Informatyka, studia dzienne, I st. semestr VI

Komputerowe systemy rozpoznawania **2018/2019**

Prowadzący: dr inż. Marcin Kacprowicz wtorek, 16.15

Data oddania: _____ Ocena: _____

Piotr Traczyk 195733

Bartosz Jurczewski 210209

Zadanie 2: Lingwistyczne podsumowania baz danych*

1. Cel

"Zadanie polega na stworzeniu aplikacji desktopowej przy użyciu dowolnego formatu bazy danych. W ogólności, aplikacja ma charakter systemu doradczego, który generuje pewną ilość podsumowań lingwistycznych dla podanej bazy, a następnie przedstawia użytkownikowi wybrane – najlepsze wg zastosowanych miar jakości – wyniki, czyli podsumowania lingwistyczne. Aplikacja umożliwiać ma automatyczne generowanie podsumowań lingwistycznych służących do tworzenia krótkich wiadomości tekstowych na podstawie dużej relacyjnej bazy danych." [1]

2. Wprowadzenie

Kwestią jaką zajmowaliśmy się w ramach projektu była analiza funkcjonowania lingwistycznych podsumowań baz danych na zbiorach rozmytych. Zbiór rozmyty jest fundamentalnym terminem wykorzystywanym przy naszym projekcie, wobec tego przytoczmy jego definicję:

* Repozytorium github: `https://github.com/jurczewski/KSR`

Definicja 1. Niech $\mathcal{X}$ będzie zbiorem, którego elementy interesują nas w sposób bezpośredni czyli jest „zbiorem zwykłym". Wówczas *zbiorem rozmytym opisanym na* $\mathcal{X}$ nazywamy każdy zbiór A postaci:

$$A = \bigcup_{x \in \mathcal{X}} \{(x, \mu_A(x))\},$$

gdzie $\mu_A(x) : \mathcal{X} \to [0, 1]$ nazywamy *funkcją przynależności do zbioru rozmytego* A.

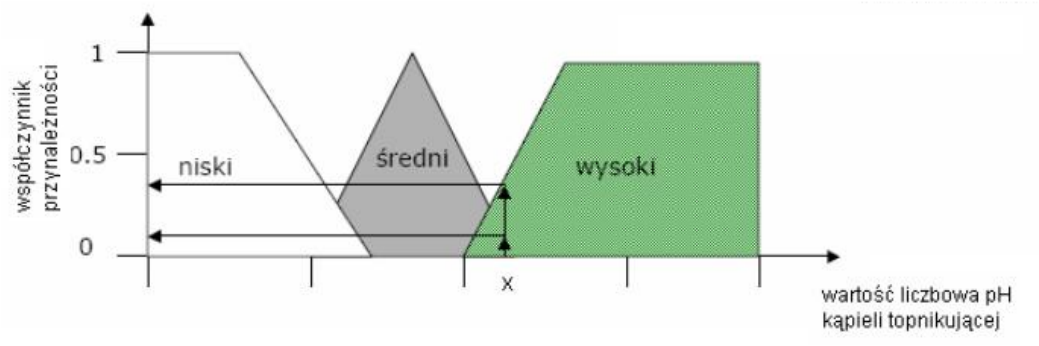
Funkcja przynależności definiuje w jakim stopniu dany element przynależy do zbioru. W zbiorach rozmytych zakres wartości jakie może ona przyjmować jest rozszerzony do przedziału [0,1]. W naszym projekcie skorzystaliśmy z funkcji przynależności *trójkątnej* oraz *trapezoidalnej*. Przytoczmy ich definicje:

Definicja 2 (Zbiór rozmyty o trójkątnej funkcji przynależności)**.** Zbiór rozmyty A opisany na przestrzeni $\mathbb{R}$ jest *liczbą rozmytą trójkątną o parametrach* a, b, c wtedy i tylko wtedy, gdy $a \leq b \leq c$ oraz:

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

Definicja 3 (Zbiór rozmyty o trapezoidalnej funkcji przynależności)**.** Zbiór rozmyty A opisany na przestrzeni $\mathbb{R}$ jest *liczbą rozmytą trapezoidalną o parametrach* a, b, c, d wtedy i tylko wtedy, gdy $a \leq b \leq c \leq d$ oraz:

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



Rysunek 1. Przykład funkcji przynależności - trójkątnej oraz trapezoidalnej [5]

3. Opis implementacji

Program został stworzony przy użyciu .NET Framework w języku C#. Graficzny interfejs powstał przy wykorzystaniu Windows Presentation Foundation. Cały projekt został napisany zgodnie ze wzorcem architektonicznym Model–view–viewmodel (MVVM), w związku z tym powstały trzy warstwy Logic, ViewModel i GUI.

3.1. Logic

Warstwa odpowiedzialna za logikę aplikacji. Klasa reprezentująca krotkę w bazie (*Player.cs*), zaimplementowane zostały: funkcje przynależności trójkątna (*TriangularFunction.cs*) oraz trapezoidalna (*TrapezoidFunction*), zmienna lingwistyczna (*LinguisticVariable.cs*), kwantyfikator (*StaticQuantifiers.cs*), zmienna, która "na sztywno" określa nasz kwantyfikator (np. słaby, przeciętny, dobry) w zależności od podanych danych (*StaticVariables.cs*), sumaryzator "i" (*And.cs*), a także sumaryzator "lub" (*Or.cs*). W tej warstwie znajduje się także klasa (*Measures.cs*), gdzie zawarliśmy wszystkie 11 miar jakości podsumowań.

3.2. ViewModel

Klasa MainViewModel przyjmuje dane wejściowe od użytkownika i reaguje na jego poczynania wywołując wybrane akcje z logiki programu oraz odpowiada za odświeżanie widoków w interfejsie graficznym.

3.3. GUI

Przejrzysty interfejs użytkownika który ma kontrole nad kwalifikatorem, sumaryzatorem (również złożonym OR/AND), generowaniem podsumowań oraz zapisem do pliku.

4. Miary jakości

Do określenia jakości podsumowania lingwistycznego zaimplementowaliśmy następujące miary jakości, których wzory zostały zaczerpnięte z [4].

- T1 (Degree of truth) - Suma przynależności wszystkich rozważanych krotek do podsumowania lingwistycznego.
- T2 (Degree of imprecision) - Określa stopień precyzyjności sumaryzatora (miarę jakości podsumowania określaną).
- T3 (Degree of covering) - Reprezentuje stopień w jakim nośnik sumaryzatora pokrywa się z nośnikiem kwalifikatora.
- T4 (Degree of appropriateness) - Definiuje jak dużo krotek przynależy do sumaryzatora, czyli czy dane podsumowanie jest właściwe dla zestawu danych.
- T5 (Length of summary) - Określa jakość podsumowania na podstawie złożoności sumaryzatora, więc im więcej składowych sumaryzatora złożonego, tym mniejsza wartość owej miary.
- T6 (Degree of quantifier imprecision) - Pokazuje w jakim stopniu precyzyjny jest kwantyfikator. Im mniejszy nośnik zbioru rozmytego tym wyższa jest jego precyzja.
- T7 (Degree of quantifier cardinality) - Opisuje stopień precyzji kwantyfikatora, im mniejsza kardynalność kwantyfikatora tym jest on bardziej precyzyjny.
- T8 (Degree of summarizer cardinality) - Opisuje stopień precyzji sumaryzatora, im mniejsza kardynalność kwantyfikatora tym jest on bardziej precyzyjny.
- T9 (Degree of qualifier imprecision) - Określa w jakim stopniu precyzyjny jest kwalifikator. Im szerszy nośnik zbioru rozmytego tym niższa jest jego precyzja, gdyż bierze pod uwagę większy zakres wartości.
- T10 (Degree of qualifier cardinality) - Opisuje stopień precyzji kwalifikatora, im większa jest kardynalność kwalifikator tym jest on mniej precyzyjny.
- T11 (Length of qualifier) - Wyznacza jakość podsumowania na podstawie złożoności kwalifikatora. Im bardziej złożony kwalifikator tym jakość podsumowania gorsza.

5. Materiały i metody

5.1. Baza danych

Do realizacji zadania wybraliśmy bazę danych *Fifa 19* [2]. Wybrana baza składa się z 18128 krotek znajdujących się w tabeli z 10 kolumnami różnych typów. Przechowuje ona statystyki piłkarzy z gry Fifa 2019.
Do tworzenia podsumowań korzystamy z kolumn:

- Age - wiek [Lata]
- Overall - ocena całkowita zawodnika
- Value - wartość zawodnika [€]
- Wage - zarobki [€]
- Height - wzrost [Stopy i Cale]
- Weight - waga [Funty]
- FKAccuracy - (free kicks accuracy) - efektywność rzutów wolnych (procent bramek z rzutów wolnych)

- SprintSpeed - szybkość w sprincie
- Stamina - wytrzymałość
- Strength - siła

	ID	Age	Overall	Value	Wage	Height	Weight	FKAccuracy	SprintSpeed	Stamina	Strength
1	158023	31.00	94.00	110.50	565.00	171.00	159.00	94.00	86.00	72.00	59.00
2	20801	33.00	94.00	77.00	405.00	186.00	183.00	76.00	91.00	88.00	79.00
3	190871	26.00	92.00	118.50	290.00	177.00	150.00	87.00	90.00	81.00	49.00
4	193080	27.00	91.00	72.00	260.00	192.00	168.00	19.00	58.00	43.00	64.00
5	192985	27.00	91.00	102.00	355.00	183.00	154.00	83.00	76.00	90.00	75.00
6	183277	27.00	91.00	93.00	340.00	174.00	163.00	79.00	88.00	83.00	66.00
7	177003	32.00	91.00	67.00	420.00	174.00	146.00	78.00	72.00	89.00	58.00
8	176580	31.00	91.00	80.00	455.00	180.00	190.00	84.00	75.00	90.00	83.00
9	155862	32.00	91.00	51.00	380.00	180.00	181.00	72.00	75.00	84.00	83.00
10	200389	25.00	90.00	68.00	94.00	186.00	192.00	14.00	60.00	41.00	78.00
11	188545	29.00	90.00	77.00	205.00	180.00	176.00	86.00	78.00	78.00	84.00
12	182521	28.00	90.00	76.50	355.00	180.00	168.00	84.00	62.00	75.00	73.00
13	182493	32.00	90.00	44.00	125.00	186.00	172.00	51.00	68.00	66.00	88.00
14	168542	32.00	90.00	60.00	285.00	174.00	148.00	77.00	64.00	78.00	52.00
15	215914	27.00	89.00	63.00	225.00	168.00	159.00	49.00	78.00	96.00	76.00
16	211110	24.00	89.00	89.00	205.00	180.00	165.00	88.00	83.00	80.00	65.00
17	202126	24.00	89.00	83.50	205.00	186.00	196.00	68.00	72.00	89.00	84.00
18	194765	27.00	89.00	78.00	145.00	177.00	161.00	78.00	85.00	83.00	62.00
19	192448	26.00	89.00	58.00	240.00	186.00	187.00	12.00	50.00	35.00	79.00
20	192119	26.00	89.00	53.50	240.00	198.00	212.00	20.00	52.00	38.00	70.00

Rysunek 2. Fragment widoku tabeli

5.2. Sumaryzatory i kwalifikatory

Poniżej zaprezentowaliśmy poszczególne sumaryzatory oraz kwalifikatory wykorzystane w naszym programie.

Etykieta	**a**	**b**	**c**	**d**
Young	15	16	17	20
Adult	19	23	27	31
Old	30	35	45	51

Tabela 1. Przyporządkowane parametry funkcji trapezoidalnej dla wieku.

Etykieta	**a**	**b**	**c**	**d**
Low Overall	46	50	55	60
Medium Overall	59	64	70	75
Good Overall	74	80	85	94

Tabela 2. Przyporządkowane parametry funkcji trapezoidalnej dla wyniku ogólnego.

Etykieta	a	b	c	
Low Value	0	30	60	100
Medium Value	99	200	400	600
High Value	599	750	850	975

Tabela 3. Przyporządkowane parametry funkcji trapezoidalnej dla wartości piłkarza.

Etykieta	a	b	c	
Low Wage	0	2	4	6
Medium Wage	5	7	9	12
High Wage	11	50	90	110
Very High Wage	109	250	400	565

Tabela 4. Przyporządkowane parametry funkcji trapezoidalnej dla zarobków.

Etykieta	a	b	c	
Short	153	160	167	172
Medium HeightTall	171	176	181	185
Tall	184	190	200	207

Tabela 5. Przyporządkowane parametry funkcji trapezoidalnej dla wzrostu.

Etykieta	a	b	c	
Light	110	110	150	155
Regular	150	160	190	200
Heavy	195	210	240	255

Tabela 6. Przyporządkowane parametry funkcji trapezoidalnej dla wagi.

Etykieta	a	b	c	
Bad Free Kick	3	15	30	40
Decent Free Kick	39	46	58	70
Good Free Kick	69	75	87	94

Tabela 7. Przyporządkowane parametry funkcji trapezoidalnej dla FKAccuracy.

Etykieta	a	b	c	
Bad Sprint	12	20	30	40
Decent Sprint	39	47	58	68
Good Sprint	67	75	85	96

Tabela 8. Przyporządkowane parametry funkcji trapezoidalnej dla szybkości sprintu.

Etykieta	a	b	c	
Bad Stamina	12	19	31	40
Decent Stamina	39	45	57	68
Good Stamina	67	75	87	96

Tabela 9. Przyporządkowane parametry funkcji trapezoidalnej dla wytrzymałości.

Etykieta	a	b	c	
Bad Strength	17	25	36	45
Decent Strength	54	65	74	80
Good Strength	79	85	90	97

Tabela 10. Przyporządkowane parametry funkcji trapezoidalnej dla siły.

5.3. Kwantyfikatory

Etykieta	Funkcja przynależności	a	b	c	d
No	Trójkątna	0	0	0.16	-
Around 20%	Trójkątna	0.14	0.2	0.26	-
Around one third	Trójkątna	0.25	0.33	0.41	-
Less than a third	Trapezoidalna	0	0	0.3	0.36
Around 40%	Trójkątna	0.34	0.4	0.46	-
Around half	Trójkątna	0.4	0.5	0.6	-
Around three quaters	Trójkątna	0.5	0.6	0.7	-
Majority	Trójkątna	0.75	0.8	0.9	-
All	Trapezoidalna	0.85	0.9	1	1

Tabela 11. Przyporządkowane parametry dla kwantyfikatora względnego.

Etykieta	Funkcja przynależności	a	b	c	d
Less than 1000	Trapezoidalna	0	0	9990	1000
Around 1500	Trójkątna	1400	1500	1600	-
Around 3000	Trójkątna	2900	3000	3100	-
More than 5000	Trapezoidalna	5000	5010	5500	-

Tabela 12. Przyporządkowane parametry dla kwantyfikatora absolutnego.

6. Badania

Nasze badania postanowiliśmy podzielić na trzy części:

1. Sprawdzenie jakie wartości przybiorą miary podsumowań dla różnych kwantyfikatorów.
2. Porównanie podsumowań bez i z kwalifikatorem.
3. Porównanie podsumowań z jednym sumaryzatorem oraz połączanie ze spójnikami OR i LUB.

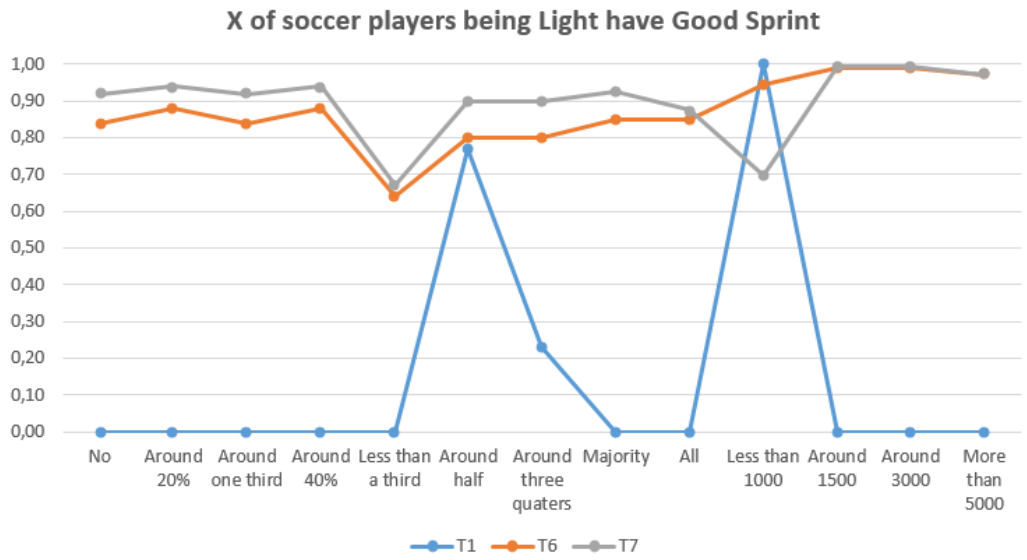
6.1. Pierwszy eksperyment

W tym badaniu wygenerowaliśmy 3 komunikaty, dotyczące odpowiednio zależności pomiędzy:

- 'BodyweightKg' a 'Sprint Speed'
- 'Height' a 'Wage'

Miary $T_2 - T_5$ oraz $T_8 - T10$ były stałe dla różnych kwalifikatorów dlatego zostały przez nas pominięte w tabeli oraz na wykresie, ponieważ kwantyfikator nie miał wpływu na ich wartość. Zostały jednak przedstawione pod tabelą porównującą wartości miar T_1, T_6 i T_7.

6.1.1. Eksperyment 1.1



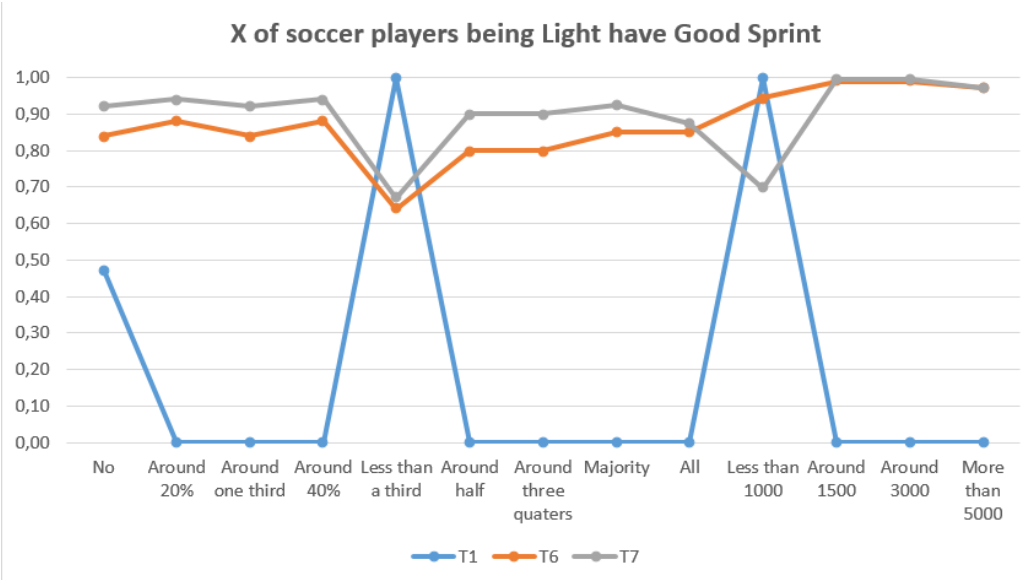
Rysunek 3. Wykres przedstawiający wyniki Eksperymentu 1.1

Kwantyfikator	T_1	T_6	T_7
No	0,00	0,84	0,92
Around 20%	0,00	0,88	0,94
Around one third	0,00	0,84	0,92
Around 40%	0,00	0,88	0,94
Less than a third	0,00	0,64	0,67
Around half	0,77	0,80	0,90
Around three quaters	0,23	0,80	0,90
Majority	0,00	0,85	0,93
All	0,00	0,85	0,88
Less than 1000	1,00	0,94	0,70
Around 1500	0,00	0,99	0,99
Around 3000	0,00	0,99	0,99
More than 5000	0,00	0,97	0,97

Tabela 13. Tabela przedstawiający wyniki Eksperymentu 1.1

Uzyskane miary, które były jednakowe dla każdego kwantyfikatora: $T_2 = 0.508$, $T_3 = 0.669$, $T_4 = 0.669$, $T_5 = 1$, $T_8 = 0.999$, $T_9 = 0.728$, $T_{10} = 0.998$, $T_{11} = 1$.

6.1.2. Eksperyment 1.2



Rysunek 4. Wykres przedstawiający wyniki Eksperymentu 1.2

Kwantyfikator	T_1	T_6	T_7
No	0,47	0,84	0,92
Around 20%	0,00	0,88	0,94
Around one third	0,00	0,84	0,92
Around 40%	0,00	0,88	0,94
Less than a third	1,00	0,64	0,67
Around half	0,00	0,80	0,90
Around three quaters	0,00	0,80	0,90
Majority	0,00	0,85	0,93
All	0,00	0,85	0,88
Less than 1000	1,00	0,94	0,70
Around 1500	0,00	0,99	0,99
Around 3000	0,00	0,99	0,99
More than 5000	0,00	0,97	0,97

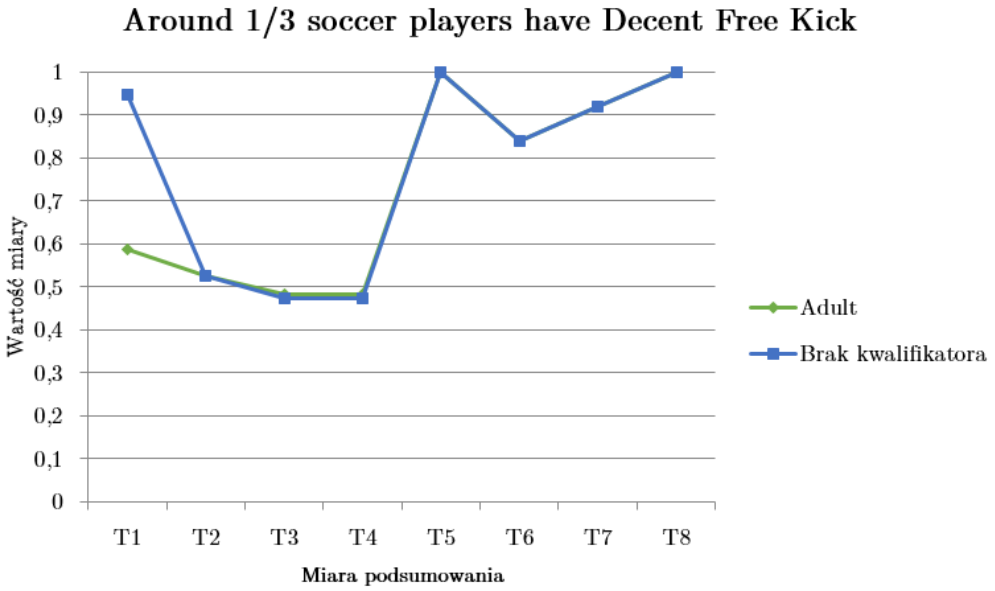
Tabela 14. Tabela przedstawiający wyniki Eksperymentu 1.1

Uzyskane miary, które były jednakowe dla każdego kwantyfikatora: $T_2 = 0.804$, $T_3 = 0.189$, $T_4 = 0.189$, $T_5 = 1$, $T_8 = 0.996$, $T_9 = 0.922$, $T_{10} = 0.999$, $T_{11} = 1$.

6.2. Drugi eksperyment

W tym badaniu wygenerowaliśmy podsumowanie dotyczące FK Accuracy. Sprawdziliśmy jak obecność kwalifikatora (Age: Adult) wpłynęła na wartości miar. Miary $T_9 - T_{11}$ zależały od kwalifikatora - nie dało się ich obliczyć bez niego, dlatego nie były porównywane w tym badaniu.

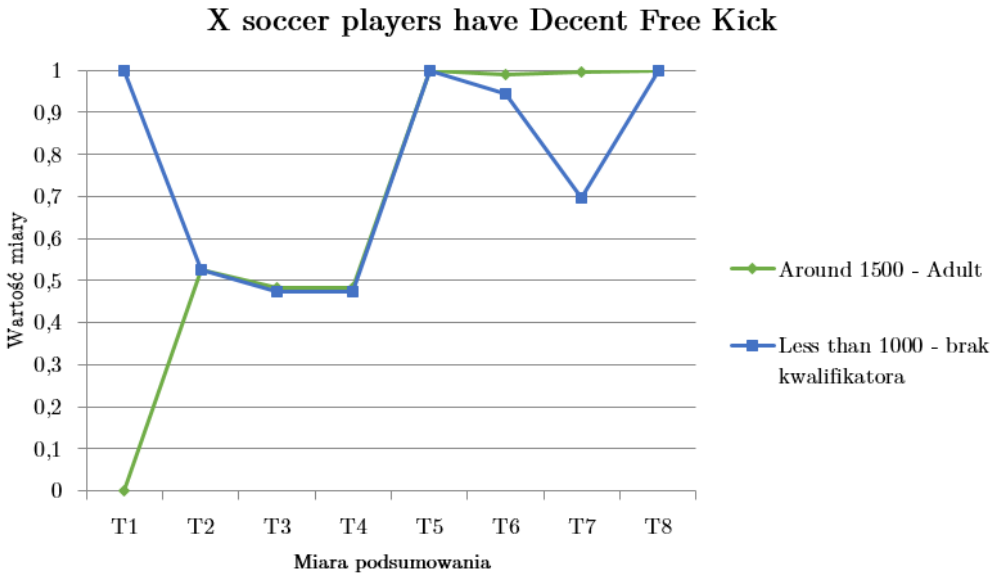
6.2.1. Eksperyment 2.1



Rysunek 5. Wykres przedstawiający wyniki Eksperymentu 2.1 - kwantyfikator względny

Miara	Słaby balans	Brak kwalifikatora
T_1	0,587	0,948
T_2	0,527	0,527
T_3	0,484	0,473
T_4	0,484	0,473
T_5	1,000	1,000
T_6	0,840	0,840
T_7	0,920	0,920
T_8	0,999	0,999

Tabela 15. Tabela przedstawiająca wyniki Eksperymentu 2.1 - kwantyfikator względny (około 2/3 piłkarzy)



Rysunek 6. Wykres przedstawiający wyniki Eksperymentu 2.1 - kwantyfikator bezwzględny

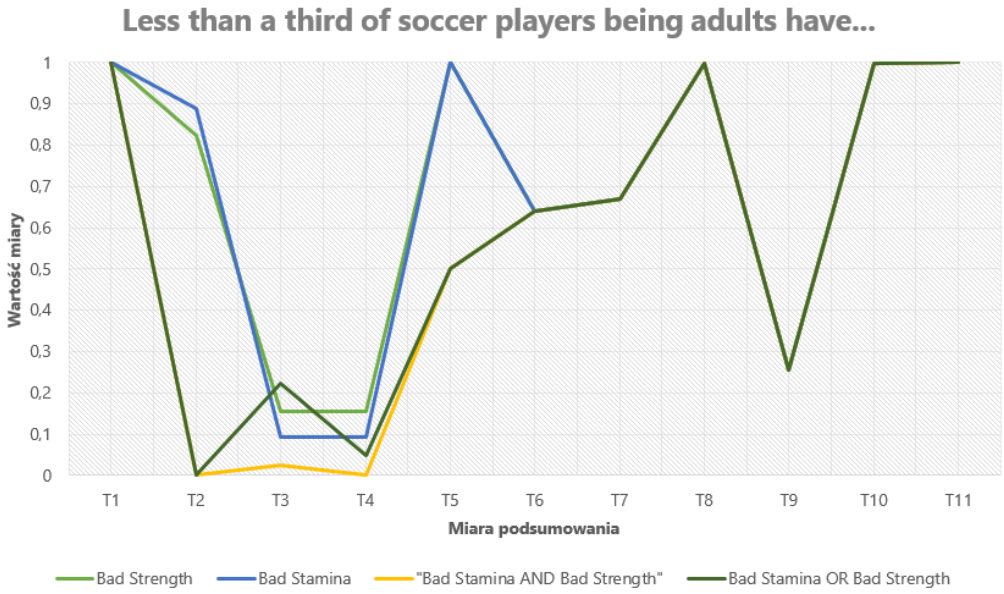
	Słaby balans	**Brak kwalifikatora**
Miara	**Około 500**	**Więcej niż 1000**
T_1	0,000	1,000
T_2	0,527	0,527
T_3	0,484	0,473
T_4	0,484	0,473
T_5	1,000	1,000
T_6	0,989	0,945
T_7	0,994	0,697
T_8	0,999	0,999

Tabela 16. Tabela przedstawiająca wyniki Eksperymentu 2.1 - kwantyfikator bezwzględny

6.3. Trzeci eksperyment

W tym badaniu wygenerowaliśmy podsumowania dotyczące dorosłych piłkarzy, którzy mają:

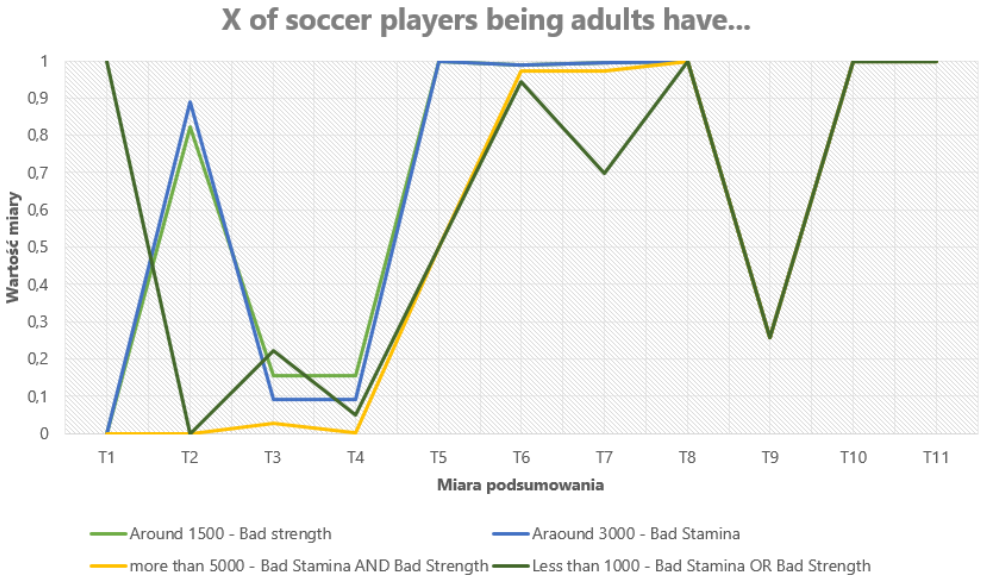
- Bad Strength
- Bad Stamina
- Bad Strength AND Bad Stamina
- Bad Strength OR Bad Stamina



Rysunek 7. Wykres przedstawiający wyniki Eksperymentu 3 - kwantyfikator względny

Miara	Bad Strength	Bad Stamina	ORAZ	LUB
T_1	1,00	1,00	1,00	1,00
T_2	0,82	0,89	0,00	0,00
T_3	0,16	0,09	0,03	0,22
T_4	0,16	0,09	0,00	0,05
T_5	1,00	1,00	0,50	0,50
T_6	0,67	0,67	0,67	0,67
T_7	0,67	0,67	0,67	0,67
T_8	1,00	1,00	1,00	1,00
T_9	0,26	0,26	0,26	0,26
T_{10}	1,00	1,00	1,00	1,00
T_{11}	1,00	1,00	1,00	1,00

Tabela 17. Tabela przedstawiająca wyniki Eksperymentu 3 - kwantyfikator względny (mniej niż 1/3 dorosłych piłkarzy)



Rysunek 8. Wykres przedstawiający wyniki Eksperymentu 3 - kwantyfikator względny

Miara	**Bad Strength**	**Bad Stamina**	**ORAZ**	**LUB**
T_1	0,00	0,00	0,00	1,00
T_2	0,82	0,89	0,00	0,00
T_3	0,16	0,09	0,03	0,22
T_4	0,16	0,09	0,00	0,05
T_5	1,00	1,00	0,50	0,50
T_6	0,99	0,99	0,97	0,70
T_7	0,99	0,99	0,97	0,70
T_8	1,00	1,00	1,00	1,00
T_9	0,26	0,26	0,26	0,26
T_{10}	1,00	1,00	1,00	1,00
T_{11}	1,00	1,00	1,00	1,00

Tabela 18. Tabela przedstawiająca wyniki Eksperymentu 3 - kwantyfikator względny (mniej niż 1/3 dorosłych piłkarzy)

7. Dyskusja

7.1. Wpływ kwantyfikatora na miary

7.2. Wpływ kwalifikatora na miary

7.3. Wpływ sumaryzatora na miary

7.4. Inne spostrzeżenia

8. Wnioski

- Podsumowania lingwistyczne są bardzo dobrą sprawa

Literatura

[1] A. Niewiadomski, *Komputerowe Systemy Rozpoznawania: Materiały, przykłady i ćwiczenia do przedmiotu.* 21 września 2009. `http://ics.p.lodz.pl/~aniewiadomski/ksr/ksr-projekt2.pdf`

[2] Baza danych - "FIFA 19 complete player dataset": `https://www.kaggle.com/karangadiya/fifa19`

[3] L. A. Zadeh, *A fuzzy-set-theoretical interpretation of linguistic hedges.* Journal of Cybernetics 1972; 2: 4–34.

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

[5] `http://zsi.tech.us.edu.pl/~nowak/si/w2_2013.pdf`