Kod szkolenia: **WEKA**

Tytuł szkolenia: **Techniczne aspekty eksploracji danych zgromadzonych w hurtowni danych z wykorzystaniem Pentaho Data Mining (WEKA)**

Dni: 4

Opis:

Adresaci szkolenia:

Szkolenie jest adresowane do programistów, architektów oraz administratorów aplikacji, którzy chcą tworzyć lub utrzymywać procesy eksploracji danych z wykorzystaniem Pentaho Data Mining (WEKA). Szkolenie jest także kierowane do osób, które chcą uzupełnić swoją wiedzę o pojęcia związane z hurtowaniami danych (DWH) oraz ich realizacją z wykorzystaniem oprogramowania Pentaho Business Intelligence Suite.

Cel szkolenia:

Uczestnicy szkolenia zdobędą przekrojową wiedzę dotyczącą projektowania, implementowania, monitorowania, uruchamiania, strojenia procesów DM, odświeżą wiedzę na temat podstawowych pojęć statystycznych, poznają najpopularniejsze algorytmy DM w szczegółach, poznają założenia hurtowni danych. Dzięki temu będą mogli wybrać właściwy zestaw narzędzi i technik dla swoich projektów. Szkolenie, poza ogólnym wprowadzeniem do pojęć teoretycznych, skupia się na stosie produktowym wybudowanym wokół Pentaho Business Intelligence a w szczególności na Pentaho Data Mining (WEKA).

Mocne strony szkolenia:

Program obejmuje zarówno ogólne wprowadzenie w tematykę DM i DWH, jak i całościowe przedstawienie stosu produktowego Pentaho Data Mining. Szkolenie jest unikalne, gdyż tematyka poruszana w jego trakcie nie jest wyczerpująco ujęta w dostępnej literaturze, a wiedza na ten temat jest mocno rozproszona. Program jest ciągle uaktualniany ze względu na szybki rozwój rozwiązań, których dotyczy szkolenie.

Wymagania:

Od uczestników wymagana jest podstawowa znajomość baz danych, podstawowa umiejętność programowania w języku Java.

Parametry szkolenia:

4*8 godzin (4*7 netto) wykładów i warsztatów, z wyraźną przewagą warsztatów. W trakcie warsztatów, oprócz prostych ćwiczeń, uczestnicy rozwiązują problemy eksploracji danych wykorzystując i strojąc algorytmy DM. Wielkość grupy: maks. 8-10 osób

Program szkolenia:

1. Wstęp
    a. Wprowadzenie do hurtowni danych:
        i. OLTP, OLAP, bazy danych, hurtownie danych, data marty
        ii. ROLAP, MOLAP, HOLAP
        iii. Normalizacja, agregacja, fakty, wymiary
        iv. SQL, MDX, XML/A
        v. ETL
        vi. BigData, BigTable, NoSQL, nierelacyjne hurtownie danych
        vii. Pozostałe
    b. Platforma Pentaho BI Suite
2. Eksploracja danych
    a. Sztuczna inteligencja, uczenie maszynowe, eksploracja danych etc.
    b. Podstawy algorytmów eksploracji danych
        i. Algorytmy
            klasyfikacja
            grupowanie
            odkrywanie wzorców i reguł asocjacji
            ograniczanie i transformacja przestrzeni atrybutów
        ii. Techniki:
            drzewa i tabele decyzyjne
            regresja liniowa
            sieci bayesa
            sieci neuronowe
            algorytmy genetyczne i ewolucyjne
        iii. Podstawowe pojęcia statystyczne
            Minimum, Maximum
            Średnia, Mediana
            Odchylenie standardowe, Wariancja
            Prawdopodobieństwo
            Korelacja
            Metryka odległości danych
            Statystyczna istotność
        iv. pozostałe
    c. Przegląd narzędzi eksploracji danych dostępnych na rynku
3. Pentaho Data Mining (WEKA)
    a. Architektura
    b. Weka Gui Chooser
        `. Explorer
        `. Experimenter
        `. Knowledge Flow

        - `. Simple CLI
        - `. Tools: ARFF Viewer, SQL Viewer etc.
        - `. Weka Light, Weka Server
    - c. Praca z Explorer'em
4. Preprocessing i praca z danymi
    - a. Format danych ARFF
    - b. Przygotowanie danych do analizy
    - c. Odpowiedni dobór atrybutów np.: korelacja atrybutów a wyniki eksploracji danych etc.
    - d. Filtrowanie i rodzaje filtrów w WEKA np.: filtrowanie, dyskretyzacja, normalizacja etc.
    - e. Wizualizacja
    - f. Przetwarzanie dużych zbiorów danych, ograniczenia JVM 32bit
    - g. Przetwarzanie strumieni oraz uczenie przyrostowe
5. Klasyfikacja
    - a. Definicja problemu klasyfikacji
    - b. Odpowiedni zbiór danych uczących i testujących a wyniki klasyfikacji
    - c. Rodzaje algorytmów klasyfikacji dostępnych w WEKA
    - d. Najpopularniejsze algorytmy klasyfikacji w szczegółach
        - i. Sieci Bayesa np.: naiwny klasyfikator bayesowski
        - ii. Regresja np.: regresja liniowa
        - iii. Drzewa i tablice decyzyjne
    - e. Walidacja krzyżowa, nadmierne dopasowanie
    - f. Interpretacja wyników klasyfikacji
6. Grupowanie
    - a. Definicja problemu grupowania
    - b. Odpowiedni zbiór danych uczących i testujących a wyniki grupowania
    - c. Rodzaje algorytmów grupowania dostępnych w WEKA
    - d. Najpopularniejsze algorytmy grupowania w szczegółach
        - i. Centroidy np.: k-średnich
        - ii. Gęstościowe np.: DBSCAN
    - e. Interpretacja wyników grupowania
7. Odkrywanie reguł asocjacyjnych
    - a. Definicja problemu odkrywania wzorców i reguł asocjacyjnych
    - b. Odpowiedni zbiór danych uczących i testujących a odkryte reguły
    - c. Rodzaje algorytmów odkrywania reguł asocjacyjnych dostępnych w WEKA
    - d. Najpopularniejsze algorytmy odkrywania reguł asocjacyjnych w szczegółach
        - i. Apriori
        - ii. Frequent Pattern Growth
    - e. Interpretacja odkrytych reguł
8. Ograniczanie i transformacja przestrzeni atrybutów
    - a. Definicja problemu selekcji, ograniczenia, transformacji atrybutów
    - b. Odpowiedni zbiór danych uczących i testujących a wybrane atrybuty
    - c. Rodzaje algorytmów ograniczania i transformacji przestrzeni atrybutów w WEKA
    - d. Najpopularniejsze algorytmy ograniczania i transformacji przestrzeni atrybutów

w szczegółach
    i. Przeszukiwania np.: BestFirst, ExhaustiveSearch, GeneticSearch
    ii. Analizy głównych składowych np.: PCA/PrincipalComponents
    iii. Maszyna wektorów nośnych np.: SVM/SVMAttributeEval
e. Interpretacja wyników

9. Pozostałe algorytmy i techniki eksploracji danych dostępne w WEKA
10. Rozbudowa możliwości WEKA
    a. Pentaho Data Mining Plug-Ins
    b. Własne algorytmy DM w WEKA
11. Wykorzystanie możliwości w połączeniu z innymi produktami Pentaho
    a. Knowledge Flow Plugin oraz Pentaho Data Integration

Sages sp. z o.o., ul. Nowogrodzka 62c, 02-002 Warszawa
tel. +48 22 203 56 00 fax: +48 22 203 56 01
e-mail: biuro@sages.com.pl web: www.sages.com.pl