SYLLABUS

Name: ***Large-scale machine learning (1000-319bBML)***
Name in Polish: **Uczenie maszynowe w dużej skali**
Name in English: ***Large-scale machine learning***

Information on course:

Course offered by department: Faculty of Mathematics, Informatics, and Mechanics
Course for department: Faculty of Mathematics, Informatics, and Mechanics

Default type of course examination report:

Examination

Language:

English

Short description:

During this class we will present techniques and tools for processing Big data. We will focus on the ones useful for machine learning practitioners. We will show the most important models and basic algorithmic techniques. We will cover how to analyze algorithms that process large data on clusters. Finally, we will introduce typical optimizations that can be useful in machine learning applications like linear regression, clustering, decision trees or neural networks.

Description:

-Distributing computation to clusters of commodity machines and distributed file system.
-MapReduce model and basic algorithmic techniques for this model. Comparing of MapReduce algorithms and typical algorithms for typical problems (matrix multiplication, multi-way join, counting triangles in large graphs).
-Total vs elapsed communication cost. Skew and methods to deal with it.
-Spark and Resilient Distributed Dataset model.
-Spark SQL and its optimizations.
-Serialization of Big data and columnar formats.
-Managed cloud data warehouse.
-Algorithms for stream pressing.
-Distributing typical machine learning algorithms, e.g., linear regression, clustering, decision trees or neural networks.
-Neural networks in large scale (data parallelism, model paralelizm).
-Learned index structores.

Bibliography:

-Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. Mining of Massive Datasets. Cambridge University Press
-Guglielmo Iozzia, Hands-On Deep Learning with Apache Spark, Packt Publishing
-Butch Quinto, Next-Generation Machine Learning with Spark: Covers XGBoost, -LightGBM, Spark NLP, Distributed Deep Learning with Keras, and More, Apress

Assessment methods and assessment criteria:

Final mark based big programming assignments, points for participation in laboratories and written exam.

Type of course

elective monographs

Prerequisites (description)

object oriented programming, computer networks, algorithms and data structures

Element of course groups in various terms:

Course group description	First term	Last term
Obligatory courses for 2nd year Machine Learning (1000-ML2-OBW)	2022	
Subjects used to compute average of grades for ERASMUS (1000-ERASMUS)	2022	

Course credits in various terms:

<without a specific program>

Type of credits	Number	First term	Last term
European Credit Transfer System (ECTS)	6	2022	