Informatyka, studia dzienne, I st. semestr VI

Komputerowe systemy rozpoznawania **2019/2020**

Prowadzący: dr hab. inż. Adam Niewiadomski prof. uczelni pon., 12:15

Data oddania: _____ Ocena: _____

Mateusz Walczak 216911
Konrad Kajszczak 216790

Zadanie 1: Ekstrakcja cech, miary podobieństwa, klasyfikacja*

1. Cel

Celem zadania było stworzenie aplikacji służącej do klasyfikacji artykułów prasowych metodą k-NN. Korzystając z różnych metod wyboru słów kluczowych i ekstrakcji wektorów cech oraz istniejących miar podobieństwa, należało porównać przypisane przez naszą aplikacje kategorie artykułów do tych faktycznych. Należało również podjąć próbę opracowania własnej miary podobieństwa i/lub metryki.

2. Wprowadzenie

Algorytm k najbliższych sąsiadów jest bardzo prostym klasyfikatorem probabilistycznym. Niekiedy mówi się, że algorytm k-NN jest leniwy. Wynika to z faktu, że nie tworzy on wewnętrznej reprezentacji danych treningowych (uczących), ale ropoczyna poszukiwanie rozwiązania dopiero podczas analizy konkretnego wzorca ze zbioru testowego.

Algorytm przechowuje zbiór wszystkich wzorców uczących, względem których obliczana jest odległość wzorca testowego, zdefiniowana poprzez odpowiednią metrykę. Następnie algorytm wybiera k wzorców treningowych, nazywanych sąsiadami, do których aktualnie badany wzorzec testowy ma

* SVN: `https://github.com/Walducha1908/KSR1`

najmniejszą odległość. Ostateczny rezultat - kategoria, do której zostanie przypisany analizowany wzorzec - stanowi najczęściej występująca kategoria wśród k najbliższych sąsiadów.

2.1. Metryki

Do obliczenia odległości pomiędzy tekstami posłużyliśmy się następującymi metrykami:

- Metryka Euklidesowa - w celu obliczenia odległości $d_e(x, y)$ między dwoma punktami x, y należy obliczyć pierwiastek kwadratowy z sumy kwadratów różnic wartości współrzędnych o tych samych indeksach, zgodnie ze wzorem:
$$d_e(x, y) = \sqrt{(y_1 - x_1)^2 + \cdots + (y_n - x_n)^2} \tag{1}$$
- Metryka uliczna (Manhattan, miejska) - w celu obliczenia odległości $d_m(x, y)$ między dwoma punktami x, y należy obliczyć sumę wartości bezwzględnych różnic współrzędnych punktów x oraz y, zgodnie ze wzorem:
$$d_m(x, y) = \sum_{k=1}^{n} |x_k - y_k| \tag{2}$$
- Metryka Czebyszewa - w celu obliczenia odległości $d_{ch}(x, y)$ między dwoma punktami x, y należy obliczyć maksymalną wartość bezwzględnych różnic współrzędnych punktów x oraz y, zgodnie ze wzorem:
$$d_{ch}(x, y) = \max_i |x_i - y_i| \tag{3}$$
- Metryka Hamminga - definiujemy jako ilość różnic pomiędzy dwoma wektorami o tej samej długości. Aby obliczyć odległość $d_h(x, y)$ między dwoma punktami x, y należy posłużyć się wzorem [1] :
$$d_h(x, y) = \sum_{i=1}^{n} |h(i)|, \tag{4}$$
gdzie
$$h(i) = \begin{cases} 0 & \text{jeśli } v_{1i} = v_{2i} \\ 1 & \text{w przeciwnym wypadku} \end{cases} \tag{5}$$
- Odległość Canberra - ważona wersja metryki ulicznej, aby obliczyć odległość $d_c(x, y)$ między dwoma punktami x, y należy posłużyć się wzorem:
$$d_c(x, y) = \sum_i \frac{|x_i - y_i|}{|x_i| + |y_i|} \tag{6}$$

2.2. Wyznaczanie słów kluczowych

Aby wyznaczyć słowa kluczowe posługujemy się poniższą metodą:

- Term frequency - metoda polegająca na zliczeniu liczby wystąpień danego słowa we wszystkich dokumentach.

Przeprowadzamy obliczenia na zbiorze wszystkich posiadanych danych (w naszym przypadku na wszystkich artykułach) i otrzymujemy zestaw par - słowo i wartość. Taki zestaw par sortujemy malejąco po wartości i wybieramy n pierwszych słów. Wybrane n słów staje się słowami kluczowymi.

Taki schemat powtarzamy l razy, gdzie l jest liczbą kategorii na jakie klasyfikujemy. Ostatecznie otrzymujemy l zestawów słów kluczowych, przy czym każdy zestaw reprezentuje inną kategorię. Otrzymane zbiory słów kluczowych oznaczamy:

$$K_1, K_2, \ldots, K_{l-1}, K_l. \tag{7}$$

Otrzymany zbiór słów kluczowych będziemy używać we wszystkich iteracjach programu. Słowa kluczowe będą niezmienne, a wszystkie przeprowadzone przez nas eksperymenty będą bazowały na tym samym zbiorze słów kluczowych.

2.3. Wyznaczanie ważonych słów kluczowych

W celach poprawienia jakości klasyfikacji wprowadzono "ważone słowa kluczowe". Tak nazwaliśmy zestaw par - słowo kluczowe i waga (wartość zmiennoprzecinkowa), z wykorzytsaniem których przeprowadziliśmy takie same eksperymenty jak z wykorzystaniem "zwykłych" słów kluczowych, opisanych w poprzednim podpunkcie.

Ważone słowa kluczowe to nic innego jak obliczony wcześniej, ten sam zestaw słów, jednak ubogacony o wagę, obliczaną zgodnie z opracowanym przez nas wzorem:

$$W_i = \left(1 - \frac{N_{W_i \in K_l}}{l-1}\right)^2, \tag{8}$$

gdzie W_i - waga i-tego słowa kluczowego, l - liczba kategorii, $N_{W_i \in K_l}$ - liczba kategorii słów kluczowych (innych od swojej własnej), w których i-te słowo kluczowe występuje.

Dla jasności przeanalizujemy przykład. Niech $l = 3$, a obliczone słowa kluczowe mają postać:

$$K_1 = \{"jesien", "ogon", "krowa"\}, \tag{9}$$

$$K_2 = \{"wiosna", "ogon", "pies"\}, \tag{10}$$

$$K_3 = \{"lato", "ogon", "krowa"\}, \tag{11}$$

Obliczmy wartości wag dla wybranych słów kluczowych z powyższego zestawu. Dla słowa "jesien" otrzymamy następującą wartość:

$$W_{jesien} = \left(1 - \frac{0}{2}\right)^2 = 1, \tag{12}$$

słowo "jesien" wystąpiło tylko w jednej, "swojej" kategorii, ma zatem największą możliwą wagę.

Dla słowa "krowa":

$$W_{krowa} = \left(1 - \frac{1}{2}\right)^2 = 0.25, \tag{13}$$

słowo "krowa" wystąpiło w jednej dodatkowej kategorii (łącznie w dwóch).

Dla słowa "ogon":

$$W_{ogon} = \left(1 - \frac{2}{2}\right)^2 = 0, \tag{14}$$

słowo "ogon" wystąpiło we wszystkich kategoriach, dlatego też uznajemy, że nie ma dla nas żadnego znaczenia, jego waga jest równa 0.

Z powyższych rozważań bardzo jasno wynika, że wagi słów kluczowych mogą osiągać wartości z przedziału $\langle 0; 1 \rangle$.

2.4. Cechy poddawane ekstrakcji

Ekstrakcja cech charakterystycznych tekstu - w tym celu tworzymy wektor cech, który opisuje tekst (w naszym przypadku artykuł) na podstawie konkretnych, zdefiniowanych cech. Poniżej znajduje się opis wszystkich cech użytych w doświadczeniu.

Przed ekstrakcją cech, tekst został odpowiednio przygotowany. Z artykułów usunięte zostały nic nie wnoszące słowa (z tzw. "stop" listy), tekst został poddany stemizacji oraz pozbawiony znaków interpunkcyjnych.

Przyjęto następujące oznaczenia:
T_i - zbiór słów do badania,
K - stały zbiór słów kluczowych[1],
$N_{K \in T}$ - liczba wystąpień elementów zbioru K w zbiorze T[2],
$C_i(T, K)$ - wartość funkcji cechy.

2.4.1. Liczba wystąpień wszystkich słów kluczowych w całym artykule

Cecha opisująca liczbę słów kluczowych, które występują w całej sekcji głównej artykułu (body).

$$C_1(T_1, K) = N_{K \in T_1}, \tag{15}$$

gdzie T_1 - zbiór słów sekcji głównej artykułu.

Przeanalizujmy przykład obliczania wartości cechy C_1. Niech zbiór słów kluczowych K ma postać:

$$K = \{"wirus", "choroba", "zaraz", "anihilacja"\}, \tag{16}$$

[1] Na który składają się zbiory $K_1, K_2, \ldots, K_{l-1}, K_l$.

[2] W przypadku ważonych słów kluczowych będzie to suma iloczynów liczby wystąpień poszczególnych elementów zbioru K w zbiorze T i odpowiadających im wag.

zaś zbiór słów do badania (zbiór słów sekcji głównej badanego artykułu testowego) T_1 prezentuje się następująco:

$$\begin{aligned} T_1 = \{&"wirus", "niszczy", "wszystko", "droga", \\ &"zaraz", "wirus", "powodowac", "choroba"\}, \end{aligned} \tag{17}$$

Najpierw w wariancie pierwszej metody ekstrakcji - wykorzystując metodę TF i zwykłe słowa kluczowe. Przeanalizujmy występowanie elementów zbioru K w zbiorze T_1:

- "wirus" - występuje 2 razy,
- "choroba" - występuje 1 raz,
- "zaraz" - występuje 1 raz,
- "anihilacja" - nie występuje ani razu.

Po dodaniu wszystkich wystąpień otrzymujemy:

$$C_1(T_1, K) = N_{K \in T_1} = 2 + 1 + 1 + 0 = 4. \tag{18}$$

Teraz zajmijmy się drugą metodą ekstrakcji - wykorzystując ważone słowa kluczowe. Załóżmy, że pary słów kluczowych wraz z obliczonymi wagami dla słów kluczowych zbioru K prezentują się następująco:

$$K_w = \{("wirus", 0.25), ("choroba", 1), ("zaraz", 0), ("anihilacja", 1)\}, \tag{19}$$

W tym przypadku zgodnie z wcześniej zaprezentowanym opisem, musimy obliczyć sumę iloczynów liczby wystąpień poszczególnych elementów zbioru K w zbiorze T_1 i odpowiadających im wag:

$$C_1(T_1, K_w) = N_{K \in T_1} = 2 \cdot 0.25 + 1 \cdot 1 + 1 \cdot 0 + 0 \cdot 1 = 0.5 + 1 + 0 + 0 = 1.5. \tag{20}$$

2.4.2. Liczba wystąpień wszystkich słów kluczowych w tytule artykułu

Cecha opisująca liczbę słów kluczowych, które występują w tytule artykułu (title).

$$C_2(T_2, K) = N_{K \in T_2}, \tag{21}$$

gdzie T_2 - zbiór słów tytułu artykułu.

2.4.3. Liczba wystąpień wszystkich słów kluczowych w sekcji daty artykułu

Cecha opisująca liczbę słów kluczowych, które występują w sekcji daty artykułu (dateline).

$$C_3(T_3, K) = N_{K \in T_3}, \tag{22}$$

gdzie T_3 - zbiór słów sekcji daty artykułu.

2.4.4. Stosunek liczby wystąpień wszystkich słów kluczowych do ogólnej liczby słów w artykule

Cecha opisująca stosunek liczby słów kluczowych, które występują w całej sekcji głównej artykułu (body), do całkowitej liczby słów występujących w części głównej.

$$C_4(T_4, K) = \frac{N_{K \in T_4}}{|T_4|}, \tag{23}$$

gdzie T_4 - zbiór słów sekcji głównej artykułu, $|T_4|$ - liczba elementów (słów) zbioru sekcji głównej artykułu.

W tym miejscu warto wspomnieć, że w przypadku ważonych słów kluczowych wartość $|T_4|$ będzię iloczynem liczby elementów zbioru sekcji głównej artykułu i maksymalnej wartości osiągalnej przez wagi. Jednak ponieważ maksymalną możliwą wartością wagi słowa kluczowego jest 1 (zgodnie z rozdziałem 2.3) to w obu przypadkach - zwykłych słów kluczowych jak i ważonych słów kluczowych - będzie to dokładnie ta sama wartość liczbowa.

2.4.5. Liczba wystąpień wszystkich słów kluczowych w pierwszych 50 słowach artykułu

Cecha opisująca liczbę słów kluczowych, które występują w pierwszych 50 słowach sekcji głównej artykułu. Jeśli artykuł jest krótszy niż 50 słów to bierzemy pod uwagę wszystkie występujące w nim słowa.

$$C_5(T_5, K) = N_{K \in T_5}, \tag{24}$$

gdzie T_5 - pierwsze 50 słów sekcji głównej artykułu.

2.4.6. Liczba wystąpień wszystkich słów kluczowych w pierwszych 10% artykułu

Cecha opisująca liczbę słów kluczowych, które występują w pierwszych 10% sekcji głównej artykułu.

$$C_6(T_6, K) = N_{K \in T_6}, \tag{25}$$

gdzie T_6 - pierwsze 10% słów sekcji głównej artykułu.

2.4.7. Liczba wystąpień wszystkich słów kluczowych w pierwszych 20% artykułu

Cecha opisująca liczbę słów kluczowych, które występują w pierwszych 20% sekcji głównej artykułu.

$$C_7(T_7, K) = N_{K \in T_7}, \tag{26}$$

gdzie T_7 - pierwsze 20% słów sekcji głównej artykułu.

2.4.8. Liczba wystąpień wszystkich słów kluczowych w pierwszych 50% artykułu

Cecha opisująca liczbę słów kluczowych, które występują w pierwszych 50% sekcji głównej artykułu.

$$C_8(T_8, K) = N_{K \in T_8}, \tag{27}$$

gdzie T_8 - pierwsze 50% słów sekcji głównej artykułu.

2.4.9. Liczba wystąpień wszystkich słów kluczowych w pierwszym paragrafie

Cecha opisująca liczbę słów kluczowych, które występują w pierwszym paragrafie sekcji głównej artykułu.

$$C_9(T_9, K) = N_{K \in T_9}, \tag{28}$$

gdzie T_9 - pierwszy paragraf sekcji głównej artykułu.

2.4.10. Liczba wystąpień wszystkich słów kluczowych w ostatnich 50 słowach artykułu

Cecha opisująca liczbę słów kluczowych, które występują w ostatnich 50 słowach sekcji głównej artykułu. Jeśli artykuł jest krótszy niż 50 słów to bierzemy pod uwagę wszystkie występujące w nim słowa.

$$C_{10}(T_{10}, K) = N_{K \in T_{10}}, \tag{29}$$

gdzie T_{10} - ostatnie 50 słów sekcji głównej artykułu.

2.4.11. Liczba wystąpień wszystkich słów kluczowych w ostatnich 10% artykułu

Cecha opisująca liczbę słów kluczowych, które występują w ostatnich 10% sekcji głównej artykułu.

$$C_{11}(T_{11}, K) = N_{K \in T_{11}}, \tag{30}$$

gdzie T_{11} - ostatnie 10% słów sekcji głównej artykułu.

2.4.12. Liczba wystąpień wszystkich słów kluczowych w ostatnim paragrafie

Cecha opisująca liczbę słów kluczowych, które występują w ostatnim paragrafie sekcji głównej artykułu.

$$C_{12}(T_{12}, K) = N_{K \in T_{12}}, \tag{31}$$

gdzie T_{12} - ostatni paragraf sekcji głównej artykułu.

3. Opis implementacji

Praca w toku

4. Materiały i metody

W tym rozdziale omówione zostaną poszczególne eksperymenty jakie wykonano z użyciem naszego programu.

Klasyfikacje artykułów przeprowadzano ze względu na dwa różne rodzaje etykiet. Pierwszym z nich była lokalizacja (place). Kategorie (etykiety) jakie wyróżniliśmy były następujące: west-germany, usa, france, uk, canada, japan. Klasyfikacja przeprowadzana była jedynie z wykorzystaniem artykułów, których pole "places" przyjmowało jedną z powyższych wartości.

Drugim rodzajem etykiet był temat (topic). Kategorie (etykiety) jakie wyróżniliśmy były następujące: earn, trade, money-supply, acq. Podobnie jak w pierwszym przypadku, klasyfikacja przeprowadzana była jedynie z wykorzystaniem artykułów, których pole "topics" przyjmowało jedną z powyższych wartości.

4.1. Wpływ liczby k sąsiadów oraz wyboru metryki na klasyfikację

Klasyfikacja tekstów została wykonana z wykorzystaniem zbioru (zwykłych) słów kluczowych. Eksperymenty wykonano z użyciem wszystkich pięciu metryk. Dla każdego przypadku testowego dokonano klasyfikacji tekstu dla następujących wartości współczynnika k:

$$k \in \{1, 3, 4, 6, 8, 10, 12, 14, 17, 20\}. \tag{32}$$

W każdym przypadku testowym zbiór treningowy stanowił 70% artykułów, zaś zbiór testowy 30% artykułów.

4.2. Wpływ podziału tekstów na zbiory treningowe i testowe na klasyfikację

Klasyfikacja tekstów została wykonana z wykorzystaniem zbioru (zwykłych) słów kluczowych. Eksperymenty przeprowadzono posługując się metryką Euklidesową. Wartość parametru k była stała i wynosiła $k = 6$. Przeprowadzono klasyfikacje dla pięciu różnych podziałów artykułów na zbiory testowe i treningowe:

- Zbiór treningowy: 40% artykułów, zbiór testowy 60%,
- Zbiór treningowy: 50% artykułów, zbiór testowy 50%,
- Zbiór treningowy: 60% artykułów, zbiór testowy 40%,
- Zbiór treningowy: 70% artykułów, zbiór testowy 60%,
- Zbiór treningowy: 80% artykułów, zbiór testowy 20%.

4.3. Wpływ konkretnych cech na klasyfikację

Klasyfikacja tekstów została wykonana z wykorzystaniem zbioru (zwykłych) słów kluczowych. Eksperymenty przeprowadzono posługując się metryką Euklidesową. Wartość parametru k była stała i wynosiła $k = 6$. W każdej iteracji programu zbiór treningowy stanowił 70% artykułów, zaś zbiór testowy 30% artykułów. Przeprowadzono klasyfikacje dla czterech różnych zestawów cech, wybranych spośród wszystkich cech omówionych w rozdziale 2.4. Wybrane zestawy cech były następujące (aby nie duplikować treści, w tym miejscu posługuję się indeksami funkcji cech z rodziału 2.4):

- Zestaw 1: C_1, C_2, C_3, C_4, C_{10}, C_{11}, C_{12},
- Zestaw 2: C_1, C_2, C_3, C_4,
- Zestaw 2: C_5, C_6, C_7, C_8, C_9,
- Zestaw 2: C_2, C_3, C_6, C_{11}.

4.4. Wpływ użycia ważonych słów kluczowych na klasyfikację

Klasyfikacja tekstów została wykonana z wykorzystaniem zbioru zwykłych oraz z użyciem ważonych słów kluczowych. Eksperymenty wykonano z użyciem wszystkich pięciu metryk. Wartość parametru k była stała i wynosiła $k = 6$. W każdym przypadku testowym zbiór treningowy stanowił 70% artykułów, zaś zbiór testowy 30% artykułów.

5. Wyniki

W tym rozdziale zamieszczono tabele oraz wykresy prezentujące wyniki przeprowadzanych przez nas eksperymentów.

5.1. Wpływ liczby k sąsiadów oraz wyboru metryki na klasyfikację

k	**places** [%]	**topics** [%]
1	81,99	91,35
3	84,99	95,06
4	85,41	94,69
6	85,79	96,23
8	85,14	95,97
10	85,07	94,90
12	85,34	95,49
14	85,19	95,38
17	85,07	95,49
20	85,04	95,70

Tabela 1. Skuteczność klasyfikacji dla metryki Euklidesowej

k	**places** [%]	**topics** [%]
1	81,64	86,25
3	84,92	88,06
4	85,36	87,74
6	85,86	88,91
8	84,35	88,75
10	84,27	88,96
12	84,35	89,28
14	84,35	87,58
17	84,20	87,53
20	84,12	87,05

Tabela 2. Skuteczność klasyfikacji dla metryki Chebysheva

k	places [%]	topics [%]
1	82,15	94,64
3	86,26	96,66
4	86,26	96,76
6	87,00	97,03
8	87,10	96,97
10	86,93	96,92
12	86,78	97,03
14	86,75	96,82
17	86,28	96,82
20	86,08	96,60

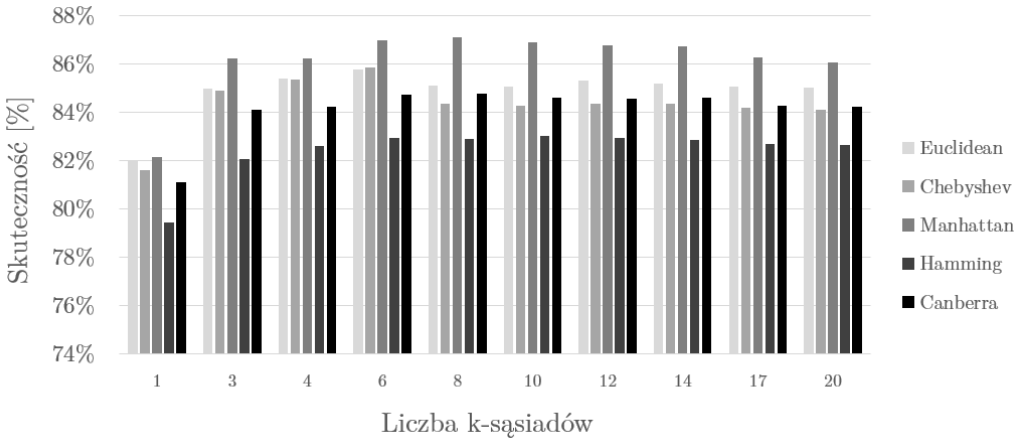
Tabela 3. Skuteczność klasyfikacji dla metryki ulicznej

k	places [%]	topics [%]
1	79,46	92,20
3	82,09	94,11
4	82,63	93,79
6	82,96	93,90
8	82,91	94,32
10	83,03	94,11
12	82,96	94,00
14	82,86	94,00
17	82,71	94,00
20	82,66	94,16

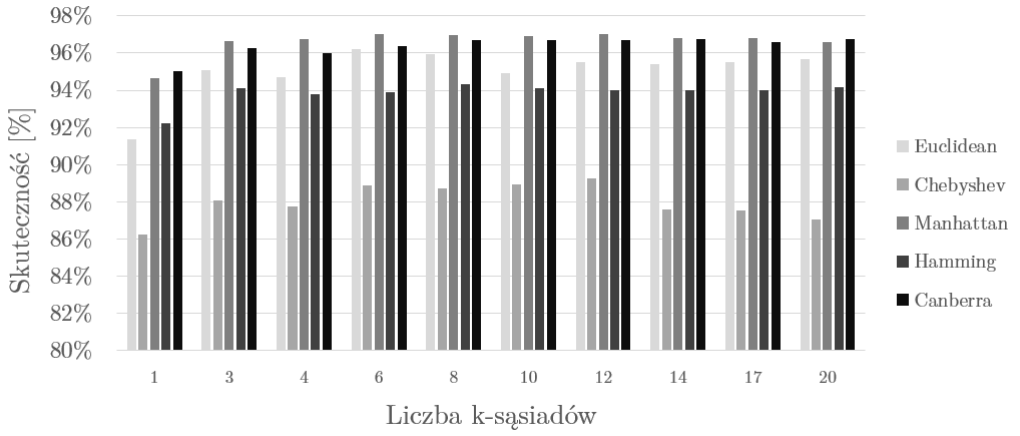
Tabela 4. Skuteczność klasyfikacji dla metryki Hamminga

k	places [%]	topics [%]
1	81,10	95,01
3	84,12	96,28
4	84,25	96,02
6	84,74	96,39
8	84,79	96,71
10	84,62	96,71
12	84,59	96,71
14	84,64	96,76
17	84,30	96,60
20	84,25	96,76

Tabela 5. Skuteczność klasyfikacji dla metryki Canberra



Rysunek 1. Wizualizacja danych z Tabel 1-5 dla kategorii "places"

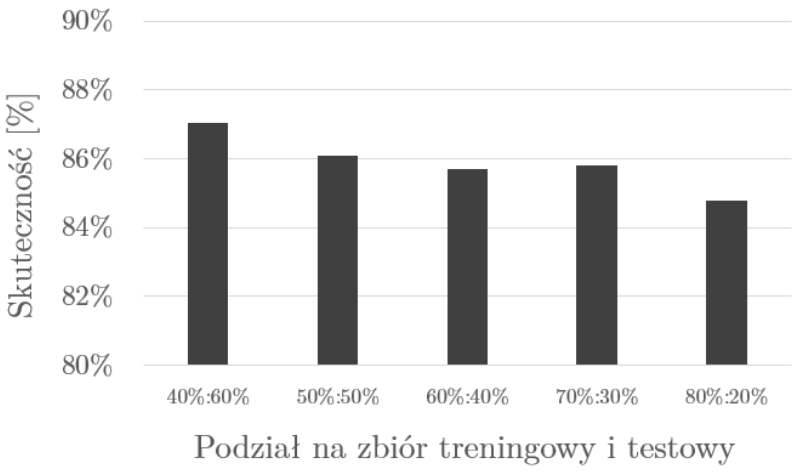


Rysunek 2. Wizualizacja danych z Tabel 1-5 dla kategorii "topics"

5.2. Wpływ podziału tekstów na zbiory treningowe i testowe na klasyfikację

Podział	**places [%]**	**topics [%]**
40:60	87,04	93,89
50:50	86,09	94,52
60:40	85,69	94,71
70:30	85,79	96,23
80:20	84,78	96,74

Tabela 6. Skuteczność klasyfikacji dla różnych podziałów artykułów (podano w kolejności treningowe:testowe)



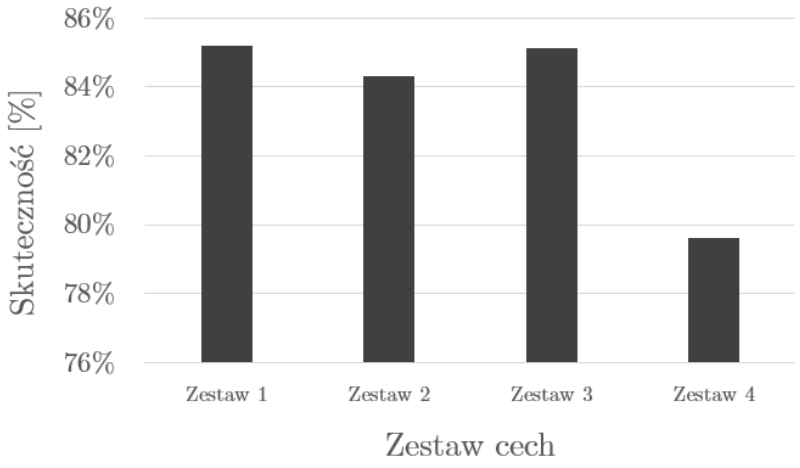
Rysunek 3. Wziualizacja danych z Tabeli 6 dla kategorii "places"

Rysunek 4. Wizualizacja danych z Tabeli 6 dla kategorii "topics"

5.3. Wpływ konkretnych cech na klasyfikację

Zestaw	**places [%]**	**topics [%]**
1	85,19	95,91
2	84,32	95,38
3	85,14	96,18
4	79,61	79,41

Tabela 7. Skuteczność klasyfikacji dla różnych zestawów cech



Rysunek 5. Wizualizacja danych z Tabeli 7 dla kategorii "places"

Rysunek 6. Wizualizacja danych z Tabeli 7 dla kategorii "topics"

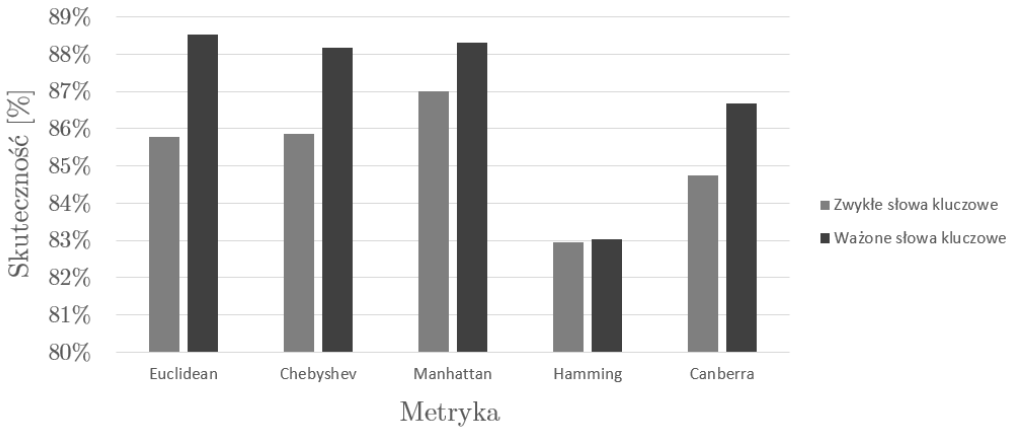
5.4. Wpływ użycia ważonych słów kluczowych na klasyfikację

Metryka	**zwykłe słowa kluczowe [%]**	**ważone słowa klczuowe [%]**
Euclidean	85,79	88,54
Chebyshev	85,86	88,17
Manhattan	87,00	88,32
Hamming	82,96	83,03
Canberra	84,74	86,68

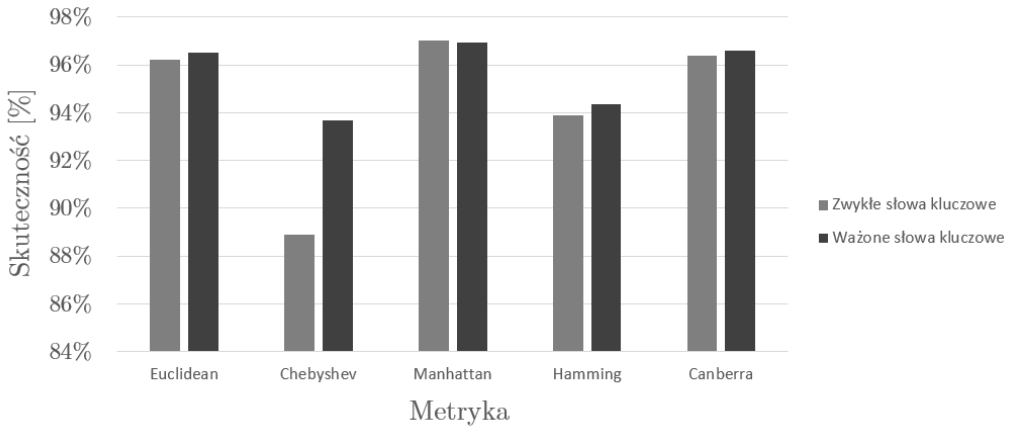
Tabela 8. Skuteczność klasyfikacji dla różnych metod ekstrakcji - zywkłe i ważone słowa kluczowe - kategoria "places"

Metryka	zwykłe słowa kluczowe [%]	ważone słowa klczuowe [%]
Euclidean	96,23	96,50
Chebyshev	88,91	93,68
Manhattan	97,03	96,92
Hamming	93,90	94,37
Canberra	96,39	96,60

Tabela 9. Skuteczność klasyfikacji dla różnych metod ekstrakcji - zywkłe i ważone słowa kluczowe - kategoria "topics"



Rysunek 7. Wizualizacja danych z Tabeli 8



Rysunek 8. Wizualizacja danych z Tabeli 9

6. Dyskusja

Praca w toku

7. Wnioski

Praca w toku

Literatura

[1] A. Niewiadomski *Materiały, przykłady i ćwiczenia do przedmiotu Komputerowe Systemy Rozpoznawania*. 19 czerwca 2012.