DATA ANALYSIS I - TASKS TO OBTAIN GRADED CREDIT

Select a data collection (either network data or vector data). The data collection must be real, either referential or custom, and must contain at least a thousand instances or thousand network vertices.

1. DATA ANALYSIS (10-20 points) (expected time required to carry out the task 4-6 hours)

The aim is to process (analyse) your selected data collection in some tool (Weka, R, NodeXL, Pajek, Gephi). Processing is understood as a detailed analysis of the data collection and the presentation of the results of this analysis, including visualization of outputs (statistical graphs, networks). The output will be in the form of a text document (PDF only). It will contain a description of the data set (where and how it was obtained, what it contains, etc.), the analysis results and the interpretation of the results (what they mean).

2. IMPLEMENTATION TASK (12-24, points) (expected time required to carry out the task - 4 hours)
The goal of the implementation (in C #, C ++, C, Java, or Python) is to bring the tasks from the seminars to a more complex code with output to the report as a text file.

a) IF YOU SELECTED VECTOR DATA (FOR DATA MINING)

1. Implementation of one of the algorithms discussed in lectures or other algorithms from the discussed areas (mainly clustering and classification).
2. The algorithm must be able to process data with at least a thousand instances and work with both categorical and numeric attributes, including missing values (such as during pre-processing of the dataset).
3. The result of the algorithm application will be a text file, which will summarize information about the inputs and outputs of the algorithm. This will include information on the range of data collection, properties of attributes, results, and measured properties (what we measured during seminars, e.g. total variance, standard deviation, etc.) and the results of the implemented algorithm.
4. The algorithm outputs will be validated using one of the tools (e.g. Weka, R) whose outputs should match the outputs of the implemented algorithm.

b) IF YOU SELECTED NETWORK DATA (FOR NETWORK ANALYSIS DATA)

1. Implementation of one of the network generation algorithms based on models discussed in lectures or other from this area (except Erdös-Renyi). Output from the algorithm application will be a text file with a list of edges (vertex pairs). Depending on the chosen task, it can be both the directed or undirected network (graph), and, both the unweighted and the weighted graph.
2. Implementation of one of the algorithms analysing graph properties (degree distribution and mean degree, clustering coefficient and average aggregation coefficient of the graph, average graph, etc.). The algorithm must be able to process a graph with at least thousands of vertices and thousands of edges.
3. The result of the algorithm application will be a text file, which will summarize information about the inputs and outputs of the algorithm. This will include information on the range of data collection, measured properties (what we measured during seminars,
4. The algorithm outputs will be validated using one of the tools (e.g. R, Pajek) whose outputs should match the outputs of the algorithm.