Map-Reduce) and data pre-processing; Query processing; Finding similar items; Graph analysis; Frequent itemset mining; Features engineering and selection; Integration of data / knowledge / methods (ensemble techniques in unsupervised, supervised and semi-supervised learning framework); Data visualization;
- Case studies and applications on heterogeneous data (logs, text, spatio-temporal data, social graphs, etc.) from real-world sources (smart phones, telecom operators, social media, satellite imagery, sensors, genomics)
- Implementing solutions in Python with additional packages: Numpy, SciPy, Networkx, Matplotlib, Orange, Scikit-learn, Pandas, PyMongo, Pydoop

Literature

1. Jure Leskovac, Anand Rajaraman, Jeffrey D. Ullman, "Mining of Massive Datasets", Cambridge University Press, 2010.
2. Pang-Ning Tan, Michael Steinbach, Vipin Kumar, "Introduction to data mining", Pearson Addison Wesley, 2006.
3. Jeffrey Dean, and Ghemawat Sanjay, "MapReduce: simplified data processing on large clusters", Communications of the ACM, 2008.
4. Santo Fortunato, "Community detection in graphs", Physics Reports, 2010.
5. Giovanni Seni, and John F. Elder, "Ensemble methods in data mining: improving accuracy through combining predictions", Synthesis Lectures on Data Mining and Knowledge Discovery, 2010.
6. Wes McKinney, Python for Data Analysis, O'Reilly Media, 2012.

Weekly teaching load				Other: 0
Lectures: 2	Exercises: 2	Other forms of teaching: 0	Student research: 0	

Teaching methodology

Lectures; revisions of the material; active students' participation in problem solving; homework assignments; application of the taught material on real-world examples.

Grading (maximum number of points 100)

Pre-exam obligations		**Points**	**Final exam**	**points**
Solved homework assignments	Course project	60 = 30 (Homework assignments) + 30 (Course project)	written exam	40