METODY ANALÝZY DAT I: ZADÁNÍ ÚKOLŮ K ZÁPOČTU

1. IMPLEMENTAČNÍ ÚKOL - Cílem implementace je dotáhnout úkoly ze cvičení do komplexnějšího kódu s výstupem do reportu v podobě textového souboru. Aplikace bude mít jednoduché GUI pro načtení datové sady a nastavení algoritmu / algoritmů.
    a) Buď DATA MINING (PŘEDPOKLÁDANÝ ČAS 6 HODIN)
        i. Implementace jednoho z algoritmů probíraných na přednáškách nebo jiného algoritmu z probíraných oblastí (především shlukování a klasifikace).
        ii. Algoritmus musí umět zpracovat data nejméně s tisícem instancí a musí pracovat jak s kategoriálními, tak numerickými atributy, a to včetně chybějících hodnot (např. v rámci předzpracování datové sady).
        iii. Výsledkem aplikace algoritmu bude textový soubor, ve kterém budou shrnuty informace o vstupech a výstupech algoritmu. A to včetně informací o rozsahu datové sady, vlastnostech atributů, výsledcích, naměřených přesnostech – rozptyl, směrodatná odchylka, kvalita shlukování apod.
        iv. Sada pro testování algoritmu může být umělá (generovaná), referenční (z webových zdrojů nebo z používaných nástrojů), nebo vlastní získaná z jiného zdroje.
        v. Výstupy algoritmu budou validovány použitím některého z nástrojů (např. Weka), jehož výstupy by měly odpovídat výstupům implementovaného algoritmu.
    b) Nebo ANALÝZA SÍTÍ (PŘEDPOKLÁDANÝ ČAS 6 HODIN)
        i. Implementace jednoho z algoritmů generování grafu na základě modelů probíraných na přednáškách nebo jiného z této oblasti (kromě Erdös-Renyi).
        ii. Výsledkem aplikace algoritmu bude textový soubor osahující seznam hran (dvojic vrcholů).
        iii. Implementace některého z algoritmů analyzujícího vlastnosti grafu (distribuce stupňů vrcholů grafu a průměrný stupeň, shlukovací koeficient vrcholu a průměrný shlukovací koeficient grafu, průměr grafu apod.).
        iv. Algoritmus musí umět zpracovat graf nejméně s tisíci vrcholy a řádově s tisíci hranami.
        v. Výsledkem aplikace algoritmu bude textový soubor obsahující shrnutí získaných informací.
        vi. Graf pro testování algoritmu může být umělý (např. generovaný vlastním algoritmem z bodu 1), referenční (z webových zdrojů nebo z používaných nástrojů), nebo vlastní získaný z jiného zdroje. Podle zvolené úlohy se může jednat jak o neorientovaný, tak o orientovaný graf, jak o nevážený (neohodnocený), tak o vážený (ohodnocený) graf.
        vii. Výstupy algoritmu budou validovány použitím některého z nástrojů (např. R), jehož výstupy by měly odpovídat výstupům algoritmu.
2. ANALYTICKÝ ÚKOL (PŘEDPOKLÁDÁNÝ ČAS 6 HODIN) - Cílem je zpracovat reálná data v některém k tomu určenému nástroji (Weka, R, NodeXL, Pajek, Gephi). Zpracováním se rozumí podrobná analýza datové sady resp. sítě a prezentace výsledků této analýzy včetně vizualizace výstupů (statistické grafy, sítě). Datová sada musí být reálná, a to buď referenční, nebo vlastní a musí obsahovat nejméně tisíc instancí resp. vrcholů sítě. Lze použít i reálnou nebo referenční sadu použitou pro testování úkolu z implementace. Výstup bude ve formě textového dokumentu (PDF), bude obsahovat popis datové sady (kde a jak byla získána, co obsahuje apod.), výsledky analýzy a interpretaci výsledků (co znamenají).