Tárgy neve: Data Mining

Tárgyfelelős neve: Buza Krisztián Antal

Tárgyfelelős tudományos fokozata: PhD, egyetemi docens

Tárgyfelelős MAB szerinti akkreditációs státusza: AT

Az oktatás célja angolul:

a) knowledge

- In order to be able to perform their work in an innovative way and do research (when necessary) in their own IT specialization, they have comprehensive and up-to-date knowledge of general mathematical and computing principles, rules and relationships, particularly – depending on their chosen specialization – in the following areas: algebraic, linear algebraic and number theory methods and applications, special fields of mathematical analysis, numerical methods and their applications; discrete mathematics, graph theory, logic and their applications; theoretical basics and applications of stochastic modelling and statistics; first-order and second-order statistical analysis, operation research; algorithmic methods in mathematics, formal models and tools in computing science, complexity and efficiency theory of algorithms, and special algorithms of application fields.
- They have comprehensive and up-to-date knowledge of the principles, methods, and procedures for designing, developing, operating, and controlling IT processes, particularly – depending on their chosen specialization – in the areas of program design methods; design, construction and management of complex software systems and databases in modern database management systems; service-oriented program design; the design, construction and management of information systems; the design and development of tools and services for the internet; the design, development and management of database systems; the design, construction and management of distributed systems, cryptography, data security and data protection.
- They have comprehensive and up-to-date knowledge of specific IT tools, particularly – depending on their chosen specialization – in the areas of numerical computing systems, model analysis, scientific computing methods, digital signal and image processing, artificial intelligence methods, software methods of operation research and optimization, modern programming languages and paradigms, the usage of modern programming languages; theoretical foundations and applications of information systems; distributed and parallel systems, expert systems; information technology and application security, geoinformatics; the construction and organization of health information systems; new methods of information management and organization, corporate (enterprise & business) information systems, services of information systems implementing corporate (enterprise & business) processes; digital signal and image processing, computer graphics; web and multimedia applications, and media informatics.
- They are familiar with the principles of business, organizational and corporate procedure, information, data, software and technical-technological architectures as well as with the methods of describing and designing these architectures.
- They have extensive knowledge on business, enabling them to perform business analysis, and to establish and run an IT enterprise.

b) skills and abilities

- They are able to formalize complex IT tasks, to identify and study their theoretical and practical background and then to solve them.
- They are able to initiate collaboration and work in a team as well as on projects with IT or other professionals.
- They are able to analyze and apply new problem-solving methods and procedures related to their IT specialization.
- They are able to apply their IT skills in a diverse, multidisciplinary professional environment.

- They are familiar with IT professional vocabulary, which enables them to express themselves at a high level, both orally and in writing, in their mother tongue and (at least) in English; i.e. they are able to participate in discussions and debates, to write reports, to work with, understand and utilize scientific and technical literature (e.g. professional books, chapters, articles etc.).
- They are able to professionally use scientific and technical information sources to obtain knowledge necessary for solving a problem, and to critically interpret and evaluate it.
- Under professional guidance, they are able to carry out scientific research on their own, and to prepare for further studies at postgraduate level.

c) attitude

- They follow professional and technological developments in their IT field.
- They are committed to critical feedback and evaluation based on self-examination.
- They are committed to lifelong learning, and they they they are open to acquiring new IT competencies.
- They accept and make their co-workers apply the ethical principles of work and organizational culture as well as those of IT scientific research.
- They share their knowledge and consider it important to disseminate professional IT results.
- They are open to proactive collaboration with IT and other professionals.

d) autonomy and responsibility

- They take responsibility for their professional decisions made in their IT-related activities.
- They undertake to meet deadlines and to have deadlines met.
- They bear responsibility for their own work as well as for the work of their colleagues they work together with in a project.

Az oktatás tartalma angolul:

This course provides an overview of the important data mining algorithms and their application possibilities and it covers the following topics:
Data types and properties, data preprocessing and cleaning techniques (handling missing and incorrect values, searching for duplicates).
Distance and similarity measures for vectors and time series (DTW: dynamic time warping).
Data reduction (dimensionality and size reduction), the curse of dimensionality (PCA, ICA, Isomap size reduction).
Classification as supervised machine learning. The related concepts and techniques: overlearning, evaluating of different classification methods, classification with unbalanced distributed data.
Classification methodologies: decision trees, rule based methods, naïve bayes approach, k-nearest neighbors method, artificial neural networks, support vector machine, Ensemble methods.
Frequent Itemset Mining problems, market basket analysis: frequent itemsets and association rules.
Clustering methods (unsupervised learnings), hierarchical and non-hierarchical methods, density-based methods.
Anomaly and outlier detection and handling.
Advanced data mining techniques, basics of semi-supervised classification, active learning, multitask classification.

During the practice the students solve different data mining exercises connected to the lectures in Python language.

A számonkérés és értékelés rendszere angolul:

continuous assessment, examination
continuous assessment during the semester (small and quick tests, home works, presentations based on literature processing, on occasion in teamwork.

Idegen nyelven történő indítás esetén az adott idegen nyelvű irodalom:

Text book, compulsory:

- Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar (2018): Introduction to Data Mining (Second Edition), ld. https://www-users.cs.umn.edu/~kumar001/dmbook/index.php

Proposed further reading:

- N. Tomasev, K. Buza, K. Marussy, P.B. Kis (2015): Hubness-aware Classification, Instance Selection and Feature Construction: Survey and Extensions to Time-Series In: U. Stanczyk, L. Jain (eds.), Feature selection for data and pattern recognition, Springer-Verlag. ld. http://www.cs.bme.hu/~buza/pdfs/marussyHubness.pdf
- Jiawei Han, Micheline Kamber, Jian Pei (2016): Data Mining, Concepts and tTechniques, third edition, http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf