Opis **zajęć (sylabus)**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Title: | | Processing massive datasets | | | | | | | | **ECTS** | **4** |
| Translation if applicable: | | Processing massive datasets | | | | | | | | | |
| Field of study: | | **Informatics and econometrics** | | | | | | | | | |
|  | |  | | | | | | | | | |
| Language: | | English | | | | Level: | | II (master) | | | |
| Form: | 🗷 regular  🞎 weekend | Status: | 🞎 primary  🗷 specialisation | 🞎 mandatory  🗷 elective | | Semester: ……2….. | | 🞎 winter semester 🗷 summer semester | | | |
|  |  | Didactic year: | | | | 2019/2020 | Numer katalogowy: | **ZIM-IE-BDA-2S-02L-16\_2** | | | |
|  | | | | | | | | | | | |
| Coordinator: | |  | | | | | | | | | |
| Lecturer: | |  | | | | | | | | | |
| Department: | |  | | | | | | | | | |
| Contracting department: | |  | | | | | | | | | |
| Assumptions, the goal and the description | | The course is devoted to the programming in the R language.A special focus will be on the discussion and exercise of programming techniques and the use of tools useful in mathematics-practice (including data analyst, statistical programmer, junior data scientist) and in scientific and research applications.Subjects of lectures and laboratory classes:Part one:  * Introduction * R Basics * Code Editors for R * Finding Help * Control Structures * Conditional Executions * Comparison Operators * Logical Operators * If Statements * Ifelse Statements * Loops * For Loop * While Loop * Apply Loop Family * For Two-Dimensional Data Sets: apply * For Ragged Arrays: tapply * For Vectors and Lists: lapply and sapply * Other Loops * Improving Speed Performance of Loops * 7 Functions  Part two:  * Parallel and distributed processing * Introduction to Map-Reduce processing * Apache Hadoop * HDFS - Distributed file systems * Apache Hadoop installation on a virtual machine (VirtualBox) * Writing Map-Reduce programs in Apache Hadoop * Parallel processing in R * Integration of R and Hadoop | | | | | | | | | |
|  | | 1. Lectures: 15h; 2. Computer labs: 30h; | | | | | | | | | |
| Didactic methods: | | Presentations, discussion, practical examples, numerical experiments, case study | | | | | | | | | |
| Assumptions: | | Basic skills in data processing, information systems, databases, statistics and data analysis. | | | | | | | | | |
| Effects: | | Knowledge:  2 – Has knowledge about the process of collecting, processing, extracting data from database systems and data warehouses.  2 – Has knowledge on mass data specifics and extracting it from source systems using IT technology and efficiently prepare it depending on the problem and the requirements. | | | Skills:  3 – Can use literature to gain required skills on analytical methods (data mining techniques)and to formulate and solve tasks at the interface between IT and business applications (in IT, industry, medicine, marketing).  2 – Has substantial skills in analysis of market phenomena and business processes, including the specificity of collecting, processing, extracting data from database systems and data warehouses. | | | | Competence: | | |
| The way of verifying the effects of education: | | Written-programming test. | | | | | | | | | |
| Form of documentation of the learning outcomes achieved: | | Test to verify ability to process data. | | | | | | | | | |
| Elements and weights of the final grade: | | **Colloquium x2 – 100% (2x50%)** | | | | | | | | | |
| Type of the classes | | Lectures + computer labs | | | | | | | | | |
| Literature:   1. Rajaraman, A., & Ullman, J. D. (2011). Mining of massive datasets. Cambridge University Press. | | | | | | | | | | | |
| Comments  The condition of passing the subject means obtaining at least 50% of points from each colloquium. | | | | | | | | | | | |

Quantitative indicators characterizing the module:

|  |  |
| --- | --- |
| Estimated total number of student work hours (contact and own work) necessary to achieve the assumed learning outcomes | **100 h** |
| The total number of ECTS points that a student receives in classes requiring direct participation of academic teachers: | **2 ECTS** |

Effects to program mapping table:

|  |  |  |  |
| --- | --- | --- | --- |
| Effect category | Effects: | Effects to program mapping | Impact\*) |
| Knowledge 1 | Has knowledge about the process of collecting, processing, extracting data from database systems and data warehouses. | K\_W06 / P7S\_WG | 2 |
| Knowledge 2 | Has knowledge on mass data specifics and extracting it from source systems using IT technology and efficiently prepare it depending on the problem and the requirements. | K\_W10 / P7S\_WG | 2 |
| Skill 1 | Can use literature to gain required skills on analytical methods (data mining techniques)and to formulate and solve tasks at the interface between IT and business applications (in IT, industry, medicine, marketing). | K\_U01 / P7S\_UW | 3 |
| Skill 2 | Has substantial skills in analysis of market phenomena and business processes, including the specificity of collecting, processing, extracting data from database systems and data warehouses. | K\_U15/ P7S\_UW | 2 |

\*)

3 – advanced,

2 – substantial,

1 – basic,