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 pozwalającej na generowanie podsumowań lingwistycznych[1] w oparciu o kwantyfikatory rozmyte[2], co oznacza opisanie danych liczbowych ze zbioru danych[3] językiem quasi-naturalnym - pozornie naturalnym.
Przykładem podsumowania lingwistycznego[1] w formie

$$Q\ P\ jest\ S\ [T] \tag{1}$$

jest: *Wiele wypadków jest przy ujemnej temperaturze [0,76]*, gdzie Q jest kwantyfikatorem lingwistycznym, P podmiotem podsumowań, S sumaryzatorem, a *T[0, 1]* stopniem prawdziwości.

Przykładem drugiego podsumowania lingwistycznego w formie:

$$Q\ P\ bedacych\ W\ jest\ S\ [T] \tag{2}$$

jest: *Wiele wypadków będących podczas deszczu, jest przy ujemnej temperaturze [0.68]*, gdzie Q jest kwantyfikatorem lingwistycznym, P podmiotem podsumowań, S sumaryzatorem, W kwantyfikatorem reprezentującym dodatkowe własności obiektów, a *T[0, 1]* stopniem prawdziwości.

Analizowany zbior danych zawiera liczbowe informacje o ponad 3 milionach wypadków samochodowych w 49 stanach Zjednoczonych Stanów Ameryki,

mających miejsce od lutego 2016 do grudnia 2020[3]. Zbiór danych składa się z 47 kolumn. W tym celu wykonania podsumowania lingwistycznego zostaną wykorzystane metody logiki rozmytej[4]. Logika rozmyta pozwala na opisanie wartości zapisanych językiem naturalnym za pomocą zrozumiałych określeń jak: mało, dużo, około połowy. W projekcie zostaną wykorzystane kwantyfikatory lingwistyczne względne takie jak: niewiele, około połowy oraz kwantyfikatory lingwistyczne absolutne takie jak: około jednego, około stu.

2. Charakterystyka podsumowywanej bazy danych

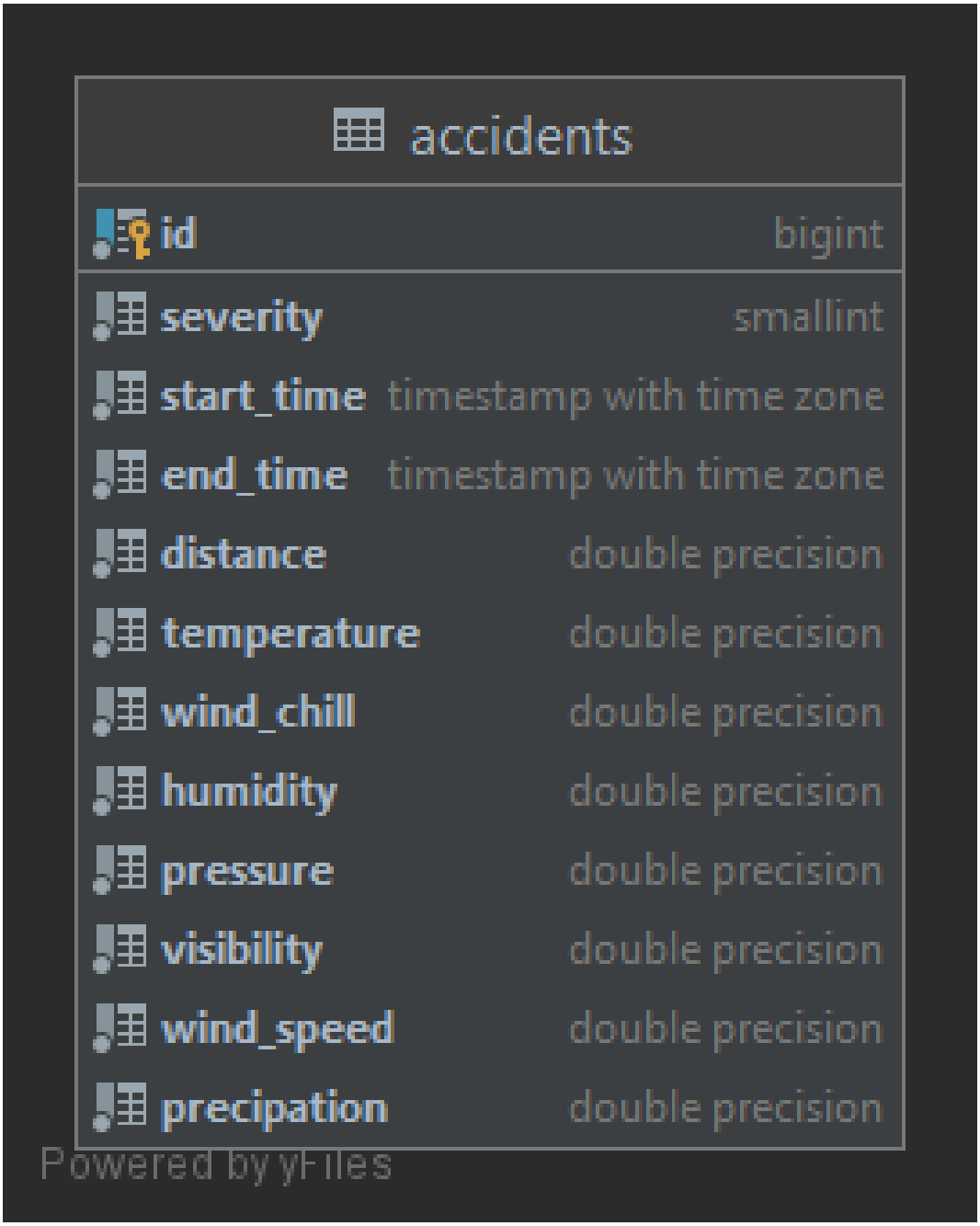
W programie został użyty zbiór danych[3] znajdujący się w pliku CSV, który został przekształcony w bazę danych.

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 następujące 11 kolumn:

— Czas rozpoczęcia - Start_Time - czas rozpoczęcia się wypadku w lokalnej strefie czasowej, przyjmuje wartości od 8 lutego 2016, do 31 grudnia 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 wyrażony w milach, na którego miał wpływ wypadek. Przyjmuje wartości zmiennoprzecinkowe od 0 do 334, gdzie zdecydowana większość danych mieści się w przedziale od 0.00 do 4.00.
— Temperatura - Temperature - temperatura powietrza wyrażona w Fahrenheit'ach, w momencie, gdy zdarzył się wypadek. Przyjmuje wartości zmiennoprzecinkowe od -16.00 do 104.00. Temperature można opisac jako bardzo zimną, zimną, umiarkowaną, ciepłą, bardzo ciepłą. Oczywiście jest to opis subiektywny.
— Temperatura odczuwalna - Wind_Chill - temperatura odczuwalna wyrażona w Fahrenheit'ach, w momencie, gdy zdarzył się wypadek. Przyjmuje wartości zmiennoprzecinkowe od -16.00 do 101.00. Temperaturę odczuwalną mozna opisać tak samo jak temperaturę.
— Wilgotność - Humidity - wilgotność powietrza wyrażona w procentach w momencie, gdy zdarzył się wypadek. Przyjmuje wartości zmiennoprzecinkowe od 4.00 do 100.00.
— Ciśnienie - Pressure - ciśnienie powietrza wyrażone w inches, w momencie, gdy zdarzył się wypadek. Przyjmuje wartości zmiennoprzecinkowe od 27.00 do 32.00. Ciśnieje można opisac jako wysokie, umiarkowane lub niskie.
— Widoczność - Visibiity - widoczność wyrażona w milach, w momencie, gdy zdarzył się wypadek. Przyjmuje wartości zmiennoprzecinkowe od 0.00 do 12.00. Widoczność mozna opisać jako dobrą, ograniczoną, słabą.

— Prędkość wiatru - Wind_Speed - prędkość wiatru wyrażona w milach na godzinę, w momencie, gdy zdarzył się wypadek. Przyjmuje wartości zmiennoprzecinkowe od 0.00 do 40.00. Wiatr mozna opisać jako słaby, umiarkowany, silny.
— Ilość opadów - Principation - ilość opadów wyrażona w inches, w momencie, gdy zdarzył się wypadek. Jeśli opady nie występowały to kolumna przyjmuje wartość nan. Przyjmuje wartości zmiennoprzecinkowe od 0.00 do 0.50.

Atrybutom nadawane są opisane zwyczajowe wartości lingwistyczne ze względu na zwiększenie przystępności i ułatwienie szybkiego zrozumienia atrybutu przez człowieka, kiedy ten atrybut nie musi być dokładnie opisany. Przykładowo temperatura, mimo że zrozumiała dla człowieka w postaci liczbowej, jest łatwiejsza do szybszego zrozumienia w postaci tekstowej, a dla ludzi nie ma dużego znaczenia czy temperatura rózni się o 1 czy 2 stopnie, wystarczy opisać ją słownie tak jak wcześniej podaliśmy jako bardzo zimną, zimną, umiarkowaną, ciepłą, bardzo ciepłą.



Rysunek 1. Tabela reprezentująca omawiane dane wykonana w DBMS Postgresql

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 [6]. Zbędne jest cytowanie definicji. Konieczne precyzyjnie 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, nauko-

wej, 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] I. Superson, A. Niewiadomski, POZYSKIWANIE WIEDZY Z RELACYJNYCH BAZ DANYCH: WIELOPODMIOTOWE PODSUMOWANIA LINGWISTYCZNE, Politechnika Łódzka

[2] A. Niewiadomski, Rozmyte metody inteligentnej interpretacji danych, tom 10, 2006, 546-547

[3] 2021 Kaggle Inc [internetowa społeczność związana z analizą danych], US Accidents (3 million records – updated) A Countrywide Traffic Accident Dataset (2016 - 2020) [przeglądany 24 kwietnia 2021], Dostępny w: `https://www.kaggle.com/sobhanmoosavi/us-accidents` Oficyna Wydawnicza EXIT, Warszawa, 2008.

[4] Zadeh L.A., A computational approach to fuzzy quantifiers in natural languages. Computers and Maths with Applications, nr 9, 1983, 149-183

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

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

[7] A. Niewiadomski, Methods for the Linguistic Summarization of Data: Applications of Fuzzy Sets and Their Extensions, Akademicka

[8] Zadeh, L. A.: 1965, 'Fuzzy sets'. Inf. and Control 8, 338–353.

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