Informatyka, studia dzienne, I st. semestr VI

Komputerowe systemy rozpoznawania 2019/2020

Prowadzący: dr hab. inż. Adam Niewiadomski, profesor uczelni pon, 12:00

Data oddania: _____ Ocena: _____

Aleksandra Wnuk 216924

Michał Kidawa 216796

Zadanie 2: Lingwistyczne podsumowania baz danych

1. Cel

Celem zadania jest stworzenie aplikacji, której główną funkcjonalnością jest lingwistyczna agregacja zawartości wybranego zbioru danych, czyli wygenerowanie w języku quasi-naturalnym opisu zawartości danych w zbiorze.

2. Wprowadzenie

Logika aplikacji bazuje na zagadnieniach z zakresu zbiorów rozmytych oraz ich zastosowaniu w generowaniu podsumowań lingwistycznych baz danych.

2.1. Podsumowania lingwistyczne

Według literatury [4] możemy wyróżnić następujące formy podsumowań lingwistycznych:

2.1.1. Podsumowania jednopodmiotowe

$$Q\ P \text{ are/have } S\ [T] \tag{1}$$

$$Q\ P \text{ being/having } W \text{ are/have } S\ [T] \tag{2}$$

gdzie Q - kwantyfikator, P - podmiot podsumowania, W - kwalifikator, S - sumaryzator, lub kilka sumaryzatorów połączonych spójnikiem AND, lub OR. T natomiast jest wartością miary, lub miar jakości podsumowania.

2.1.2. Podsumowania wielopodmiotowe

$$Q\ P_1 \text{ compared to } P_2 \text{ are/have } S\ [T] \tag{3}$$

$$Q\ P_1 \text{ compared to } P_2 \text{ being/having } W \text{ are/have } S\ [T] \tag{4}$$

$$Q\ P_1 \text{ being/having } W \text{ compared to } P_2 \text{ are/have } S\ [T] \tag{5}$$

$$\text{More } P_1 \text{ than } P_2 \text{ are/have } S\ [T] \tag{6}$$

gdzie P_1 i P_2 to pierwszy i drugi podmiot podsumowania, a reszta symboli jak w punkcie 2.1.1.

2.2. Kwantyfikator

Jest to określenie ilości, zdefiniowane przez pewien zbiór rozmyty. Wyróżniamy kwantyfikatory względne np.: "Wiele", "Kilka" i bezwzględne np. "Ponad 40000". W naszym programie funkcje przynależności kwantyfikatorów względnych zostały zaimplementowane zgodnie z przykładami w literaturze [2].

2.3. Sumaryzatory i Kwalifikatory

Zarówno sumaryzatory, jak i kwalifikatory są określeniami odnoszącymi się do dodanego atrybutu podmiotu np.: dla naszego programu, gdzie podmiotem są biegi w wyścigach konnych sumaryzator / kwalifikator to etykiety zmiennych lingwistycznych np. "Waga konia średnia", lub "Wiek konia młody" wraz z ich przestrzeniami rozważań oraz funkcjami przynależności.

2.4. Miary jakości podsumowań lingwistycznych

Dla każdego podsumowania można wyznaczyć miarę jego jakości. Podstawową miarą jakości jest *degree of truth* - poziom prawdziwości danego podsumowania, który przyjmuje wartość z przedziału [0,1]. Poza poziomem prawdziwości określanym jako T1, można wyznaczyć również inne miary (od T2 do T11). Każda z nich dotyczy innej właściwości wygenerowanego podsumowania.

— T1 - stopień prawdziwości
— T2 - stopień nieprecyzyjności
— T3 - stopień pokrycia
— T4 - stopień trafności
— T5 - długość podsumowania
— T6 - stopień nieprecyzyjności kwantyfikatora
— T7 - stopień kardylności względnej kwantyfikatora

— T8 - stopień kardylności względnej sumaryzatora
— T9 - stopień nieprecyzyjności kwalifikatora
— T10 - stopień kardylności względnej kwalifikatora
— T11 - długość kwalifikatora

Posiadając dane o wartości wszystkich miar (od T1 do T11) jesteśmy w stanie wznaczyć ogólną jakość podsumowania (podsumowanie optymalne), którą można wyliczyć za pomocą średniej ważonej, gdzie suma wszystkich wag musi być równa 1.

$$T = \sum_{i=1}^{11} w_i T_i \tag{7}$$

gdzie w_i - znormalizowane miary jakości, których suma daje 1, T_i - wartość miary jakości i.

2.5. Funkcje przynależności

Wykorzystaliśmy trzy podstawowe funkcje przynależności:

2.5.1. Funkcja trójkątna

Definiujemy zbiór rozmyty o trójkątnej funkcji przynależności o parametrach a, b, m wtedy i tylko wtedy, gdy $a \leq m \leq b$ oraz:

$$\mu_A(x) = \begin{cases} 0 & \text{gdy } x \in (-\infty, a] \\ (x-a)/(m-a) & \text{gdy } x \in (a, m) \\ 1 & \text{gdy } x = m \\ (b-x)/(b-m) & \text{gdy } x \in (m, b) \\ 0 & \text{gdy } x \in [b, +\infty) \end{cases}$$

2.5.2. Funkcja trapezoidalna

Definiujemy zbiór rozmyty o trapezoidalnej funkcji przynależności o parametrach a, b, m, n wtedy i tylko wtedy, gdy $a \leq m \leq n \leq b$ oraz:

$$\mu_A(x) = \begin{cases} 0 & \text{gdy } x \in (-\infty, a] \\ (x-a)/(m-a) & \text{gdy } x \in (a, m) \\ 1 & \text{gdy } x \in [m, n] \\ (b-x)/(b-n) & \text{gdy } x \in (n, b) \\ 0 & \text{gdy } x \in [b, +\infty) \end{cases}$$

2.5.3. Funkcja gaussowska

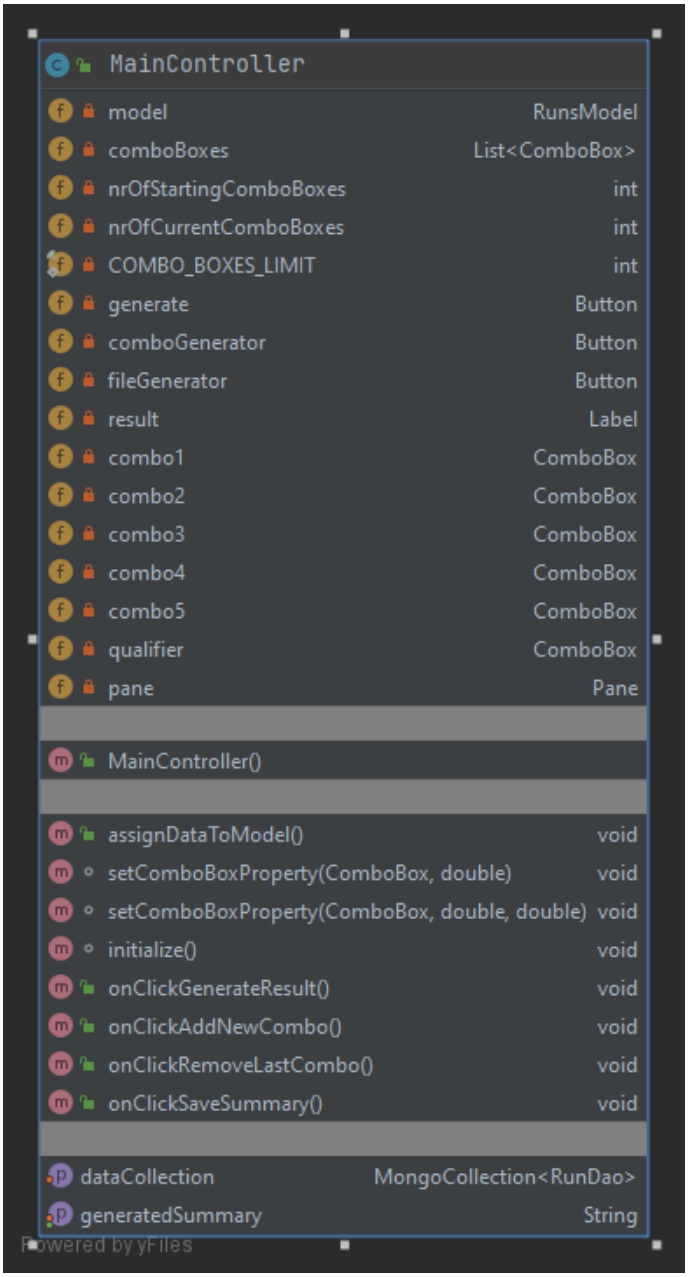
Zbiory o gaussowskiej funkcji przynależności charakteryzują się funkcjami przynależności o kształcie krzywej Gaussa:
$\mu_A(x) = e^{-(\frac{x-a}{b})^2}$

3. Opis implementacji

Aplikacja została zaimplementowana w języku Java w wersji 8, jako aplikacja z interfejsem graficznym utworzonym za pomocą biblioteki JavaFX. Przy implementacji programu wykorzystaliśmy wzorzec architektoniczny MVC - Model-Widok-Kontroler. Baza danych opisywana w dalszych sekcjach sprawozdania to nierelecyjna baza No-SQL MongoDB.

3.1. Pakiet controller

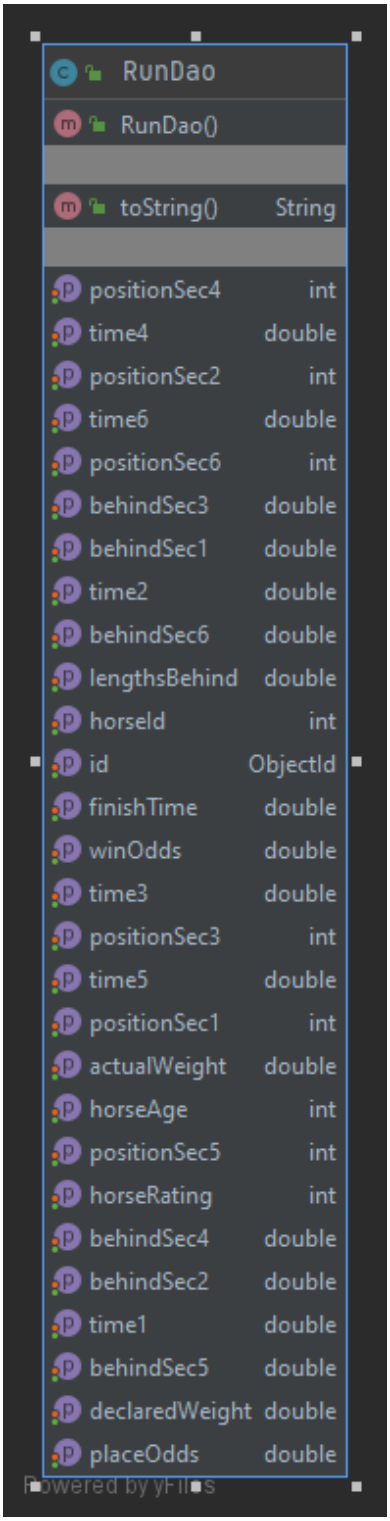
Pakiet zawiera klasę MainController - główny kontroler aplikacji połączony z widokiem generowania podsumowań lingwistycznych.



Rysunek 1. Diagram UML pakietu controller

3.2. Pakiet dao

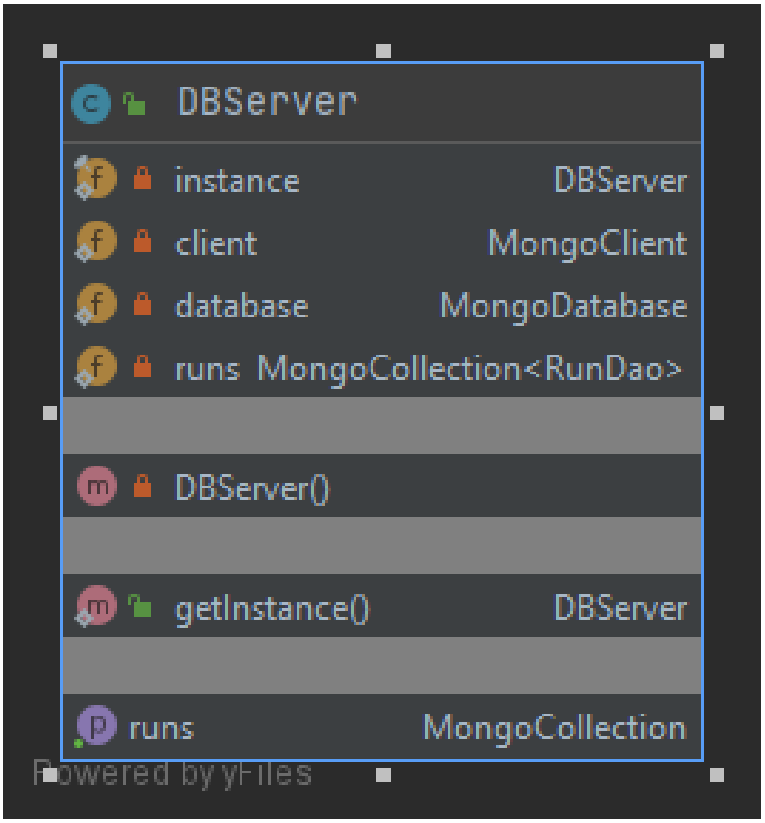
Pakiet zawiera klasę RunsDao - odpowiada za mapowanie obiektowo-relacyjne elementów z bazy danych.



Rysunek 2. Diagram UML pakietu dao

3.3. Pakiet service

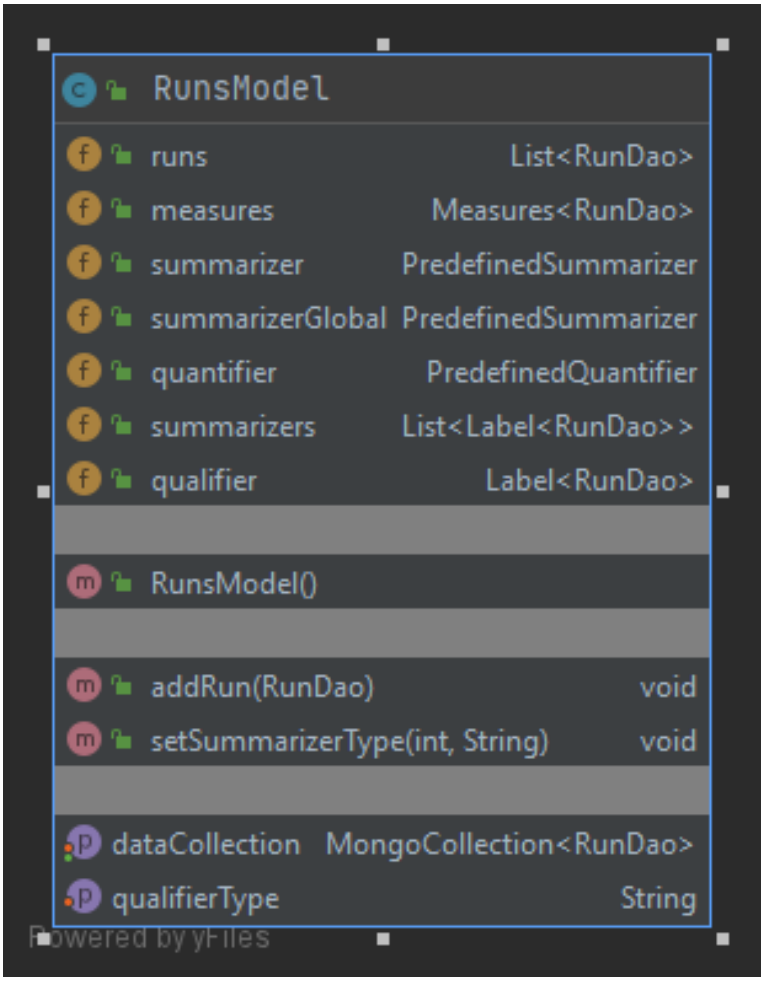
Pakiet ten zawiera klasę DBServer - zaimplementowaną zgodnie ze wzorcem projektowym Singleton klasę umożliwiającą nawiązanie połączenia z klastrem MongoDB znajdującym się w chmurze i dostęp do kolekcji danych.



Rysunek 3. Diagram UML pakietu service

3.4. Pakiet model

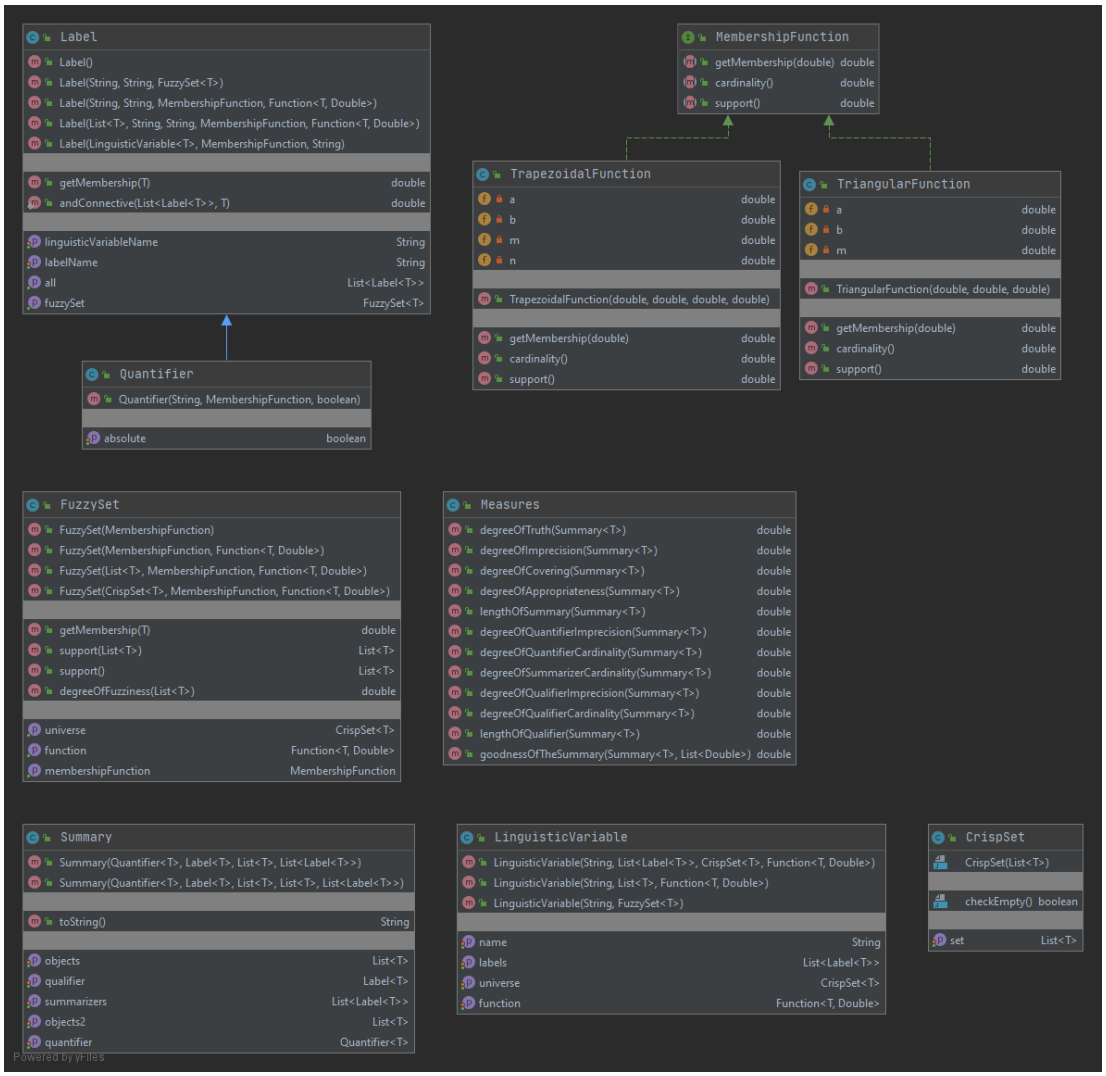
Pakiet zawiera klasę RundsModel, która przechowuje wszystkie obiekty odpowiedzialne za logikę biznesową aplikacji.



Rysunek 4. Diagram UML pakietu model

3.5. Pakiet fuzzylogic

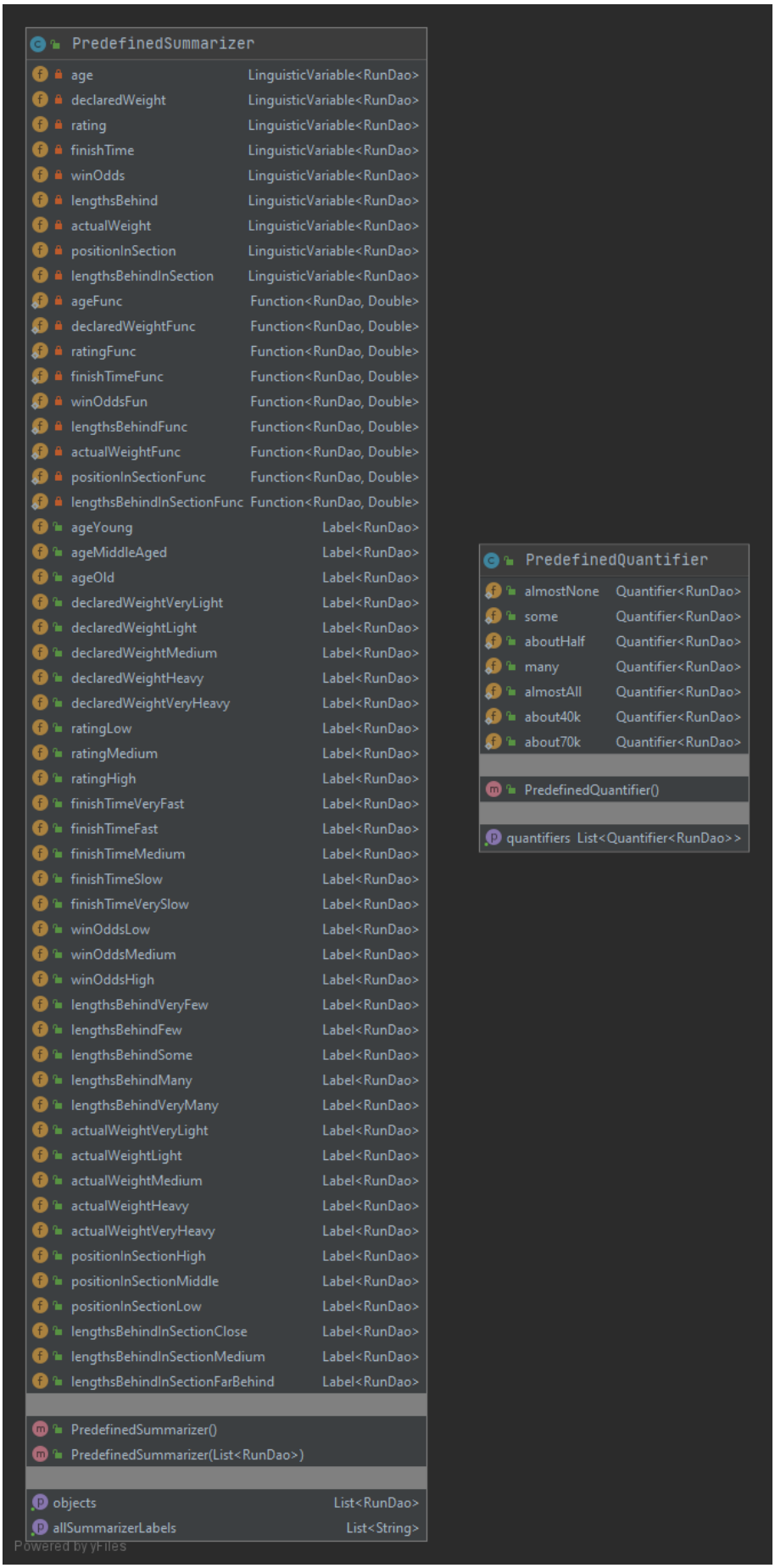
Pakiet ten zawiera klasy odpowiedzialne za implementację biblioteki realizującej obliczenia rozmyte. Dla uproszczenia implementacji zmienna lingwistyczna została rozłożona na pojedyncze zbiory rozmyte. Np. instancja klasy LinguisticVariable będzie zawierała zbiór rozmyty z jedną funkcją przynależności, np. biegi koni w wieku młodym.



Rysunek 5. Diagram UML pakietu fuzzylogic

3.6. Pakiet fuzzyruns

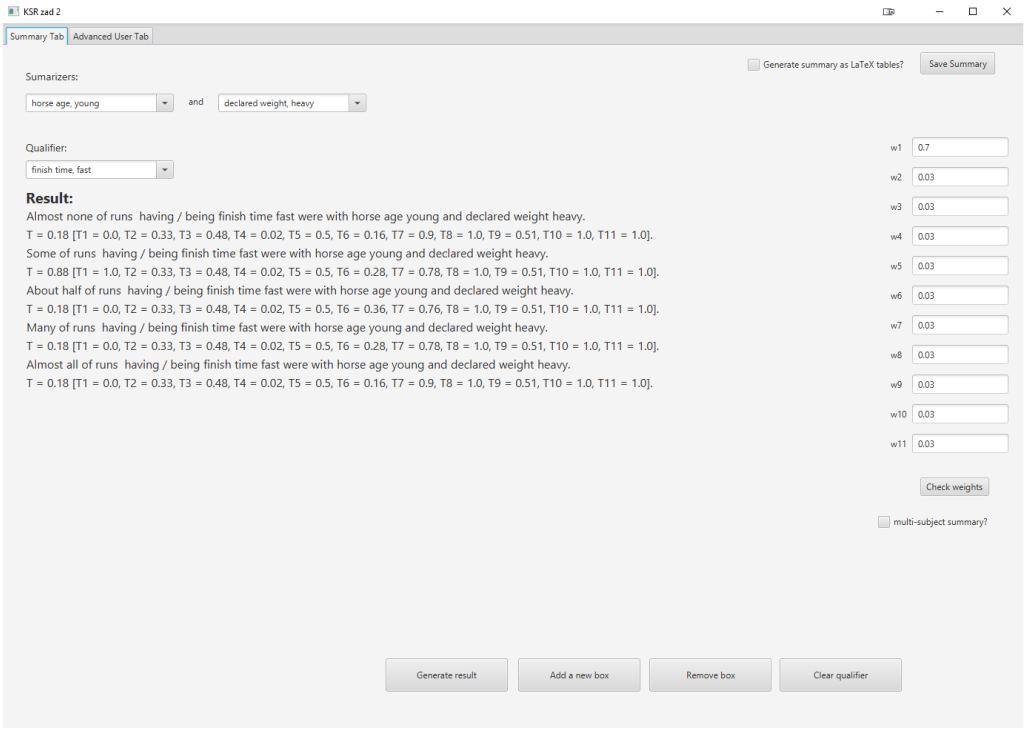
Pakiet zawiera klasy implementujące zdefiniowane przez nas kwantyfikatory oraz sumaryzatory, które są polami tych klas.



Rysunek 6. Diagram UML pakietu fuzzyruns

3.7. Interfejs użytkownika

3.7.1. Podstawowe funkcjonalności



Rysunek 7. Zrzut ekranu przedstawiający zakładkę z podstawowymi funkcjonalnościami aplikacji

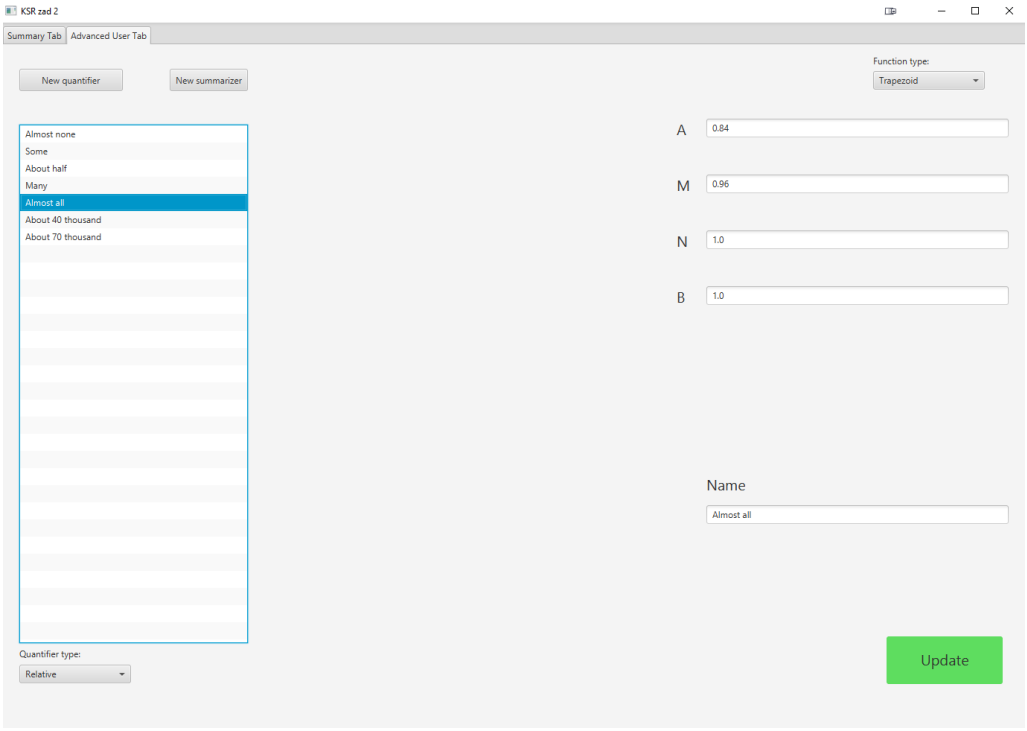
Po uruchomieniu, stworzony przez nas program przenosi użytkownika do zakładki umożliwiającej generowanie podsumowań lingwistycznych. Domyślnie tworzone jest jedno pole z wyborem dostępnych sumaryzatorów, jednak za pomocą odpowiedniego przycisku istnieje możliwość zwiększenia liczby sumaryzatorów do 5.
Poniżej za pomocą rozwijanej listy istnieje możliwość wyboru kwalifikatora wykorzystanego w podsumowaniu.
Po prawej stronie znajdują się pola tekstowe zawierające wartośći poszczególnych wagi miar jakości wykorzystywane przy obliczaniu podsumowania optymalnego.
Poniżej pól tekstowych z wagami znajduje się pole wyboru. Jeśli zostanie wybrane umożliwi użytkownikowi wygenerowanie podsumowań wielopodpiotowych. Baza została podzielona na podmioty względem atrybutu horse_type. Po zaznaczeniu opcji podsumowania wielopodmiotowego wyświetlą się dwie rozwijane listy z możliwymi wyborami podmiotów.
Po naciśnięciu przycisku z napisem "Generate result" pod napisem "Result" pojawi się wygenerowane podsumowanie lingwistyczne.

3.8. Zapisywanie do pliku

Po utworzeniu podsumowania lingwistycznego istnieje możliwość zapisania go do pliku tesktowego. Aby to zrobić należy nacisnąć guzik "Save

Summary" znajdujący się w prawym górnym rogu.
Aby ułatwić pracę ze sprawozdanie zaimplementowaliśmy również funkcjonalność pozwalającą na generowanie wyników podsumowań, w formie tabeli zgodnych ze standardem języka LaTeX.

3.9. Funkcjonalności użytkownika zaawansowanego



Rysunek 8. Zrzut ekranu przedstawiający zakładkę z funkcjonalnościami użytkownika zaawansowanego

Wchodząc w zakładkę użytkownika zaawansowanego, istnieje możliwość utworzenia nowych kwantyfikatorów, kwalifikatorów oraz sumaryzatorów, a także edycja już istniejących. W zależności od wciśniętego przycisku ("New quantifier" / "New summarizer") wyświetlona zostanie lista zawierająca zdefiniowane już w programie kwantyfikatory, bądź kwalifikatory i sumaryzatory. Po naciśnięciu na jeden z elementów listy, pola tekstowe z parametrami zostaną automatycznie wypełnione. Jeżeli zmieniona zostanie wartość jakiegoś z parametrów, a następnie użytkownik naciśnie przycisk "Update", nastąpi aktualizacja danego elementu. Jeżeli natomiast w pola tekstowa związane z nazwą zostanie wpisana nowa pozycja, to po uzupełnieniu odpowiednich pól i naciśnięciu przycisku "Create" zostanie utworzony nowy kwantyfikator bądź sumaryzator / kwalifikator. Po wprowadzeniu jakichkolwiek zmian cały interfejs zostanie zaktualizowany i wprowadzone nowe elementy będą dostępne do wykorzystania przez użytkownika przy generowaniu podsumowań lingwistycznych. Jeżeli jedno z niezbędnych pól nie zostało wypełnione użytkownik zostanie poinformowany o tym powiadomieniem.

4. Materiały i metody

Baza wybrana przez nas do realizacji zadania to zbiór danych nt. wyścigów konnych w Hong Kongu w latach 1997 - 2005. Baza dostępna jest na licencji CC0 w serwisie kaggle.com [1]. Baza składa się z dwóch plików w formacie csv, jeden opisujący wyścigi (jeden rekord odpowiada jednemu wyścigowi, który odbył się w danym dniu), natomiast drugi przedstawiający indywidualne biegi danego konia w danym wyścigu. Na potrzeby zadania wykorzystana została jedynie druga tabela. Liczba rekordów znajdujących się w tabeli wynosi 79447. Kolumny znajdujące się w tabeli:

— race_id - identyfikator wyścigu
— horse_no - numer przypisany danemu koniowi w danym wyścigu
— horse_id - identyfikator konia
— result - pozycja, którą zajął dany koń
— won - wartość binarna opisująca, czy dany koń wygrał wyścig (1 jeśli tak, 0 jeśli nie)
— lengths_behind - pozycja końcowa danego konia określona jako liczba długości konia za zwycięzcą
— horse_age - wiek konia w dniu wyścigu
— horse_country - kraj pochodzenia konia
— horse_type - typ konia, jeden z terminów:
  — Brown - koń gniady (umaszczenie)
  — Colt - ogier poniżej 4 roku życia, niewykastrowany
  — Filly - klacz poniżej 4 roku życia
  — Gelding - ogier wykastrowany (wałach)
  — Grey - koń siwy (umaszczenie)
  — Horse - dorosły ogier niewykastrowany
  — Mare - klacz powyżej 4 roku życia
  — Rig - koń z wnętrostwem
  — Roan - deresz (umaszczenie)

  W większości przypadków w tej kolumnie znajduje się określenie niedotyczące umaszczenia, mimo że określenia te się nie wykluczają, np. koń może być jednocześnie dereszem i ogierem powyżej 4 roku życia, ale w bazie znajduje się tylko jeden termin.
— horse_rating - ocena konia wg HKJC (Hong Kong Jockey Club)
— horse_gear - opis sprzętu posiadanego przez konia
— declared_weight - zadeklarowana waga konia i dżokeja w funtach
— actual_weight - rzeczywista waga niesiona przez konia w funtach
— draw - miejsce startowe konia
— position_sec1 - pozycja danego konia na pierwszym odcinku wyścigu
— position_sec2 - pozycja danego konia na drugim odcinku wyścigu
— position_sec3 - pozycja danego konia na trzecim odcinku wyścigu
— position_sec4 - pozycja danego konia na czwartym odcinku wyścigu
— position_sec5 - pozycja danego konia na piątym odcinku wyścigu
— position_sec6 - pozycja danego konia na szóstym odcinku wyścigu
— behind_sec1 - pozycja danego konia wyrażona w długościach za prowadzącym na pierwszym odcinku wyścigu

— behind_sec2 - pozycja danego konia wyrażona w długościach za prowadzącym na drugim odcinku wyścigu
— behind_sec3 - pozycja danego konia wyrażona w długościach za prowadzącym na trzecim odcinku wyścigu
— behind_sec4 - pozycja danego konia wyrażona w długościach za prowadzącym na czwartym odcinku wyścigu
— behind_sec5 - pozycja danego konia wyrażona w długościach za prowadzącym na piątym odcinku wyścigu
— behind_sec6 - pozycja danego konia wyrażona w długościach za prowadzącym na szóstym odcinku wyścigu
— time1 - czas, jaki zajęło danemu koniowi przebiegnięcie pierwszego odcinka wyścigu (w sekundach)
— time2 - czas, jaki zajęło danemu koniowi przebiegnięcie drugiego odcinka wyścigu (w sekundach)
— time3 - czas, jaki zajęło danemu koniowi przebiegnięcie trzeciego odcinka wyścigu (w sekundach)
— time4 - czas, jaki zajęło danemu koniowi przebiegnięcie czwartego odcinka wyścigu (w sekundach)
— time5 - czas, jaki zajęło danemu koniowi przebiegnięcie piątego odcinka wyścigu (w sekundach)
— time6 - czas, jaki zajęło danemu koniowi przebiegnięcie szóstego odcinka wyścigu (w sekundach)
— finish_time - czas końcowy danego konia (w sekundach)
— win_odds - szansa na wygraną danego konia na początku wyścigu
— place_odds - szansa na zdobycie pierwszego, drugiego lub trzeciego miejsca na początku wyścigu
— trainer_id - identyfikator trenera konia
— jockey_id - identyfikator dżokeja

Atrybuty możliwe do rozmycia:

1. lengths_behind - wartości od 0.0 do 8.75, skok o 0.25
2. horse_age - wartości od 2 do 10, skok o 1
3. horse_rating - wartości od 10 do 138, skok o 1
4. declared_weight - wartości od 693 do 1369, skok o 1
5. actual_weight - wartości od 103 do 133, skok o 1
6. position_sec$n$ - wartości od 1 do 14, skok o 1
7. behind_sec$n$ - wartości od 0.0 do 98.75 (z wyłączeniem pojedynczych rekordów z wartością 999.0), skok o 0.25
8. time$n$ - wartości od 12.39 do 73.74 (z wyłączeniem pojedynczych rekordów z wartością 999.0), skok o 0.01
9. finish_time - wartości od 55.16 do 163.58, skok o 0.01
10. win_odds - wartości od 1.0 do 99.9, skok o 0.1
11. place_odds - wartości od 1.0 do 92.0, skok o 0.1

Wyścigi miały długość od 1000m do 2400m (co jest opisane w drugiej tabeli bazy), jednak dzięki podziałowi wyścigów na odcinki możliwe jest porów-

nywanie ze sobą wszystkich rekordów bez uwzględniania całkowitej długości wyścigu. Jedynym atrybutem, który może być problematyczny do uwzględnienia bez brania pod uwagę całkowitej długości wyścigu jest finish_time.

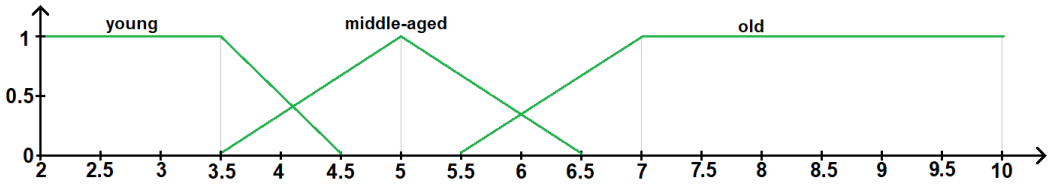
Zmienne lingwistyczne:

— L1 = ⟨horse_age, H(horse_age) = {young, middle-aged, old}, X = [2, 10]⟩

$$\mu_{young}(x) = \begin{cases} 1 & \text{dla } x \in [3, 3.5] \\ 4.5 - x & \text{dla } x \in (3.5, 4.5] \end{cases}$$

$$\mu_{middle-aged}(x) = \begin{cases} \frac{2}{3}x - \frac{7}{3} & \text{dla } x \in [3.5, 5) \\ 1 & \text{dla } x = 5 \\ \frac{13}{3} - \frac{2}{3}x & \text{dla } x \in (5, 6.5] \end{cases}$$

$$\mu_{old}(x) = \begin{cases} \frac{2}{3}x - \frac{11}{3} & \text{dla } x \in (5.5, 7] \\ 1 & \text{dla } x \in (7, 10] \end{cases}$$



Rysunek 9. Funkcja przynależności dla zmiennej L1

— L2 = ⟨horse_rating, H(horse_rating) = {low rating, medium rating, high rating}, X = [10, 138]⟩

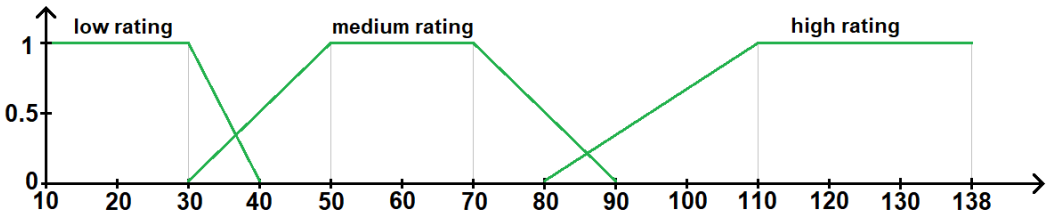
$$\mu_{low-rating}(x) = \begin{cases} 1 & \text{dla } x \in [10, 30] \\ 4 - \frac{1}{10}x & \text{dla } x \in (30, 40] \end{cases}$$

$$\mu_{medium-rating}(x) = \begin{cases} \frac{1}{20}x - \frac{3}{2} & \text{dla } x \in [30, 50) \\ 1 & \text{dla } x \in [50, 70) \\ \frac{9}{2} - \frac{1}{20}x & \text{dla } x \in (70, 90] \end{cases}$$

$$\mu_{high-rating}(x) = \begin{cases} \frac{1}{30}x - \frac{8}{3} & \text{dla } x \in [80, 110) \\ 1 & \text{dla } x \in [110, 138] \end{cases}$$

— L3 = ⟨declared_weight, H(declared_weight) = {very light, light, medium, heavy, very heavy}, X = [693, 1396]⟩

$$\mu_{very-light}(x) = \begin{cases} 1 & \text{dla } x \in [693, 800] \\ 17 - \frac{1}{50}x & \text{dla } x \in (800, 850] \end{cases}$$



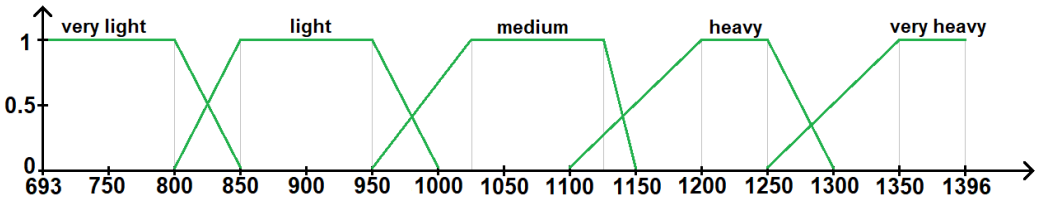
Rysunek 10. Funkcja przynależności dla zmiennej L2

$$\mu_{light}(x) = \begin{cases} \frac{1}{50}x - 16 & \text{dla } x \in [800, 850) \\ 1 & \text{dla } x \in [850, 950) \\ 20 - \frac{1}{50}x & \text{dla } x \in (950, 1000] \end{cases}$$

$$\mu_{medium}(x) = \begin{cases} \frac{1}{75}x - \frac{38}{3} & \text{dla } x \in [950, 1025) \\ 1 & \text{dla } x \in [1025, 1125) \\ 46 - \frac{1}{25}x & \text{dla } x \in (1125, 1150] \end{cases}$$

$$\mu_{heavy}(x) = \begin{cases} \frac{1}{100}x - 11 & \text{dla } x \in [1100, 1200) \\ 1 & \text{dla } x \in [1200, 1250) \\ 26 - \frac{1}{50}x & \text{dla } x \in (1250, 1300] \end{cases}$$

$$\mu_{very-heavy}(x) = \begin{cases} \frac{1}{100}x - \frac{25}{2} & \text{dla } x \in [1250, 1350) \\ 1 & \text{dla } x \in [1350, 1396] \end{cases}$$



Rysunek 11. Funkcja przynależności dla zmiennej L3

— L4 = ⟨finish_time, H(finish_time) = {very fast, fast, medium, slow, very slow}, X = [55.16, 163.58]⟩

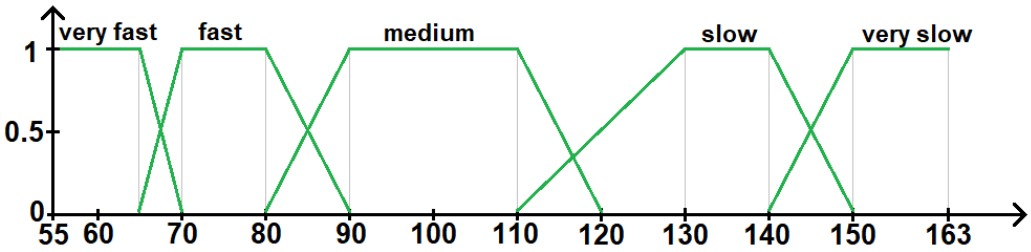
$$\mu_{very-fast}(x) = \begin{cases} 1 & \text{dla } x \in [55.16, 65.0] \\ 14 - \frac{1}{5}x & \text{dla } x \in (65.0, 70.0] \end{cases}$$

$$\mu_{fast}(x) = \begin{cases} \frac{1}{5}x - 13 & \text{dla } x \in [65.0, 70.0) \\ 1 & \text{dla } x \in [850, 950) \\ 9 - \frac{1}{10}x & \text{dla } x \in (80.0, 90.0] \end{cases}$$

$$\mu_{medium}(x) = \begin{cases} \frac{1}{10}x - 8 & \text{dla } x \in [80.0, 90.0) \\ 1 & \text{dla } x \in [90.0, 110.0) \\ 12 - \frac{1}{10}x & \text{dla } x \in (110.0, 120.0] \end{cases}$$

$$\mu_{slow}(x) = \begin{cases} \frac{1}{20}x - \frac{11}{2} & \text{dla } x \in [110.0, 130.0) \\ 1 & \text{dla } x \in [130.0, 140.0) \\ 15 - \frac{1}{10}x & \text{dla } x \in (140.0, 150.0] \end{cases}$$

$$\mu_{very-slow}(x) = \begin{cases} \frac{1}{10}x - 140 & \text{dla } x \in [140.0, 150.0) \\ 1 & \text{dla } x \in [150.0, 163.58] \end{cases}$$



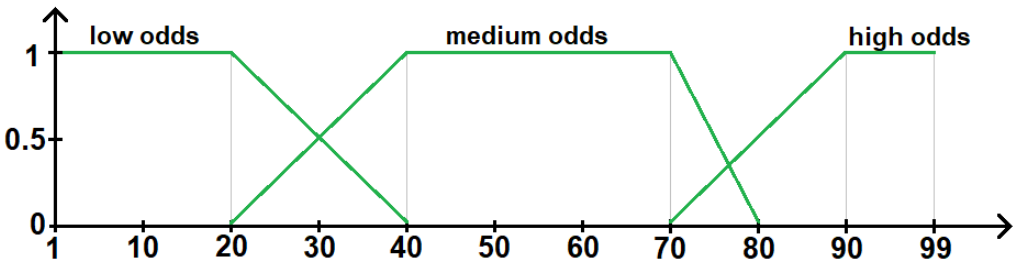
Rysunek 12. Funkcja przynależności dla zmiennej L4

— L5 = ⟨win_odds, H(win_odds) = {low odds, medium odds, high odds}, X = [1.0, 99.9]⟩

$$\mu_{low-odds}(x) = \begin{cases} 1 & \text{dla } x \in [1.0, 20.0] \\ 2 - \frac{1}{20}x & \text{dla } x \in (20.0, 40.0] \end{cases}$$

$$\mu_{medium-odds}(x) = \begin{cases} \frac{1}{20}x - 1 & \text{dla } x \in [20.0, 40.0) \\ 1 & \text{dla } x \in [40.0, 70.0) \\ 8 - \frac{1}{10}x & \text{dla } x \in (70.0, 80.0] \end{cases}$$

$$\mu_{high-odds}(x) = \begin{cases} \frac{1}{20}x - \frac{7}{2} & \text{dla } x \in [70.0, 90.0) \\ 1 & \text{dla } x \in [90.0, 99.9] \end{cases}$$



Rysunek 13. Funkcja przynależności dla zmiennej L5

Regułą gramatyczną G, która generuje terminy w zbiorach H(L) jest proste wyliczenie możliwych etykiet.

5. Wyniki

Dla podsumowania optymalnego zostały ustalone następujące wagi odpowiednich miar jakości:

— $w_1 = 0.7$
— $w_2 = 0.03$
— $w_3 = 0.03$
— $w_4 = 0.03$
— $w_5 = 0.03$
— $w_6 = 0.03$
— $w_7 = 0.03$
— $w_8 = 0.03$
— $w_9 = 0.03$
— $w_{10} = 0.03$
— $w_{11} = 0.03$

Program umożliwia zmianę wag poprzez GUI.

5.1. Podsumowanie 1

Almost none of runs were with horse age young.
Some of runs were with horse age young.
About half of runs were with horse age young.
Many of runs were with horse age young.
Almost all of runs were with horse age young.
About 40 thousand of runs were with horse age young.
About 70 thousand of runs were with horse age young.

Kwantyfikator	T	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11
Almost none	0.12	0.0	0.1	0.9	0.0	1.0	0.16	0.9	1.0	0.0	0.0	0.0
Some	0.12	0.0	0.1	0.9	0.0	1.0	0.28	0.78	1.0	0.0	0.0	0.0
About half	0.12	0.0	0.1	0.9	0.0	1.0	0.36	0.76	1.0	0.0	0.0	0.0
Many	0.69	0.82	0.1	0.9	0.0	1.0	0.28	0.78	1.0	0.0	0.0	0.0
Almost all	0.16	0.06	0.1	0.9	0.0	1.0	0.16	0.9	1.0	0.0	0.0	0.0
About 40 thousand	0.12	0.0	0.1	0.9	0.0	1.0	0.25	0.81	1.0	0.0	0.0	0.0
About 70 thousand	0.82	1.0	0.1	0.9	0.0	1.0	0.25	0.81	1.0	0.0	0.0	0.0

Tabela 1. Miary jakości

5.2. Podsumowanie 2

Almost none of runs were with horse rating medium rating.
Some of runs were with horse rating medium rating.
About half of runs were with horse rating medium rating.
Many of runs were with horse rating medium rating.
Almost all of runs were with horse rating medium rating.
About 40 thousand of runs were with horse rating medium rating.

About 70 thousand of runs were with horse rating medium rating.

Kwantyfikator	T	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11
Almost none	0.12	0.0	0.05	0.95	0.0	1.0	0.16	0.9	1.0	0.0	0.0	0.0
Some	0.12	0.0	0.05	0.95	0.0	1.0	0.28	0.78	1.0	0.0	0.0	0.0
About half	0.12	0.0	0.05	0.95	0.0	1.0	0.36	0.76	1.0	0.0	0.0	0.0
Many	0.12	0.0	0.05	0.95	0.0	1.0	0.28	0.78	1.0	0.0	0.0	0.0
Almost all	0.43	0.44	0.05	0.95	0.0	1.0	0.16	0.9	1.0	0.0	0.0	0.0
About 40 thousand	0.12	0.0	0.05	0.95	0.0	1.0	0.25	0.81	1.0	0.0	0.0	0.0
About 70 thousand	0.82	1.0	0.05	0.95	0.0	1.0	0.25	0.81	1.0	0.0	0.0	0.0

Tabela 2. Miary jakości

5.3. Podsumowanie 3

Almost none of runs were with win odds high odds.
Some of runs were with win odds high odds.
About half of runs were with win odds high odds.
Many of runs were with win odds high odds.
Almost all of runs were with win odds high odds.
About 40 thousand of runs were with win odds high odds.
About 70 thousand of runs were with win odds high odds.

Kwantyfikator	T	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11
Almost none	0.35	0.32	0.86	0.14	0.0	1.0	0.16	0.9	1.0	0.0	0.0	0.0
Some	0.14	0.03	0.86	0.14	0.0	1.0	0.28	0.78	1.0	0.0	0.0	0.0
About half	0.12	0.0	0.86	0.14	0.0	1.0	0.36	0.76	1.0	0.0	0.0	0.0
Many	0.12	0.0	0.86	0.14	0.0	1.0	0.28	0.78	1.0	0.0	0.0	0.0
Almost all	0.12	0.0	0.86	0.14	0.0	1.0	0.16	0.9	1.0	0.0	0.0	0.0
About 40 thousand	0.12	0.0	0.86	0.14	0.0	1.0	0.25	0.81	1.0	0.0	0.0	0.0
About 70 thousand	0.12	0.0	0.86	0.14	0.0	1.0	0.25	0.81	1.0	0.0	0.0	0.0

Tabela 3. Miary jakości

5.4. Podsumowanie 4

Almost none of runs were with horse age young and finish time medium.
Some of runs were with horse age young and finish time medium.
About half of runs were with horse age young and finish time medium.
Many of runs were with horse age young and finish time medium.
Almost all of runs were with horse age young and finish time medium.
About 40 thousand of runs were with horse age young and finish time medium.
About 70 thousand of runs were with horse age young and finish time medium.

Kwantyfikator	T	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11
Almost none	0.1	0.0	0.3	0.49	0.0	0.5	0.16	0.9	1.0	0.0	0.0	0.0
Some	0.51	0.58	0.3	0.49	0.0	0.5	0.28	0.78	1.0	0.0	0.0	0.0
About half	0.3	0.28	0.3	0.49	0.0	0.5	0.36	0.76	1.0	0.0	0.0	0.0
Many	0.1	0.0	0.3	0.49	0.0	0.5	0.28	0.78	1.0	0.0	0.0	0.0
Almost all	0.1	0.0	0.3	0.49	0.0	0.5	0.16	0.9	1.0	0.0	0.0	0.0
About 40 thousand	0.1	0.0	0.3	0.49	0.0	0.5	0.25	0.81	1.0	0.0	0.0	0.0
About 70 thousand	0.1	0.0	0.3	0.49	0.0	0.5	0.25	0.81	1.0	0.0	0.0	0.0

Tabela 4. Miary jakości

5.5. Podsumowanie 5

Almost none of runs were with horse age young and win odds low odds and lengths behind winner some and horse rating medium rating.
Some of runs were with horse age young and win odds low odds and lengths behind winner some and horse rating medium rating.
About half of runs were with horse age young and win odds low odds and lengths behind winner some and horse rating medium rating.
Many of runs were with horse age young and win odds low odds and lengths behind winner some and horse rating medium rating.
Almost all of runs were with horse age young and win odds low odds and lengths behind winner some and horse rating medium rating.
About 40 thousand of runs were with horse age young and win odds low odds and lengths behind winner some and horse rating medium rating.
About 70 thousand of runs were with horse age young and win odds low odds and lengths behind winner some and horse rating medium rating.

Kwantyfikator	T	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11
Almost none	0.08	0.0	0.25	0.29	0.02	0.13	0.16	0.9	1.0	0.0	0.0	0.0
Some	0.78	1.0	0.25	0.29	0.02	0.13	0.28	0.78	1.0	0.0	0.0	0.0
About half	0.08	0.0	0.25	0.29	0.02	0.13	0.36	0.76	1.0	0.0	0.0	0.0
Many	0.08	0.0	0.25	0.29	0.02	0.13	0.28	0.78	1.0	0.0	0.0	0.0
Almost all	0.08	0.0	0.25	0.29	0.02	0.13	0.16	0.9	1.0	0.0	0.0	0.0
About 40 thousand	0.08	0.0	0.25	0.29	0.02	0.13	0.25	0.81	1.0	0.0	0.0	0.0
About 70 thousand	0.08	0.0	0.25	0.29	0.02	0.13	0.25	0.81	1.0	0.0	0.0	0.0

Tabela 5. Miary jakości

5.6. Podsumowanie 6

Almost none of runs having / being horse rating high rating were with win odds low odds.
Some of runs having / being horse rating high rating were with win odds low odds.
About half of runs having / being horse rating high rating were with win odds low odds.

Many of runs having / being horse rating high rating were with win odds low odds.
Almost all of runs having / being horse rating high rating were with win odds low odds.

Kwantyfikator	T	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11
Almost none	0.21	0.0	0.24	0.75	0.01	1.0	0.16	0.9	1.0	0.94	1.0	1.0
Some	0.21	0.0	0.24	0.75	0.01	1.0	0.28	0.78	1.0	0.94	1.0	1.0
About half	0.21	0.0	0.24	0.75	0.01	1.0	0.36	0.76	1.0	0.94	1.0	1.0
Many	0.91	1.0	0.24	0.75	0.01	1.0	0.28	0.78	1.0	0.94	1.0	1.0
Almost all	0.21	0.0	0.24	0.75	0.01	1.0	0.16	0.9	1.0	0.94	1.0	1.0

Tabela 6. Miary jakości

5.7. Podsumowanie 7

Almost none of runs having / being win odds high odds were with finish time fast.
Some of runs having / being win odds high odds were with finish time fast.
About half of runs having / being win odds high odds were with finish time fast.
Many of runs having / being win odds high odds were with finish time fast.
Almost all of runs having / being win odds high odds were with finish time fast.

Kwantyfikator	T	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11
Almost none	0.21	0.0	0.51	0.59	0.1	1.0	0.16	0.9	1.0	0.86	1.0	1.0
Some	0.21	0.0	0.51	0.59	0.1	1.0	0.28	0.78	1.0	0.86	1.0	1.0
About half	0.92	1.0	0.51	0.59	0.1	1.0	0.36	0.76	1.0	0.86	1.0	1.0
Many	0.21	0.0	0.51	0.59	0.1	1.0	0.28	0.78	1.0	0.86	1.0	1.0
Almost all	0.21	0.0	0.51	0.59	0.1	1.0	0.16	0.9	1.0	0.86	1.0	1.0

Tabela 7. Miary jakości

5.8. Podsumowanie 8

Almost none of runs having / being actual weight light were with declared weight medium.
Some of runs having / being actual weight light were with declared weight medium.
About half of runs having / being actual weight light were with declared weight medium.
Many of runs having / being actual weight light were with declared weight medium.
Almost all of runs having / being actual weight light were with declared

weight medium.

Kwantyfikator	T	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11
Almost none	0.21	0.0	0.24	0.78	0.02	1.0	0.16	0.9	1.0	0.88	1.0	1.0
Some	0.21	0.0	0.24	0.78	0.02	1.0	0.28	0.78	1.0	0.88	1.0	1.0
About half	0.31	0.14	0.24	0.78	0.02	1.0	0.36	0.76	1.0	0.88	1.0	1.0
Many	0.76	0.79	0.24	0.78	0.02	1.0	0.28	0.78	1.0	0.88	1.0	1.0
Almost all	0.21	0.0	0.24	0.78	0.02	1.0	0.16	0.9	1.0	0.88	1.0	1.0

Tabela 8. Miary jakości

5.9. Podsumowanie 9

Almost none of runs having / being lengths behind in section close behind were with lengths behind winner very few.
Some of runs having / being lengths behind in section close behind were with lengths behind winner very few.
About half of runs having / being lengths behind in section close behind were with lengths behind winner very few.
Many of runs having / being lengths behind in section close behind were with lengths behind winner very few.
Almost all of runs having / being lengths behind in section close behind were with lengths behind winner very few.

Kwantyfikator	T	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11
Almost none	0.2	0.0	0.56	0.48	0.05	1.0	0.16	0.9	1.0	0.46	1.0	1.0
Some	0.2	0.0	0.56	0.48	0.05	1.0	0.28	0.78	1.0	0.46	1.0	1.0
About half	0.76	0.8	0.56	0.48	0.05	1.0	0.36	0.76	1.0	0.46	1.0	1.0
Many	0.2	0.0	0.56	0.48	0.05	1.0	0.28	0.78	1.0	0.46	1.0	1.0
Almost all	0.2	0.0	0.56	0.48	0.05	1.0	0.16	0.9	1.0	0.46	1.0	1.0

Tabela 9. Miary jakości

5.10. Podsumowanie 10

Almost none of runs having / being horse age old were with horse rating medium rating and position in section in the middle of the scoreboard.
Some of runs having / being horse age old were with horse rating medium rating and position in section in the middle of the scoreboard.
About half of runs having / being horse age old were with horse rating medium rating and position in section in the middle of the scoreboard.
Many of runs having / being horse age old were with horse rating medium rating and position in section in the middle of the scoreboard.
Almost all of runs having / being horse age old were with horse rating medium rating and position in section in the middle of the scoreboard.

Kwantyfikator	T	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11
Almost none	0.19	0.0	0.33	0.36	0.1	0.5	0.16	0.9	1.0	0.96	1.0	1.0
Some	0.89	1.0	0.33	0.36	0.1	0.5	0.28	0.78	1.0	0.96	1.0	1.0
About half	0.19	0.0	0.33	0.36	0.1	0.5	0.36	0.76	1.0	0.96	1.0	1.0
Many	0.19	0.0	0.33	0.36	0.1	0.5	0.28	0.78	1.0	0.96	1.0	1.0
Almost all	0.19	0.0	0.33	0.36	0.1	0.5	0.16	0.9	1.0	0.96	1.0	1.0

Tabela 10. Miary jakości

5.11. Podsumowanie 11 - wielopodmiotowe w formie 1

Almost none of runs by Mare horses compared to runs by Gelding horses were with horse age old.
Some of runs by Mare horses compared to runs by Gelding horses were with horse age old.
About half of runs by Mare horses compared to runs by Gelding horses were with horse age old.
Many of runs by Mare horses compared to runs by Gelding horses were with horse age old.
Almost all of runs by Mare horses compared to runs by Gelding horses were with horse age old.

Kwantyfikator	T1
Almost none	0.0
Some	0.0
About half	0.72
Many	0.0
Almost all	0.0

Tabela 11. Miary jakości

5.12. Podsumowanie 12 - wielopodmiotowe w formie 1

Almost none of runs by Gelding horses compared to runs by Rig horses were with finish time very fast.
Some of runs by Gelding horses compared to runs by Rig horses were with finish time very fast.
About half of runs by Gelding horses compared to runs by Rig horses were with finish time very fast.
Many of runs by Gelding horses compared to runs by Rig horses were with finish time very fast.
Almost all of runs by Gelding horses compared to runs by Rig horses were with finish time very fast.

Kwantyfikator	T1
Almost none	0.0
Some	0.0
About half	0.37
Many	0.44
Almost all	0.0

Tabela 12. Miary jakości

5.13. Podsumowanie 13 - wielopodmiotowe w formie 2

Almost none of runs by Mare horses compared to runs by Gelding horses having / being horse age young were with finish time slow.
Some of runs by Mare horses compared to runs by Gelding horses having / being horse age young were with finish time slow.
About half of runs by Mare horses compared to runs by Gelding horses having / being horse age young were with finish time slow.
Many of runs by Mare horses compared to runs by Gelding horses having / being horse age young were with finish time slow.
Almost all of runs by Mare horses compared to runs by Gelding horses having / being horse age young were with finish time slow.

Kwantyfikator	T1
Almost none	0.0
Some	0.0
About half	0.0
Many	1.0
Almost all	0.0

Tabela 13. Miary jakości

5.14. Podsumowanie 14 - wielopodmiotowe w formie 2

Almost none of runs by Mare horses compared to runs by Horse horses having / being declared weight heavy were with finish time very slow.
Some of runs by Mare horses compared to runs by Horse horses having / being declared weight heavy were with finish time very slow.
About half of runs by Mare horses compared to runs by Horse horses having / being declared weight heavy were with finish time very slow.
Many of runs by Mare horses compared to runs by Horse horses having / being declared weight heavy were with finish time very slow.
Almost all of runs by Mare horses compared to runs by Horse horses having / being declared weight heavy were with finish time very slow.

Kwantyfikator	T1
Almost none	0.0
Some	0.0
About half	0.0
Many	0.0
Almost all	0.73

Tabela 14. Miary jakości

5.15. Podsumowanie 15 - wielopodmiotowe w formie 3

Almost none of runs by Filly horses having / being horse age young compared to runs by Mare horses were with declared weight medium.
Some of runs by Filly horses having / being horse age young compared to runs by Mare horses were with declared weight medium.
About half of runs by Filly horses having / being horse age young compared to runs by Mare horses were with declared weight medium.
Many of runs by Filly horses having / being horse age young compared to runs by Mare horses were with declared weight medium.
Almost all of runs by Filly horses having / being horse age young compared to runs by Mare horses were with declared weight medium.

Kwantyfikator	T1
Almost none	0.0
Some	0.0
About half	1.0
Many	0.0
Almost all	0.0

Tabela 15. Miary jakości

5.16. Podsumowanie 16 - wielopodmiotowe w formie 3

Almost none of runs by Gelding horses having / being finish time medium compared to runs by Mare horses were with horse rating low rating.
Some of runs by Gelding horses having / being finish time medium compared to runs by Mare horses were with horse rating low rating.
About half of runs by Gelding horses having / being finish time medium compared to runs by Mare horses were with horse rating low rating.
Many of runs by Gelding horses having / being finish time medium compared to runs by Mare horses were with horse rating low rating.
Almost all of runs by Gelding horses having / being finish time medium compared to runs by Mare horses were with horse rating low rating.

Kwantyfikator	T1
Almost none	0.06
Some	0.81
About half	0.0
Many	0.0
Almost all	0.0

Tabela 16. Miary jakości

5.17. Podsumowanie 17 - wielopodmiotowe w formie 4

More runs by Filly horses than runs by Mare horses were with horse age young.

Kwantyfikator	T1
More	0.15

Tabela 17. Miary jakości

5.18. Podsumowanie 18 - wielopodmiotowe w formie 4

More runs by Gelding horses than runs by Mare horses were with lengths behind in section close behind.

Kwantyfikator	T1
More	0.13

Tabela 18. Miary jakości

5.19. Podsumowanie 19 - wielopodmiotowe w formie 4

More runs by Mare horses than runs by Filly horses were with finish time fast.

Kwantyfikator	T1
More	0.17

Tabela 19. Miary jakości

6. Dyskusja

Na podstawie przeprowadzonych eksperymentów możemy zauważyć, że miarą jakości niosącą najwięcej wartości jest stopień prawdziwości podsumowania, zatem jego waga przy obliczaniu podsumowania optymalnego powinna być najwyższa.

Kwantyfikatory bezwzględne mają swoje zastosowanie jedynie, gdy znana jest liczba elementów w bazie danych. Np. gdy kwantyfikatory będą zdefiniowane jako "około 40 tysięcy", "około 70 tysięcy" i "około 100 tysięcy" dla bazy, w której znajduje się tylko 20 tysięcy elementów, to stopień prawdziwości podsumowania z tymi kwantyfikatorami nigdy nie wyniesie więcej niż 0.
Należy również wspomnieć o wadze wiedzy eksperckiej podczas definiowania etykiet oraz funkcji przynależności dla zmiennych lingwistycznych. Bez odpowiedniej wiedzy wygenerowane podsumowania nie będą niosły żadnej znaczącej treści.
Dla podsumowań wielopodmiotowych konieczne jest, aby elementy w bazie były zróżnicowane względem danego atrybutu. W przypadku naszej bazy biegi koni różnego typu nie różniły się od siebie znacząco, dlatego wyniki otrzymane w podsumowaniach od 11 do 19 nie są tak wartościowe jak pozostałe podsumowania. Przyczynił się do tego również fakt, że liczność grup reprezentujących typy koni jest bardzo różna - ok. 95% stanowią biegi koni typu Gelding, podczas gdy pozostałe 5% rozłożone jest na aż 8 innych typów.

7. Wnioski

— Miara T1 jest najważniejszą miarą podczas wyznaczania podsumowania optymalnego.
— Kwantyfikatory względne powinny mieć szeroki zakres lub uwzględniać liczbę elementów w bazie danych.
— Wiedza ekspercka jest niezbędna podczas definiowania etykiet zmiennych lingwistycznych.
— Aby podsumowania wielopodmiotowe były wartościowe, konieczne jest, żeby elementy w bazie danych różniły się względem jakiegoś atrybutu.

Literatura

[1] https://www.kaggle.com/gdaley/hkracing
[2] Niewiadomski, Adam. Methods for the Linguistic Summarization of Data: Applications of Fuzzy Sets and Their Extensions. Akademicka Oficyna Wydawnicza EXIT, Warszawa 2008. ISBN: 978-83-60434-40-6
[3] Niewiadomski, Adam. Zbiory rozmyte typu 2. Zastosowania w reprezentacji informacji. Akademicka Oficyna Wydawnicza EXIT, 2019. ISBN: 978-83-7837-595-1
[4] J. Kacprzyk, R. R. Yager, S. Zadrożny. A Fuzzy Logic Based Approach To Linguistic Summaries of Databes, International Journal of Appliead Mathematics and Computer Sciences, 10:813-834, 2000.