Informatyka, studia STACJONARNE, I st. semestr VI

Komputerowe systemy rozpoznawania **2020/2021**

Prowadzący: prof. dr hab. inż. Adam Niewiadomski poniedziałek, 12:00

Data oddania: _____ Ocena: _____

Julia Szymańska 224441

Przemysław Zdrzalik 224466

Projekt 2. Podsumowania lingwistyczne relacyjnych baz danych

1. Cel

Celem projektu jest stworzenie aplikacji lingwistycznie agregującej zawartości zbioru danych. Jako wynik działania aplikacji zostanie wygenerowany opis zawartości danych liczbowych w zbiorze w języku quasi-naturalnym.

2. Charakterystyka podsumowywanej bazy danych

W programie został użyty zbiór danych znajdujący się w pliku CSV, który został przekształcony w bazę danych.

Zbiór danych zawiera informacje o ponad 3 milionach wypadków samochodowych w 49 stanach Zjednoczonych Stanów Ameryki, mających miejsce od lutego 2016 do grudnia 2020. Spośród 47 kolumn znajdujących się w zbiorze danych, wybraliśmy następujących 11 kolumn:

— Dotkliwość - Severity - wpływ wypadku na ruch na drodze, przyjmuje wartości całkowite od 1 do 4 włącznie, gdzie 1 oznacza najmniejszy wpływ na ruch drogowy, natomaist 4 oznacza największy wpływ.

— Czas rozpoczęcia - Start_Time - czas rozpoczęcia się wypadku w lokalnej strefie czasowej, przyjmuje wartości od 8 lutego 2016, do 31 grudnai 2020. Wartość kolumny zostanie zamieniona na wartość całkowitą oznaczającą liczbę sekund od początku 1970 roku.

— Czas zakończenia - End_Time - czas zakończenia się wypadku w lokalnej strefie czasowej, przyjmuje wartości od 8 lutego 2016, do 1 stycznia 2021.

Wartość kolumny zostanie zamieniona na wartość całkowitą oznaczającą liczbę sekund od początku 1970 roku.

— Odległość - Distance - długość odcinka ulicy, na którego miał wpływ wypadek wyrażony w milach. Przyjmuje wartości zmiennoprzecinkowe od 0 do 334, gdzie zdecydowana większość danych mieści się w przedziale od 0 do 6.67.
— Temperatura - Temperature - temperatura powietrza w momencie, gdy zdarzył się wypadek wyrażona w Fahrenheit'ach. Przyjmuje wartości zmiennoprzecinkowe od -16 do 104.
— Temperatura odczuwalna - Wind_Chill - temperatura odczuwalna w momencie, gdy zdarzył się wypadek wyrażona w Fahrenheit'ach. Przyjmuje wartości zmiennoprzecinkowe od -16 do 101.
— Wilgotność - Humidity - wilgotność powietrza w momencie, gdy zdarzył się wypadek wyrażona w procentach. Przyjmuje wartości zmiennoprzecinkowe od 4 do 100.
— Ciśnienie - Pressure - ciśnienie powietrza w momencie, gdy zdarzył się wypadek wyrażona w inches. Przyjmuje wartości zmiennoprzecinkowe od 27 do 32.
— Widoczność - Visibiity - widoczność w momencie, gdy zdarzył się wypadek wyrażona w milach. Przyjmuje wartości zmiennoprzecinkowe od 0 do 12.
— Prędkość wiatru - Wind_Speed - prędkość wiatru w momencie, gdy zdarzył się wypadek wyrażona w milach na godzinę. Przyjmuje wartości zmiennoprzecinkowe od 0 do 40.
— Ilość opadów - Principation - ilość opadów w momencie, gdy zdarzył się wypadek wyrażona w inches. Jeśli wyapdy nie występowały to kolumna przyjmuje wartość nan. Przyjmuje wartości zmiennoprzecinkowe od 0 do 0.5.

Krótki opis bazy danych wybranej do podsumowywania, źródło, opis treści, użyteczność/zastosowania. Liczba rekordów (min. 10 000 i koniecznie wszystkie tego samego typu), liczba atrybutów możliwych do rozmycia (min. 10), czyli o stosunkowo dużej liczbie możliwych wartości. Zwyczajowe wartości lingwistyczne nadawane wybranym atrybutom oraz dlaczego istnieje zwyczaj, zapotrzebowanie/inne powody „przekałdania" tych danych na język naturalny (a nie formalny) [1, 3].
Realizacja bazy w wybranym DBMS. Rysunek lub tabela (fragment tabeli BD).
Sekcja uzupełniona jako efekt zadania Tydzień 08 wg Harmonogramu Zajęć na WIKAMP KSR.

3. Atrybuty i liczności obiektów wyrażone zmiennymi lingwistycznymi

Zmienne lingwistyczne dla wybranych 10 atrybutów z bazy danych, przedstawione w formie wykresów funkcji przynależności i wzorów analitycznych, wymienione etykiety oraz objaśnione wszystkie symbole ułatwiające czytelnikowi ich zrozumienie [2]. Zbędne jest cytowanie definicji. Konieczne precy-

zyjnie podane przestrzenie rozważań każdej zmiennej lingwistycznej, wzory i wykresy dla każdej wartości/etykiety.
Jw. kwantyfikatory lingwistyczne – opisane etykietami, wykresami funkcji przynależności i wzorami analitycznymi. Uzasadnione wiedzą dziedzinową wybrane zakresy i etykiety. Precyzyjnie podane przestrzenie rozważań każdego kwantyfikatora lingwistycznego/rozmytego, wzory i wykresy dla każdej wartości/etykiety. Opisy własne z przypisami do literatury, tak by inżynier innej specjalności zrozumiał dalszy opis tego konkretnego ćwiczenia/eksperymentu.
Sekcja uzupełniona jako efekt zadania Tydzień 09 wg Harmonogramu Zajęć na WIKAMP KSR.

4. Narzędzia obliczeniowe: projekt (wybór, implementacja) i diagram UML pakietu obliczeń rozmytych. Diagram UML generatora podsumowań

4.1. Diagram pakietu obliczeń rozmytych

Diagram UML i zwięzły opis pakietu obliczeń rozmytych: źródło pakietu (zewnętrzny/własny/hybrydowy), przypis do literatury. Krótka charakterystyka najważniejszych klas i podstawowych dla zadania ich metod.
Sekcja uzupełniona jako efekt zadania Tydzień 10 wg Harmonogramu Zajęć na WIKAMP KSR.

4.2. Diagram UML generatora podsumowań. Krótka instrukcja użytkownika

Diagram UML generatora podsumowań (warstwy obliczeniowej oraz interfejsu użytkownika). Krótki ilustrowany opis jak użytkownik może korzystać z aplikacji, w szczególności wprowadzać parametry podsumowań, odczytywać wyniki oraz definiować własne etykiety i kwantyfikatory. Wersja JRE i inne wymogi niezbędne do uruchomienia aplikacji przez użytkownika na własnym komputerze.
Sekcja uzupełniona jako efekt zadania Tydzień 11 wg Harmonogramu Zajęć na WIKAMP KSR.

5. Jednopodmiotowe podsumowania lingwistyczne. Miary jakości, podsumowanie optymalne

Wyniki kolejnych eksperymentów wg punktów 2.-4. opisu projektu 2. Listy podsumowań jednopodmiotowych i tabele/rankingi podsumowań dla danych atrybutów obowiązkowe i dokładnie opisane w „captions” (tytułach), konieczny opis kolumn i wierszy tabel. Dla każdego podsumowania podane miary jakości oraz miara jakości podsumowania optymalnego.
Sekcja uzupełniona jako efekt zadania Tydzień 11 wg Harmonogramu Zajęć na WIKAMP KSR.

6. Wielopodmiotowe podsumowania lingwistyczne i ich miary jakości

Wyniki kolejnych eksperymentów wg punktów 2.-4. opisu projektu 2. Uzasadnienie i metoda podziału zbioru danych na rozłączne podmioty. Listy podsumowań wielopodmiotowych i tabele/rankingi podsumowań dla danych atrybutów obowiązkowe i dokładnie opisane w „captions" (tytułach), konieczny opis kolumn i wierszy tabel. Konieczne uwzględnienie wszystkich 4-ch form podsumowań wielopodmiotowych.

** Możliwe sformułowanie zagadnienia wielopodmiotowego podsumowania optymalnego **.

Ewentualne wyniki realizacji punktu „na ocenę 5.0" wg opisu Projektu 2. i ich porównanie do wyników z części obowiązkowej.

Sekcja uzupełniona jako efekt zadania Tydzień 12 wg Harmonogramu Zajęć na WIKAMP KSR.

7. Dyskusja, wnioski

Dokładne interpretacje uzyskanych wyników w zależności od parametrów klasyfikacji opisanych w punktach 3.-4 opisu Projektu 2. Szczególnie istotne są wnioski o charakterze uniwersalnym, istotne dla podobnych zadań. Omówić i wyjaśnić napotkane problemy (jeśli były). Każdy wniosek/problem powinien mieć poparcie w przeprowadzonych eksperymentach (odwołania do konkretnych wyników: tabel i miar jakości). Ocena które wybrane kwantyfikatory, sumaryzatory, kwalifikatory i/lub ich miary jakości mają małe albo duże znaczenie dla wiarygodności i jakości otrzymanych agregacji/podsumowań. Dla końcowej oceny jest to najważniejsza sekcja sprawozdania, gdyż prezentuje poziom zrozumienia rozwiązywanego problemu.

** Możliwości kontynuacji prac w obszarze logiki rozmytej i wnioskowania rozmytego, zwłaszcza w kontekście pracy inżynierskiej, magisterskiej, naukowej, itp. **

Sekcja uzupełniona jako efekt zadań Tydzień 11 i Tydzień 12 wg Harmonogramu Zajęć na WIKAMP KSR.

8. Braki w realizacji projektu 2.

Wymienić wg opisu Projektu 2. wszystkie niezrealizowane obowiązkowe elementy projektu, ewentualnie podać merytoryczne (ale nie czasowe) przyczyny tych braków.

Literatura

[1] A. Niewiadomski, Zbiory rozmyte typu 2. Zastosowania w reprezentowaniu informacji. Seria „Problemy współczesnej informatyki" pod redakcją L. Rutkowskiego. Akademicka Oficyna Wydawnicza EXIT, Warszawa, 2019.

[2] S. Zadrożny, Zapytania nieprecyzyjne i lingwistyczne podsumowania baz danych, EXIT, 2006, Warszawa

[3] A. Niewiadomski, Methods for the Linguistic Summarization of Data: Applications of Fuzzy Sets and Their Extensions, Akademicka Oficyna Wydawnicza EXIT, Warszawa, 2008.

Literatura zawiera wyłącznie źródła recenzowane i/lub o potwierdzonej wiarygodności, możliwe do weryfikacji i cytowane w sprawozdaniu.