Course Unit Descriptor

Study Programme: Applied Mathematics – Data Science

Course Unit Title: Large scale data mining

Course Unit Code: MDS09

Name of Lecturer(s): Dušan Jakovetić, Miloš Radovanović, Vladimir Kurbalija

Type and Level of Studies: Master studies

Course Status (compulsory/elective): Compulsory

Semester (winter/summer): Winter

Language of instruction: English

Mode of course unit delivery (face-to-face/distance learning): Face-to-face

Number of ECTS Allocated: 5

Prerequisites: Pattern recognition and machine learning, Graph theory

Course Aims:

- Introducing the methods for large-scale computational data analysis
- Learning programming skills and tools for storing, querying, and analyzing 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 visualization
- Ability to work and solve problems in a variety of data intensive areas

Syllabus:

Theory

- Data storage (Files, SQL, noSQL, Map-Reduce) and data pre-processing; Query processing; Finding similar items; Graph analysis; Frequent item set 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)

Practice

- Implementing solutions in Python with additional packages: Numpy, SciPy, Networkx, Matplotlib, Orange, Scikit-learn, Pandas, PyMongo, Pydoop

Required Reading:

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 Contact Hours:	**Lectures:** 2	**Practical work:** 2

Teaching Methods:

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

Knowledge Assessment (maximum of 100 points):

Pre-exam obligations	points	**Final exam**	points
Active class participation	30	written exam	40
Practical work	30	oral exam	
Preliminary exam(s)		…….	
Seminar(s)			

The methods of knowledge assessment may differ; the table presents only some of the options: written exam, oral exam, project presentation, seminars, etc.