Informatyka stosowana, studia dzienne, II st. semestr II

Analiza danych złożonych **2020/2021**

Prowadzący: dr hab. inż. Agnieszka Duraj poniedziałek, 11:45

Etap 3: Badanie podobieństwa oraz przynależności zbiorów tekstowych

Autorzy:

Paweł Galewicz 234053

Karol Podlewski 234106

Spis treści

1. Cel

Zadanie polegało na analizie zbioru danych pod kątem podobieństwa atrybutów tekstowych oraz przynależności atrybutów numerycznych do określonych etykiet.

2. Implementacja

Program został stworzony w języku Python w wersji 3.8.6. Wyniki zostały zapisane w skoroszycie Excel za pomocą biblioteki *openpyxl*. Do wygenerowania wykresów przynależności wykorzystano bibliotekę *matplotlib*. Podobieństwo atrybutów tekstowych sprawdzano z wykorzystaniem wzoru Niewiadomskiego. Dla atrybutów numerycznych określono etykiety na podstawie charakterystyki zbioru i nadano im trapezowe funkcje przynależności.

3. Opis teoretyczny

W sprawozdaniu wykorzystano następujące zagadnienia.

3.1. Zbiory rozmyte

Niech X będzie przestrzenią o skończonej liczbie elementów, wówczas zbiorem rozmytym A określonym na przestrzeni X nazywamy zbiór par w postaci

$$A = \{< x, \mu_A(x) >: x \in X\}, x \in X$$

gdzie $\mu_A(x)$ jest funkcją przynależności wyrażoną jako

$$\mu_A(x) : X \to [0, 1]$$

i określającą stopień przynależności elementu x do zbioru A.

Rozszerzeniem zbioru rozmytego jest przedziałowy zbiór rozmyty, dla którego definiowane są dwie funkcje przynależności - dolna i górna. Oznaczają one odpowiednio minimalny i maksymalny stopień przynależności elementu.

$$A = \{< x, \underline{\mu_A(x)}, \overline{\mu_A(x)} >: x \in X\}, x \in X$$

gdzie

$$\underline{\mu_A(x)}, \overline{\mu_A(x)} : X \to [0; 1]$$

oraz

$$0 \leqslant \underline{\mu_A(x)} \leqslant \overline{\mu_A(x)} \leqslant 1$$

3.2. Podobieństwo tekstu

Podobieństwo słów s_1 i s_2 zdefiniować można następującym wzorem:

$$\mu_{RS}(s_1, s_2) = \frac{2}{N^2 + N} \sum_{i=1}^{N(s_1)} \sum_{j=1}^{N(s_1)-i+1} h(i, j)$$

w którym N definiowane jest jako maksymalna długość słów s_1 i s_2, zaś $h(i, j)$ przyjmuje wartość 1 gdy podciąg i-elementowy liter występujący w słowie s_1 i rozpoczynający się od j-tego miejsca w słowie s_1 występuje co najmniej raz w słowie s_2, a 0, jeżeli podciąg i-elementowy liter występujący w słowie s_1 i rozpoczynający się od j-tego miejsca w słowie s_1 nie występuje w słowie s_2.

Korzystając z powyższego wzoru zdefiniować możemy funkcję podobieństwa zdań z_1 i z_2, wyrażoną następującym wzorem:

$$\mu_{RZ}(z_1, z_2) = \frac{1}{N} \sum_{i=1}^{N(z_1)} max_{j \in \{1,2,\ldots,N\}} g(s_i, s_j)$$

gdzie N jest maksymalną liczbą słów z_1 i z_2, a $g(s_i, s_j)$ jest funkcją podobieństwa słów s_i i s_2.

3.3. Współczynnik R

Współczynnik r atrybutów R, S dla których etykiety zdefiniowane są odpowiednio a_1, a_2, a_3 i b_1, b_2, b_3 wyrażony jest wzorem:

$$r = \frac{\sum_{i=1}^{n} [\mu_R(a_i) * \mu_S(b_i)]}{\sum_{i=1}^{n} \mu_R(a_i)}$$

4. Zbiór danych

Do zadania wykorzystano zbiór wpisów na portalu Tweeter - dalej zwanych *tweetami* - dotyczących zmian klimatycznych. Zbiór pobrano z platformy Kaggle [1]. Zbiór zawiera ok. 400 rekordów zawierających treść wpisu oraz metadane dotyczące jego oraz jego autora. Do analizy wybrano następujące atrybuty:

— twitter_name – nazwa użytkownika z portalu Tweeter, autora tweeta
— text – tekst tweeta
— followers – liczba osób obserwujących autora tweeta

— likes – liczba polubień tweeta
— polarity – odbiór tweeta przez innych użytkowników prezentowana liczbą z przedziału [-1, 1], gdzie -1 oznacza negatywny odbiór, zaś 1 oznacza pozytywny odbiór

Na potrzeby wyznaczenia etykiet oraz wzorów przynależności dla liczbowych atrybutów: followers, likes, polarity wyznaczono statystyki zaprezentowane Tabeli 1.

Tabela 1. Statystyki wybranych atrybutów numerycznych

	followers	likes	polarity
mean	11472.44	6.74	0.04
std	95007.77	46.87	0.14
min	0	0	-0.5
25%	132	0	0
50%	1069	0	0
75%	4191.25	2	0.12
max	1720089	88.0	0.6

5. Badania

W ramach analizy przeprowadzono następujące zadania.

5.1. Porównanie tekstu tweetów

Do analizy wykorzystano atrybut text. W ramach badania porównane zostały wszystkie tweety ze sobą i na podstawie wyników stworzono macierz, w której w kolumnami i rzędami są tweety, a na przecięciu wpisany jest wyliczony współczynnik podobieństwa tweetów. Wynikowa tabela znajduje się w arkuszu *Tweets similarity*, a jej poglądową część prezentuje Tabela 2.

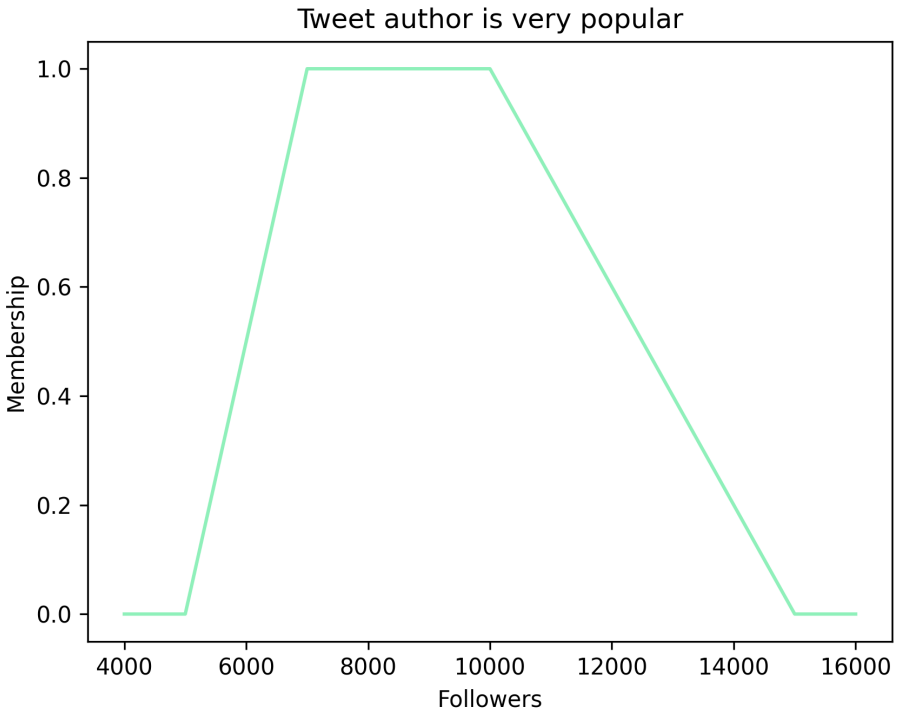
Tabela 2. Porównanie tekstu tweetów

Similarity	Tweet 1	Tweet 2	Tweet 3	Tweet 4	Tweet 5		Tweet 396
Tweet 1	1	0,342888	0,299971	0,251233	0,226598		0,381816
Tweet 2	0,209888	1	0,312124	0,225024	0,189221		0,311165
Tweet 3	0,214364	0,312775	1	0,402763	0,346178	...	0,326685
Tweet 4	0,113415	0,130878	0,243279	1	0,468254		0,157893
Tweet 5	0,12073	0,122297	0,212279	0,490781	1		0,161802
			...				
Tweet 396	0,277615	0,348182	0,346694	0,265106	0,262948		1

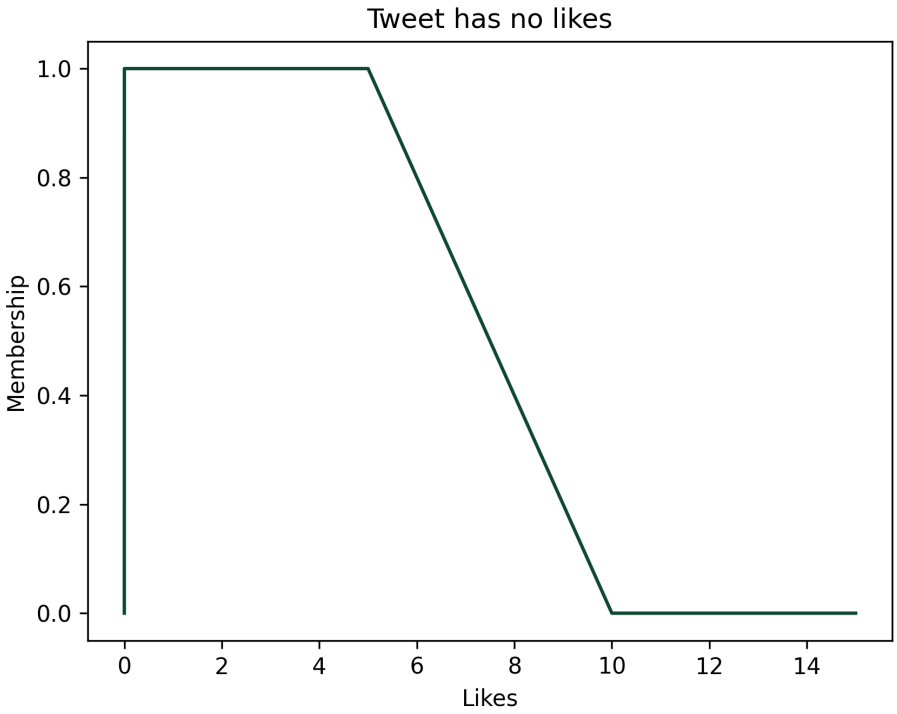
5.2. Przynależność z wykorzystaniem zbiorów rozmytych

Do badań przynależności wykorzystano atrybuty followers, likes, polarity. Każdemu z atrybutów nadano etykietę, do którego stworzono funkcję przynależności w następujący sposób:

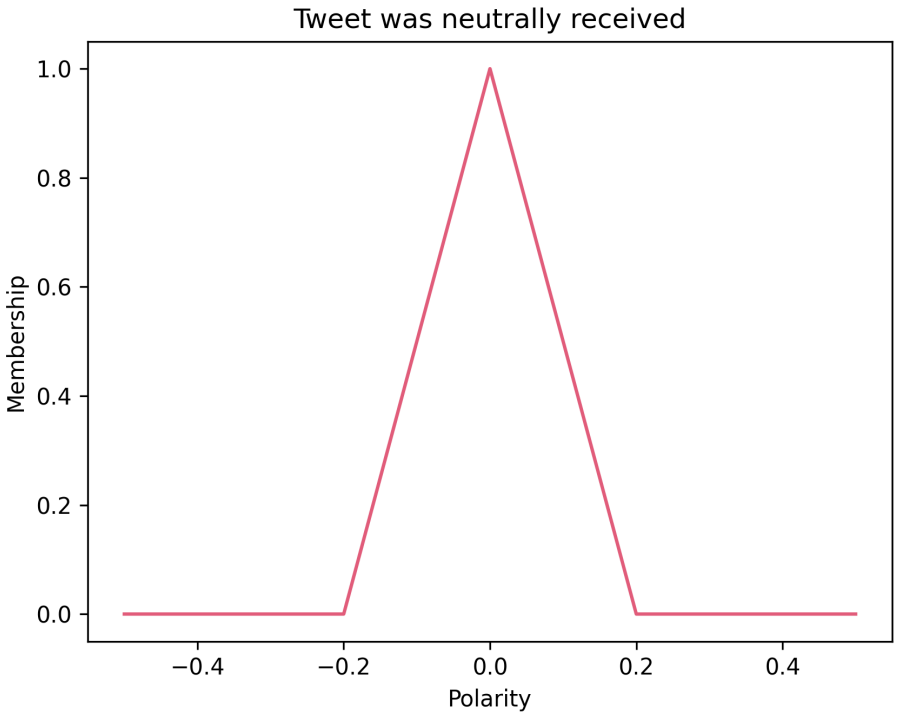
— Dla atrybutu followers etykietę "Tweet author is very popular", której funkcję przynależności prezentuje Rysunek 1
— Dla atrybutu likes etykietę "Tweet has no likes", której funkcję przynależności prezentuje Rysunek 2
— Dla atrybutu polarity etykietę "Tweet was neutrally received", której itemize przynależności prezentuje Rysunek 3



Rysunek 1. Funkcja przynależności dla atrybutu followers



Rysunek 2. Funkcja przynależności dla atrybutu likes



Rysunek 3. Funkcja przynależności dla atrybutu polarity

Dla każdego tweeta wyliczono wartości przynależności funkcji i zapisano w arkuszu *Fuzzy membership*, którego poglądową część zaprezentowano w Tabeli 3

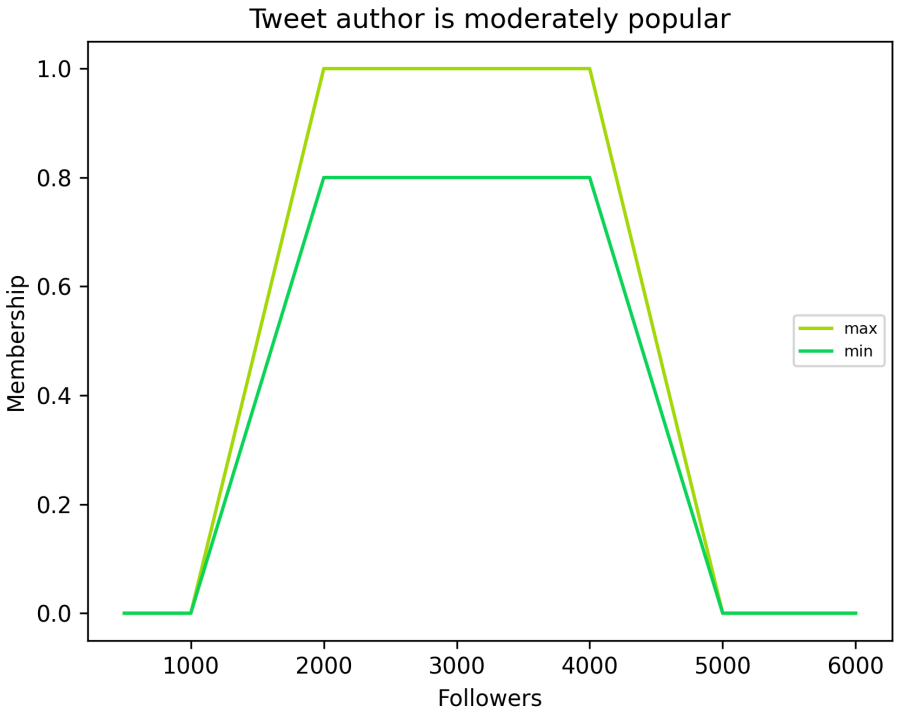
Tabela 3. Wartości funkcji przynależności tweetów

Author	Tweet	Tweet has no likes	Tweet was neutrally recieved	Tweet author is very popular
ECOWARRIORSS	{Treść}	0	0,73	0
ElsevierEnergy	{Treść}	0	0	0,80
siwarr5	{Treść}	1	0	0
EDITORatWORK	{Treść}	0	1	0
EDITORatWORK	{Treść}	1	1	0
mapsofworld	{Treść}	0	0,39	0,15
EnvirHealthNews	{Treść}	1	0	0,95
			...	
TheDailyClimate	{Treść}	0	0,44	0

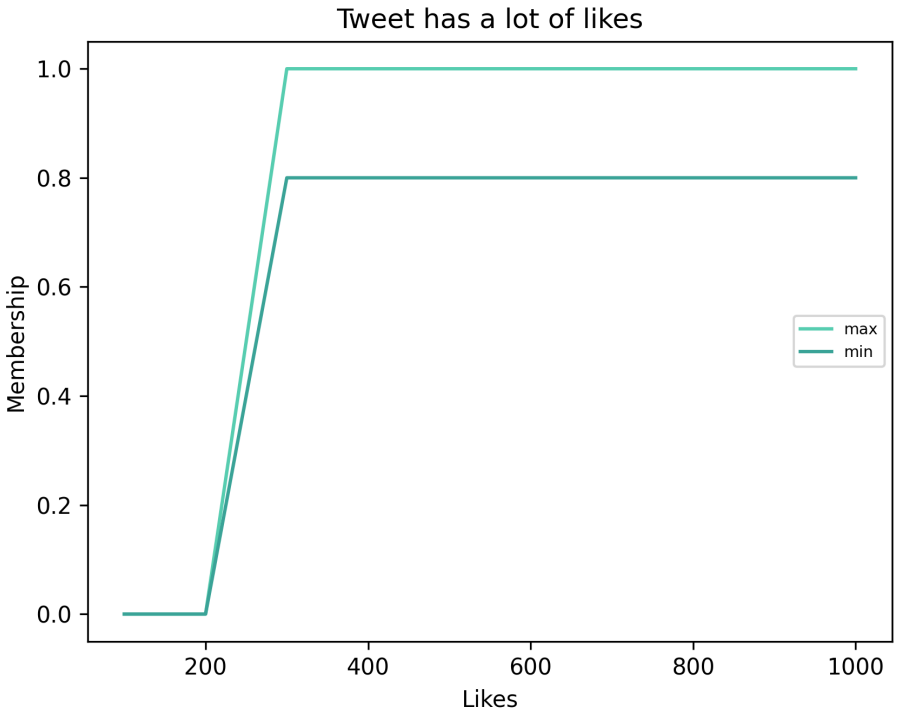
5.3. Przynależność z wykorzystaniem przedziałowych zbiorów rozmytych

Dla tych samych atrybutów nadano inne etykiety, dla których tym razem stworzono funkcje przynależności z wykorzystaniem przedziałowych zbiorów rozmytych. Etykiety prezentują się następująco:

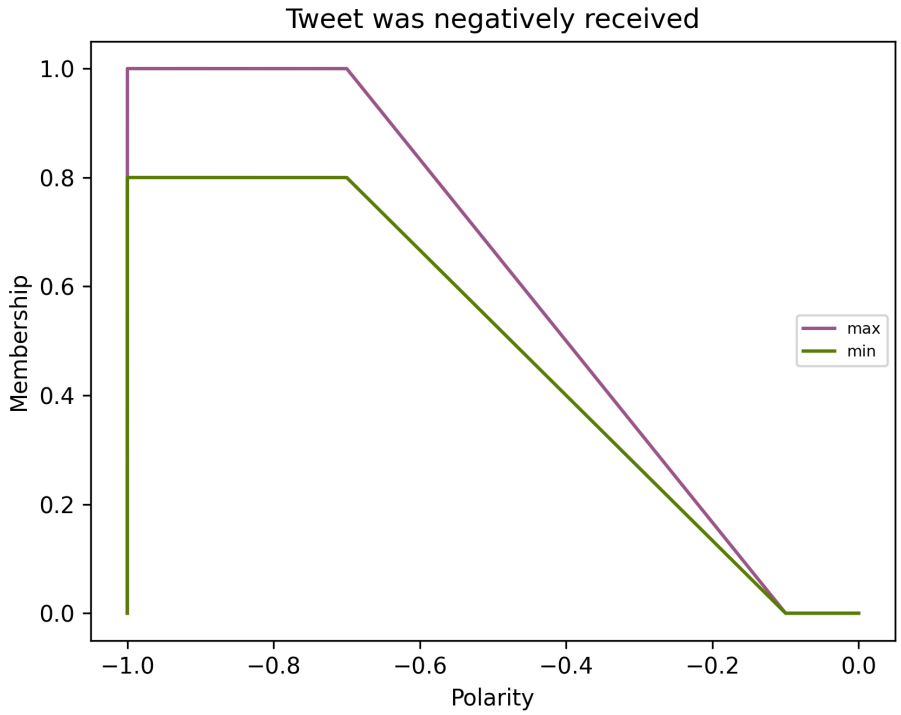
— Dla atrybutu followers etykietę "Tweet author is moderately popular", której funkcję przynależności pokazano na Rysuneku 4
— Dla atrybutu likes etykietę "Tweet has a lot of likes", której funkcję przynależności pokazano na Rysuneku 5
— Dla atrybutu polarity etykietę "Tweet was negatively received", której funkcję przynależności pokazano na Rysuneku 6



Rysunek 4. Przedziałowa funkcja przynależności dla atrybutu followers



Rysunek 5. Przedziałowa funkcja przynależności dla atrybutu likes



Rysunek 6. Przedziałowa funkcja przynależności dla atrybutu polarity

Ponownie dla każdego tweeta wyliczono wartości przedziałowych funkcji przynależności i zapisano w arkuszu *Fuzzy interval membership*. Poglądowa część zaprezentowana jest w Tabeli 4

Tabela 4. Wartości przedziałowych funkcji przynależności tweetów

Author	Tweet	Tweet has a lot of likes (min)	Tweet has a lot of likes (max)	Tweet was negatively received (min)	Tweet was negatively received (max)	Tweet author is moderately popular (min)	Tweet author is moderately popular (max)
ECOWARRIORSS	{Treść}	0	0	0	0	0	0
ElsevierEnergy	{Treść}	0	0	0	0	0	0
siwarr5	{Treść}	0	0	0	0	0	0
EDITORatWORK	{Treść}	0	0	0	0	0,65	0,81
EDITORatWORK	{Treść}	0	0	0	0	0,65	0,81
mapsofworld	{Treść}	0	0	0,03	0,04	0	0
EnvirHealthNews	{Treść}	0	0	0	0	0	0
				...			
TheDailyClimate	{Treść}	0	0	0	0	0	0

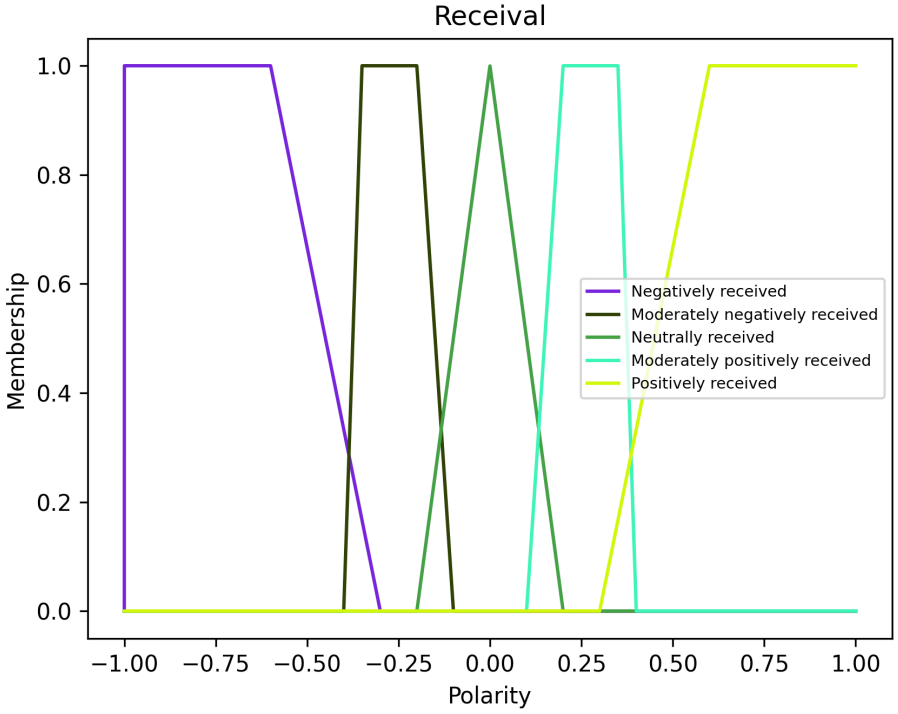
5.4. Wyliczenie współczynnika R

Do wyliczenia współczynnika R wybrano atrybuty polarity i followers. Każdemu z nich nadano 5 etykiet i stworzono funkcje przynależności:

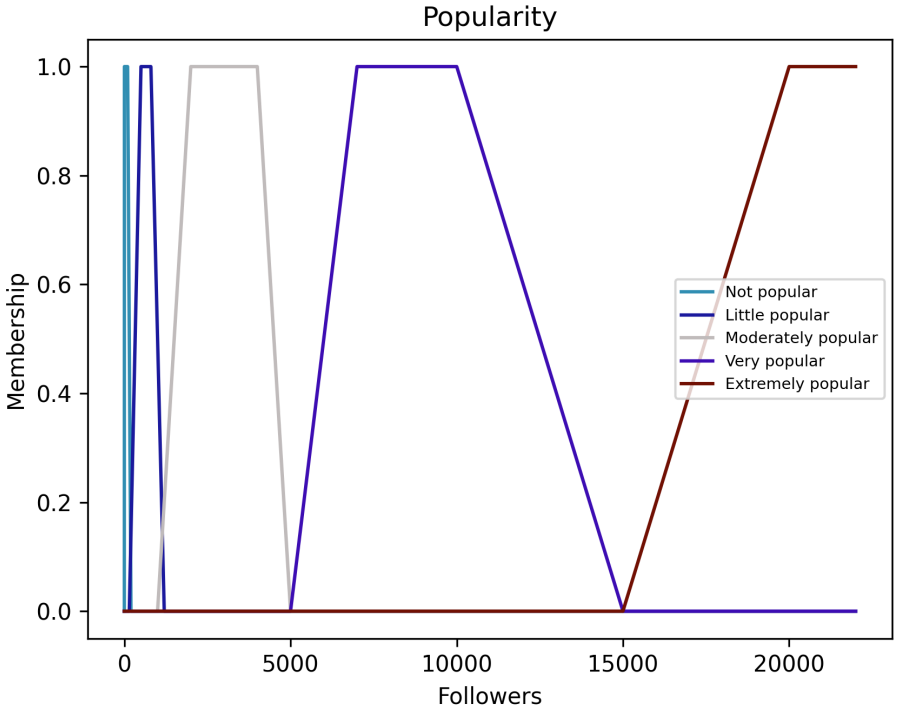
- Dla atrybutu polarity nadano etykiety:
  - Negatively received
  - Moderately negatively received
  - Neutrally received
  - Moderately positively received
  - Positively receive

  których funkcje przynależności pokazano na Rysunku 7
- Dla atrybutu followers nadano etykiety:
  - Not popular
  - Little popular
  - Moderately popular
  - Very popular
  - Extremely popular

  których funkcje przynależności pokazano na Rysunku 8



Rysunek 7. Funkcje przynależności dla atrybutu polarity



Rysunek 8. Funkcje przynależności dla atrybutu followers

Dla każdego tweeta wyliczono współczynnik R i zapisano w arkuszu *Receival to popularity R coef*. Część wyników pokazuje Tabela 5

Tabela 5. Wartości przedziałowych funkcji przynależności tweetów

Author	Tweet	Receival to popularity R coefficient
ECOWARRIORSS	{Treść}	0
ElsevierEnergy	{Treść}	0,37
siwarr5	{Treść}	0
EDITORatWORK	{Treść}	0,81
EDITORatWORK	{Treść}	0,81
mapsofworld	{Treść}	0
EnvirHealthNews	{Treść}	0,95
	...	
TheDailyClimate	{Treść}	0

6. Wnioski

1. Wykorzystana metoda porównywania tekstu nie jest symetryczna
2. Do skonstruowania odpowiedniej etykiety i funkcji przynależności wymagana jest wiedza ekspercka

Literatura

[1] *Zbiór danych z wpisami z portalu Tweeter*, https://www.kaggle.com/joseguzman/climate-sentiment-in-twitter?select=Climate_twitter.csv [dostęp: 15.11.2020]