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 1: Ekstrakcja cech, miary podobieństwa, klasyfikacja*

1. Cel

Celem zadania było stworzenie aplikacji do klasyfikacji tekstów metodą k-NN, korzystając z różnych sposobów ekstrakcji wektorów cech i metryk. Dodatkowo przeanalizowanie dla jakich parametrów program ma największą skuteczność.

2. Wprowadzenie

Zagadnieniem, wokół którego skupiona jest nasza aplikacja jest klasyfikacja statystyczna, która jest rodzajem algorytmu statystycznego przydzielającego elementy do klas, bazując na cechach owych elementów. W ramach przeprowadzanego eksperymentu zaimplementowaliśmy klasyfikator k-najbliższych sąsiadów.

Algorytm k najbliższych sąsiadów, nazywany także algorytmem k-NN, należy do grupy algorytmów leniwych, czyli takich, które nie tworzą wewnętrznej reprezentacji danych uczących, lecz szukają rozwiązania dopiero w momencie pojawienia się wzorca testującego. Przechowuje wszystkie wzorce uczące, względem których wyznacza odległość wzorca testowego [2]. Metoda

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

k-NN wyznacza k sąsiadów, do których badany element ma najmniejszą odległość w danej metryce, a następnie wyznacza wynik w oparciu o najczęstszy element, wśród k najbliższych. W przypadku naszego projektu odległość definiujemy jako skalę podobieństwa tekstów.

W ramach zadania zostały użyte 2 metody ekstrakcji cech:

- Inverse document frequency - metoda polegająca na wyznaczeniu, czy dane słowo występuje powszechnie we wszystkich dokumentach. Jest to logarytmicznie skalowana odwrotna część dokumentów zawierających wybrane słowo (uzyskana poprzez podzielenie całkowitej liczby dokumentów przez liczbę dokumentów zawierających ten termin). Obliczana jest z poniższego wzoru:
$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$
- Term frequency - metoda polegająca na zliczeniu częstości występowania danego słowa w dokumencie. Obliczana jest z poniższego wzoru:
$$tf_{i,j} = \frac{n_{i,j}}{\sum_k nk, j}$$

Do obliczenia odległości tekstów posłużyliśmy się 3 metrykami:

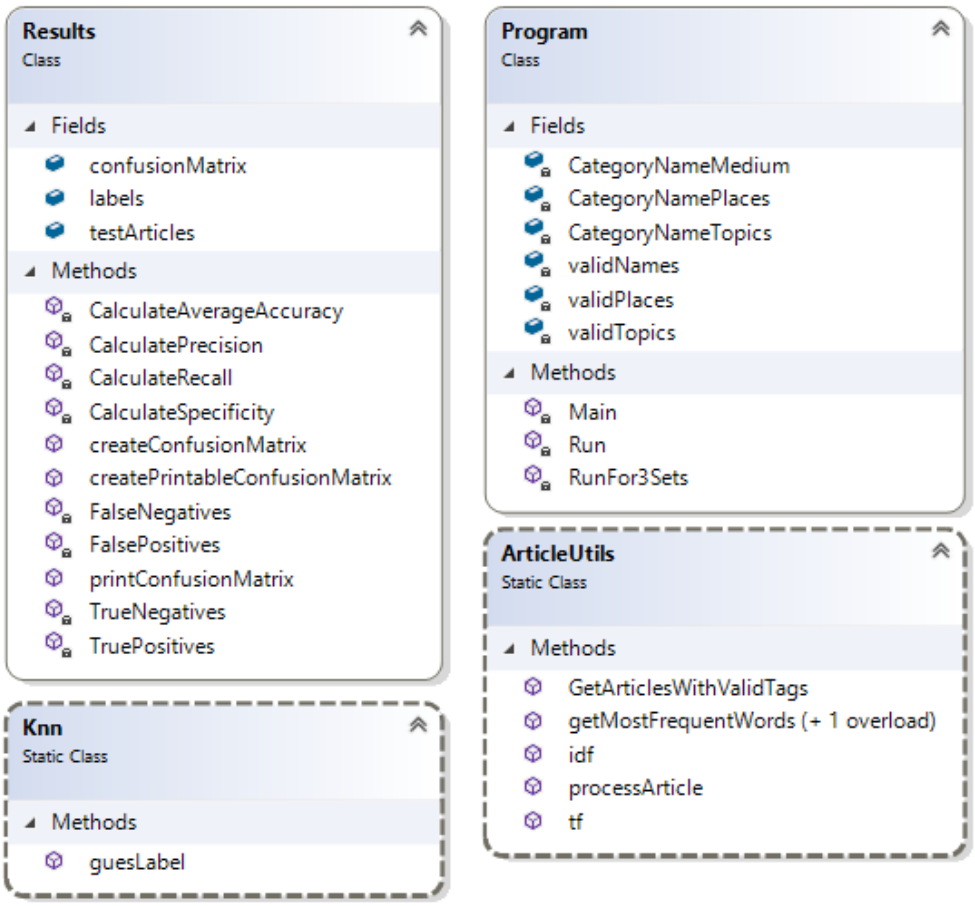
- metryka Euklidesowa - w celu obliczenia odległości $d_e(x, y)$ między dwoma punktami x, y należy obliczyć pierwiastek kwadratowy z sumy drugich potęg różnic wartości współrzędnych o tych samych indeksach, zgodnie ze wzorem:
$$d_e(x, y) = \sqrt{(y_1 - x_1)^2 + \cdots + (y_n - x_n)^2}$$
- metryka uliczna (Manhattan, miejska) - w celu obliczenia odległości $d_e(x, y)$ między dwoma punktami x, y należy obliczyć sumę wartości bezwzględnych różnic współrzędnych punktów x oraz y, zgodnie ze wzorem:
$$d_m(x, y) = \sum_{k=1}^{n} |x_k - y_k|$$
- metryka Czebyszewa - w celu obliczenia odległości $d_e(x, y)$ między dwoma punktami x, y należy obliczyć maksymalną wartość bezwzględnych różnic współrzędnych punktów x oraz y, zgodnie ze wzorem:
$$d_{ch}(x, y) = \max_i |x_i - y_i|$$

3. Opis implementacji

Program został w całości stworzony w języku C# (.NET Framework, v4.6.1).

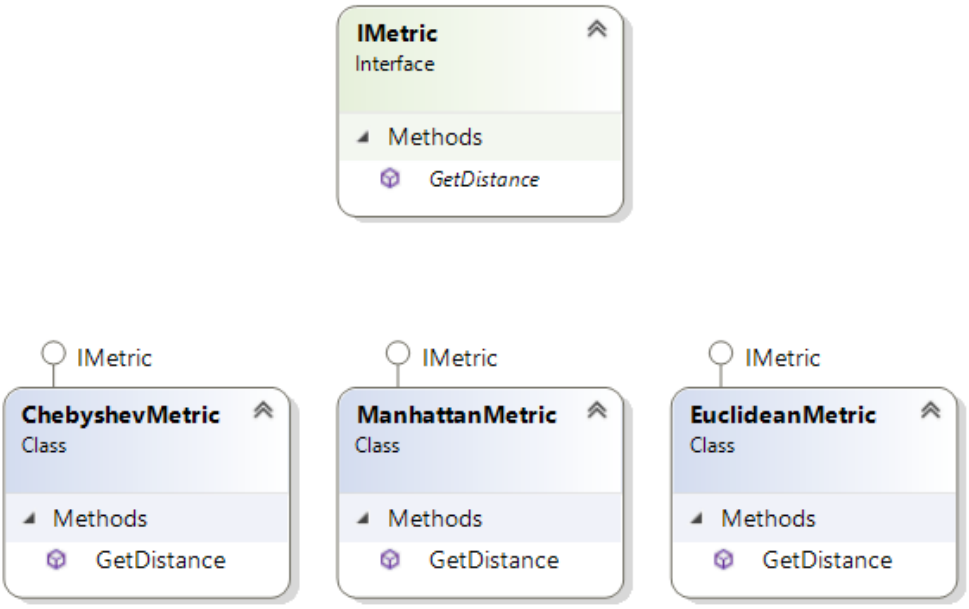
W programie znajdują się następujące klasy:

- Knn - implementacja algorytmu K-nn
- Program - klasa główna
- ArticleUtilis - klasa pomocnicza
- Results - klasa odpowiedzialna za dokładne zestawienie rezultatów



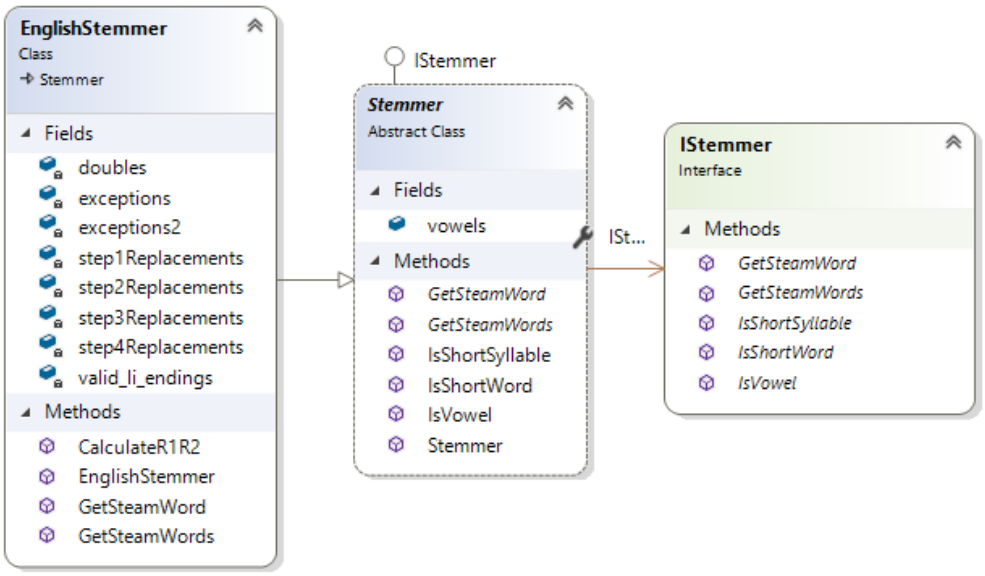
Rysunek 1. Diagram UML wygenerowany dla klas ogólnych.

- IMetric - interfejs dla metryk, klasy implementujące owy interfejs:
  - ChebyshewMetric - klasa odpowiedzialna za metrykę Czebyszewa
  - ManhatattanMetric - klasa odpowiedzialna za metrykę uliczną
  - EuclideanMetric - klasa odpowiedzialna za metrykę euklidesową



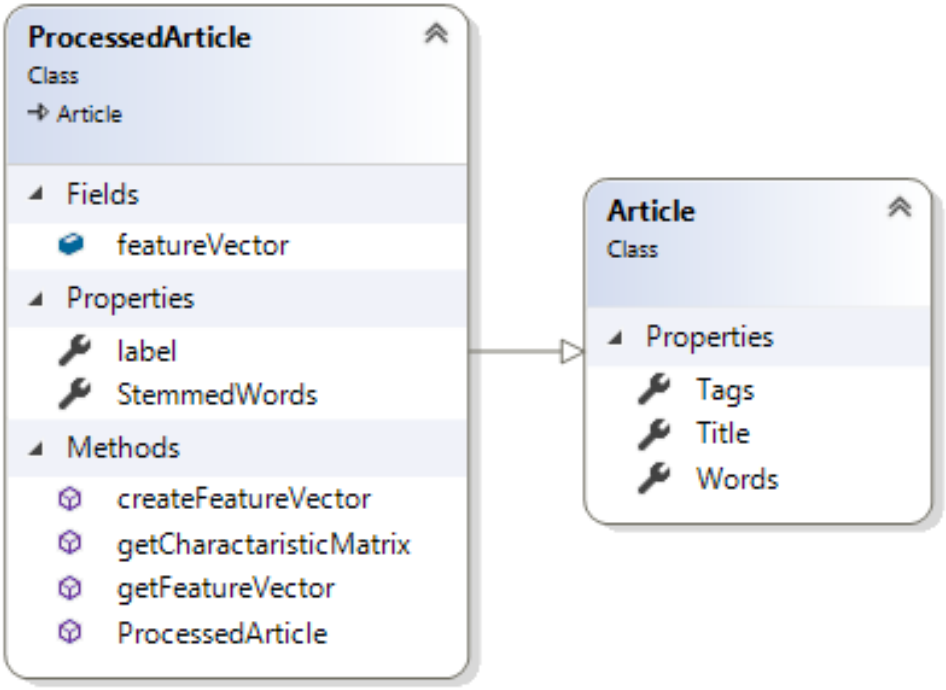
Rysunek 2. Diagram UML wygenerowany dla klas dotyczących metryk.

- IStemmer - interfejs reprezentujący stemmer
- Stemmer - reprezentacja stemera
- EnglishStemer - stemer dla języka angielskiego



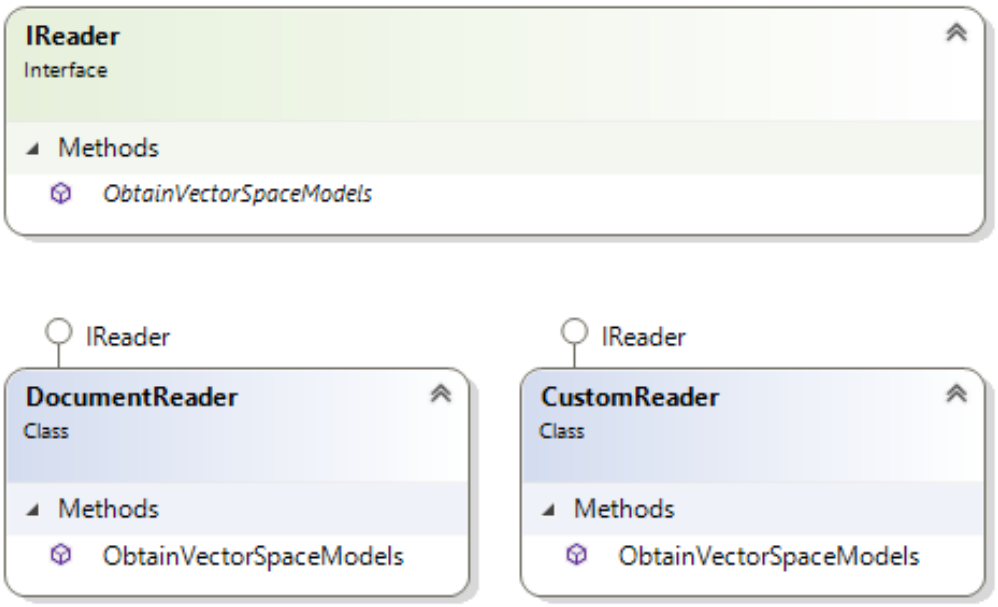
Rysunek 3. Diagram UML wygenerowany dla klas dotyczących procesu stemizacji.

- Article - klasa reprezentująca pojedyńczy artykuł
- ProcessedArticle - klasa dziedzicząca po Article, reprezentuje przetworzony artykuł



Rysunek 4. Diagram UML wygenerowany dla klas reprezentujących artykuły.

- IReader - interfejs dla klas wczytujących dane, klasy implementujące owy interfejs:
  - DocumentReader - klasa odpowiedzialna za wczytanie plików z zestawu [4]
  - CustomReader - klasa odpowiedzialna za wczytanie plików z przygotowanego przez nas zestawu [5]



Rysunek 5. Diagram UML wygenerowany dla klas wczytujących dane.

4. Materiały i metody

Klasyfikacja tekstów została wykonana wszystkimi dostępnymi metodami ekstrakcji cech dla wszystkich trzech metryk. Dla każdego przypadku testowego dokonano klasyfikacji tekstu dla k $\in \{2, 3, 5, 7, 10, 15, 20\}$ najbliższych sąsiadów. Wyniki porównano z faktyczną etykietą danego artykułu.

Przyjelismy trzy różne proporcje zbioru treningowego do zbioru testowego, było to:

— zbiór treningowy 20%, zbiór testowy 80%
— zbiór treningowy 60%, zbiór testowy 40%
— zbiór treningowy 80%, zbiór testowy 20%

Klasyfikacja dotycząca lokalizacji przeprowadzana była jedynie na danych, których pole places przyjmowało jedną z wartości: west-germany, usa, france, uk, canada, japan.

Klasyfikacja dotycząca tematów przeprowadzana była jedynie na danych, które pole topics przyjmowało jedną z wartości: gold, cocoa, sugar, coffe, grain.

Dane które przygotowaliśmy do analizy były opisy dzieł kultury. Posiadały one jedną kategorię - medium. Pole przyjmowało wartość: book lub movie.

5. Wyniki

5.1. Term frequency

5.1.1. Metryka Euklidesowa

k	places [%]	topics [%]	medium [%]
2	80,3	42,4	33,8
3	83,8	46,7	62,5
5	84	38	75
7	83,6	37	48,8
10	83,3	35,9	37,5
15	82,8	35,9	47,5
20	82	35,9	48,8

Tabela 1. Skuteczność klasyfikacji dla podziału: zbiór treningowy 20%, zbiór testowy 80%

k	places [%]	topics [%]	medium [%]
2	81,3	54,3	45
3	85,3	58,7	47,5
5	85,4	47,8	57,5
7	85,2	47,8	57,5
10	84,5	45,7	52,5
15	83,6	39,1	67,5
20	83,3	21,7	67,5

Tabela 2. Skuteczność klasyfikacji dla podziału: zbiór treningowy 60%, zbiór testowy 40%

k	places [%]	topics [%]	medium [%]
2	84,6	47,8	75
3	86,4	56,5	80
5	86,9	34,8	65
7	87	34,8	80
10	86,6	43,5	70
15	85,5	43,5	90
20	85,6	34,8	85

Tabela 3. Skuteczność klasyfikacji dla podziału: zbiór treningowy 80%, zbiór testowy 20%

5.1.2. Metryka uliczna

k	places [%]	topics [%]	medium [%]
2	79,6	35,9	78,8
3	82,1	48,9	62,5
5	81,8	35,9	53,8
7	81,7	37	53,8
10	81,8	35,9	47,5
15	81,3	35,9	46,2
20	80,7	35,9	48,8

Tabela 4. Skuteczność klasyfikacji dla podziału: zbiór treningowy 20%, zbiór testowy 80%

k	places [%]	topics [%]	medium [%]
2	79,9	52,2	50
3	83,9	54,3	47,5
5	83,7	52,2	67,5
7	83,5	47,8	70
10	82,7	47,8	60
15	82	45,7	67,5
20	81,7	23,9	47,5

Tabela 5. Skuteczność klasyfikacji dla podziału: zbiór treningowy 60%, zbiór testowy 40%

k	places [%]	topics [%]	medium [%]
2	84,1	47,8	65
3	85,1	47,8	90
5	85,6	43,5	65
7	84,7	30,4	80
10	84,4	34,8	50
15	83,6	47,8	50
20	83,3	39,1	85

Tabela 6. Skuteczność klasyfikacji dla podziału: zbiór treningowy 80%, zbiór testowy 20%

5.1.3. Metryka Czebyszewa

k	places [%]	topics [%]	medium [%]
2	80,4	42,4	53,8
3	82,4	45,7	53,8
5	83,1	37	68,8
7	83,2	45,7	56,2
10	82,9	35,9	61,3
15	82,3	35,9	52,5
20	81,6	35,9	38,8

Tabela 7. Skuteczność klasyfikacji dla podziału: zbiór treningowy 20%, zbiór testowy 80%

k	places [%]	topics [%]	medium [%]
2	78,2	43,5	40
3	82,3	47,8	57,5
5	82,8	45,7	77,5
7	83	45,7	80
10	82,8	43,5	45
15	82,6	32,6	60
20	82	28,3	50

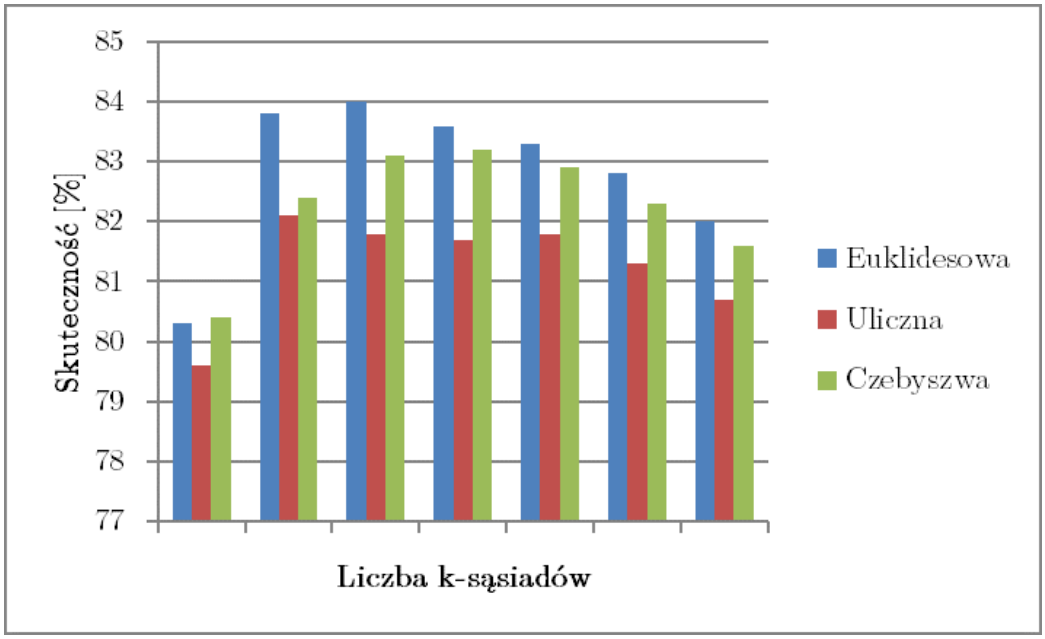
Tabela 8. Skuteczność klasyfikacji dla podziału: zbiór treningowy 60%, zbiór testowy 40%

k	places [%]	topics [%]	medium [%]
2	82,1	30,4	60
3	84,6	39,1	70
5	84,7	26,1	75
7	84,5	30,4	45
10	84,4	30,4	40
15	84,3	26,1	75
20	83,3	26,1	60

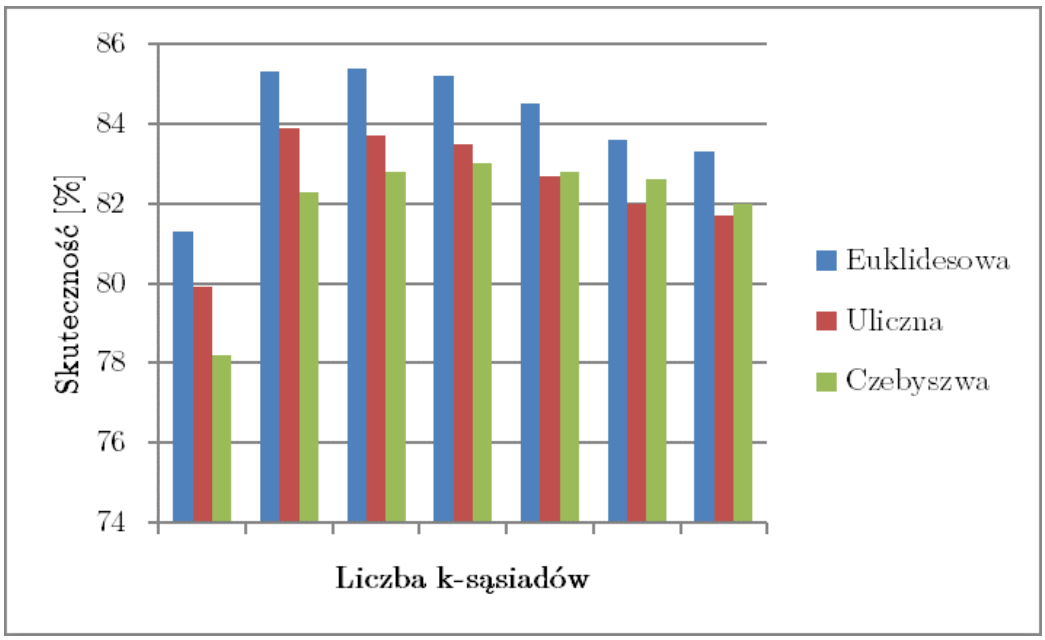
Tabela 9. Skuteczność klasyfikacji dla podziału: zbiór treningowy 80%, zbiór testowy 20%

5.2. Term frequency - podsumowanie wykresami

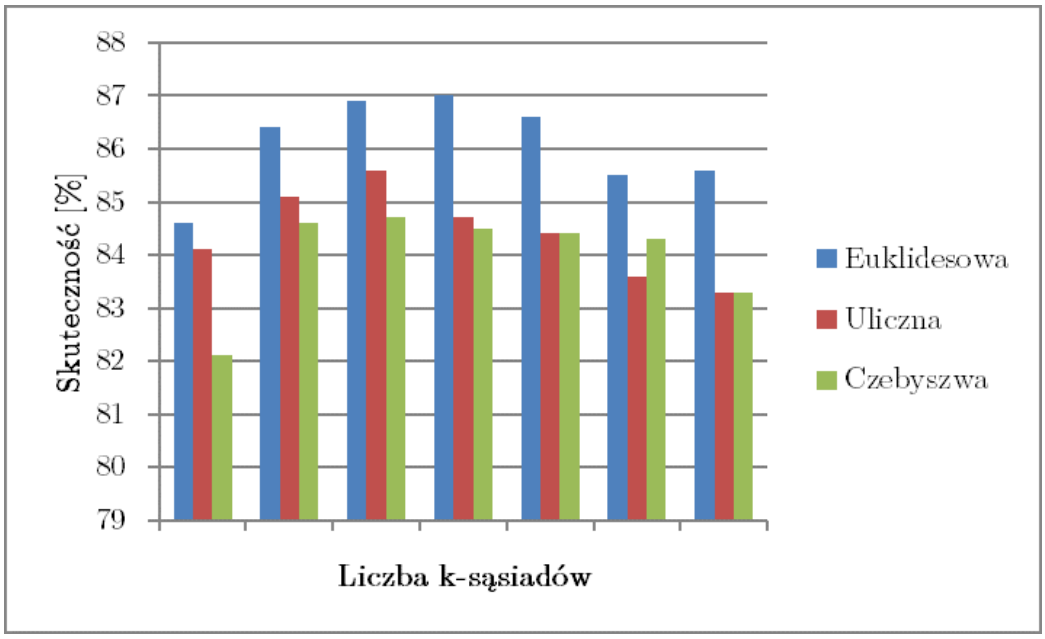
5.2.1. Places



Rysunek 6. Dane z tabel 1-9 dla kategorii places, zbiór treningowy 20%, zbiór testowy 80%

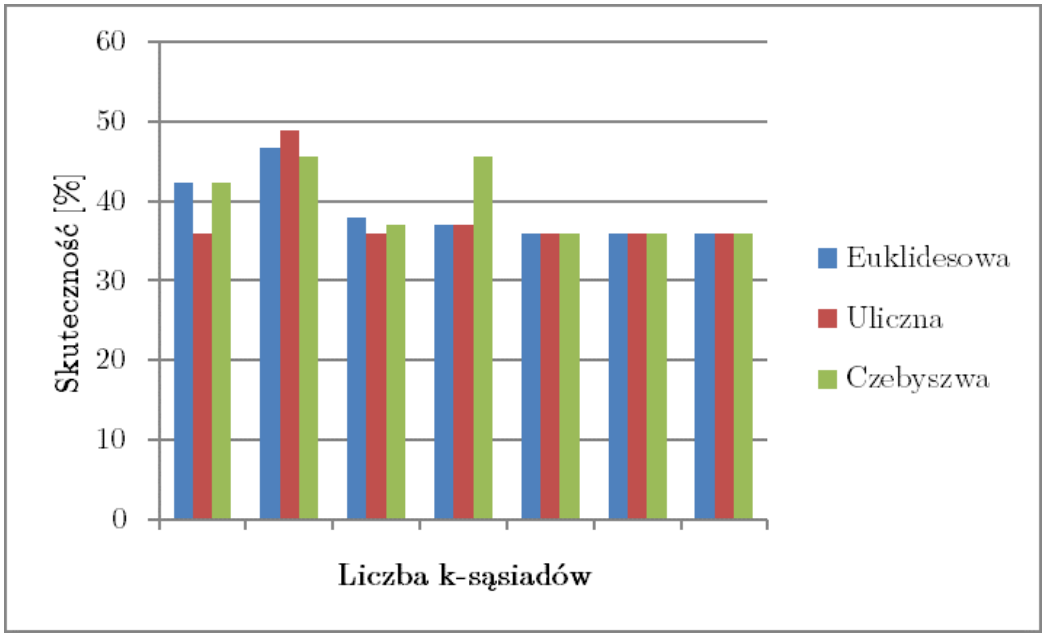


Rysunek 7. Dane z tabel 1-9 dla kategorii places, zbiór treningowy 60%, zbiór testowy 40%

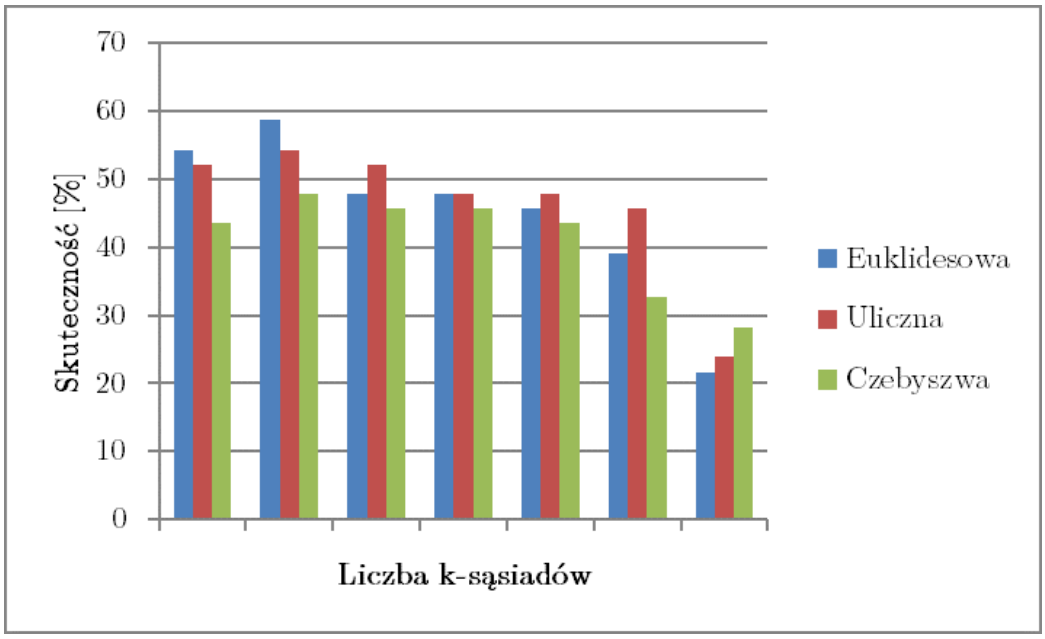


Rysunek 8. Dane z tabel 1-9 dla kategorii places, zbiór treningowy 80%, zbiór testowy 20%

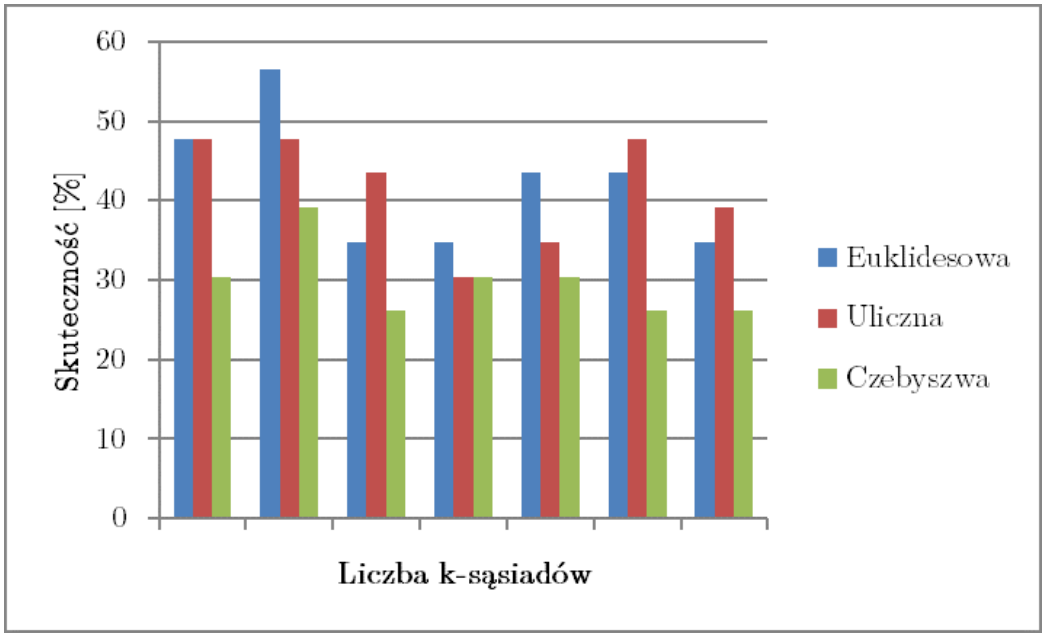
5.2.2. Topics



Rysunek 9. Dane z tabel 1-9 dla kategorii topics, zbiór treningowy 20%, zbiór testowy 80%

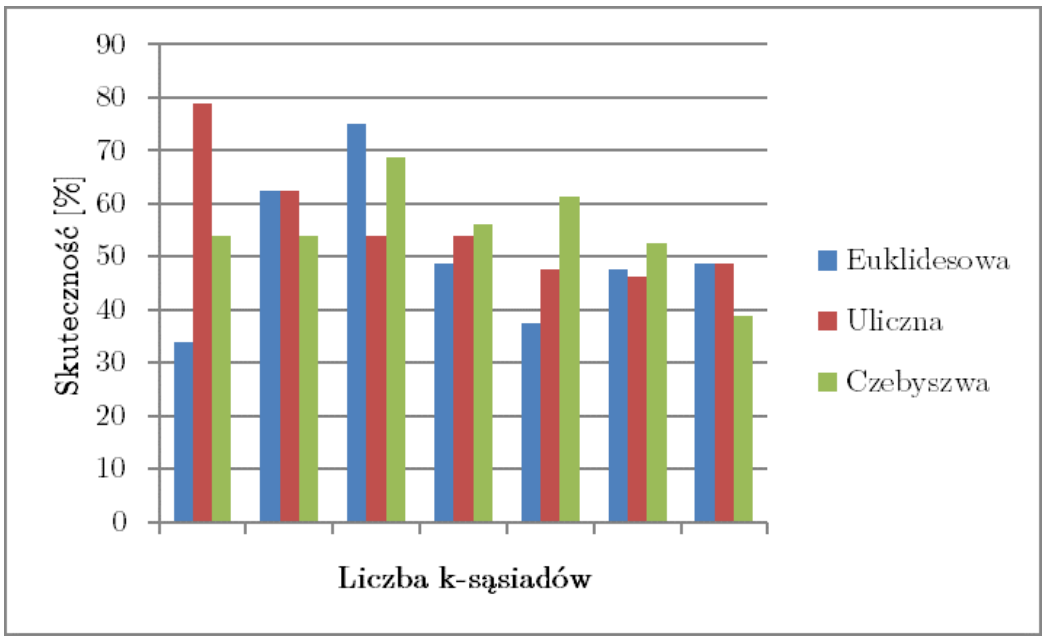


Rysunek 10. Dane z tabel 1-9 dla kategorii topics, zbiór treningowy 60%, zbiór testowy 40%

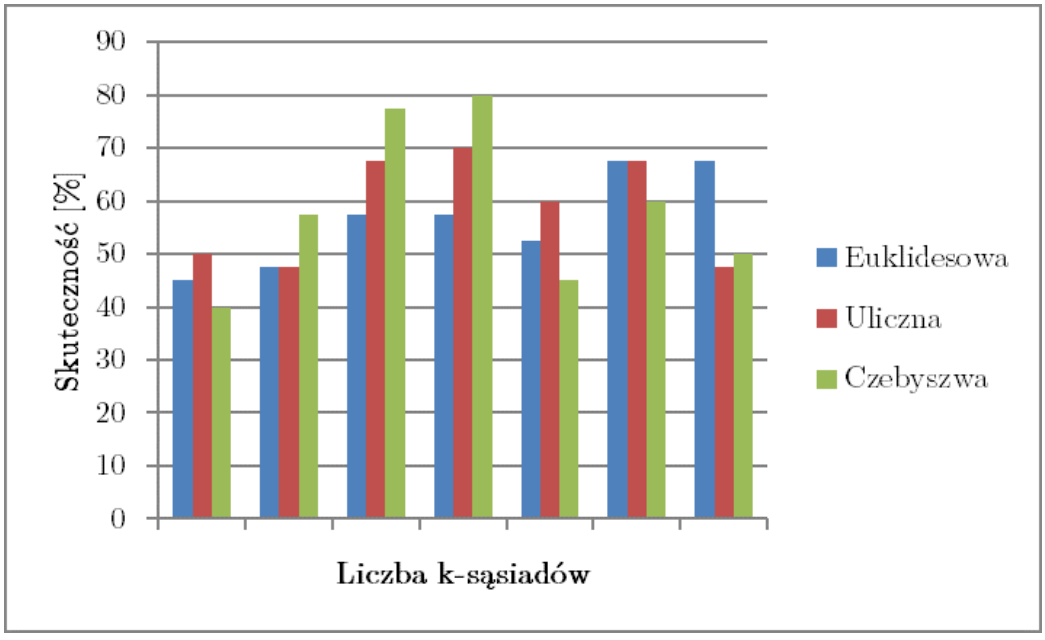


Rysunek 11. Dane z tabel 1-9 dla kategorii topics, zbiór treningowy 80%, zbiór testowy 20%

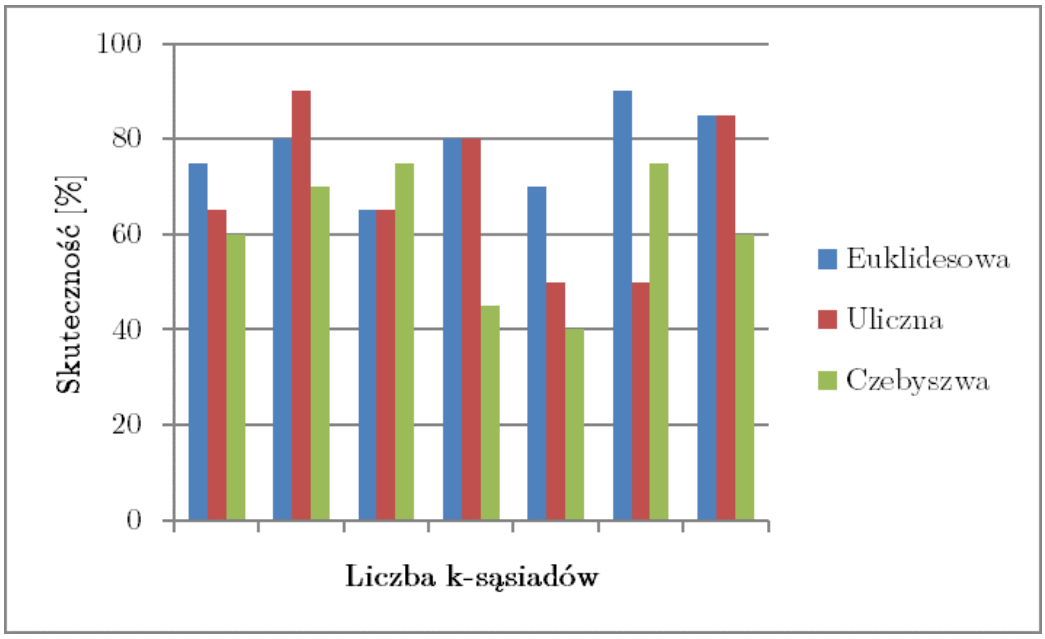
5.2.3. Medium



Rysunek 12. Dane z tabel 1-9 dla kategorii medium, zbiór treningowy 20%, zbiór testowy 80%



Rysunek 13. Dane z tabel 1-9 dla kategorii medium, zbiór treningowy 60%, zbiór testowy 40%



Rysunek 14. Dane z tabel 1-9 dla kategorii medium, zbiór treningowy 80%, zbiór testowy 20%

5.3. Inverse document frequency

5.3.1. Metryka Euklidesowa

k	places [%]	topics [%]	medium [%]
2	78,7	32,6	57,5
3	82,6	31,5	43,8
5	82,9	29,3	65
7	82,3	33,7	66,2
10	81,7	33,7	51,2
15	81,3	35,9	46,2
20	81	35,9	46,2

Tabela 10. Skuteczność klasyfikacji dla podziału: zbiór treningowy 20%, zbiór testowy 80%

k	places [%]	topics [%]	medium [%]
2	78,6	34,8	57,5
3	82,3	32,6	62,5
5	83,5	28,3	45
7	83,7	28,3	67,5
10	83,4	28,3	30
15	82,8	21,7	42,5
20	82,6	26,1	45

Tabela 11. Skuteczność klasyfikacji dla podziału: zbiór treningowy 60%, zbiór testowy 40%

k	places [%]	topics [%]	medium [%]
2	80,5	30,4	35
3	83,8	34,8	65
5	83,8	21,7	55
7	83,8	26,1	45
10	83,5	26,1	50
15	83,7	43,5	50
20	83,1	34,8	70

Tabela 12. Skuteczność klasyfikacji dla podziału: zbiór treningowy 80%, zbiór testowy 20%

5.3.2. Metryka uliczna

k	places [%]	topics [%]	medium [%]
2	77,8	31,5	50
3	81,2	34,8	63,7
5	81,7	33,7	52,5
7	81,3	34,8	45
10	81,2	35,9	57,5
15	81	35,9	47,5
20	81	35,9	47,5

Tabela 13. Skuteczność klasyfikacji dla podziału: zbiór treningowy 20%, zbiór testowy 80%

k	places [%]	topics [%]	medium [%]
2	78,8	19,6	60
3	81,8	19,6	52,5
5	82,6	19,6	65
7	82,8	17,4	67,5
10	82,2	19,6	50
15	81,9	19,6	70
20	81,5	26,1	75

Tabela 14. Skuteczność klasyfikacji dla podziału: zbiór treningowy 60%, zbiór testowy 40%

k	places [%]	topics [%]	medium [%]
2	79,2	26,1	50
3	83,3	30,4	50
5	83,4	26,1	45
7	83,3	21,7	60
10	82,9	26,1	65
15	82,8	39,1	45
20	82,2	26,1	70

Tabela 15. Skuteczność klasyfikacji dla podziału: zbiór treningowy 80%, zbiór testowy 20%

5.3.3. Metryka Czebyszewa

k	places [%]	topics [%]	medium [%]
2	77,6	31,5	62,5
3	80,7	31,5	51,2
5	81,5	40,2	58,8
7	81,2	33,7	70
10	81	34,8	47,5
15	80,5	35,9	57,5
20	79,6	35,9	50

Tabela 16. Skuteczność klasyfikacji dla podziału: zbiór treningowy 20%, zbiór testowy 80%

k	**places** [%]	**topics** [%]	**medium** [%]
2	79,8	41,3	47,5
3	81,8	63	42,5
5	81,6	28,3	62,5
7	81,7	39,1	70
10	81,6	34,8	65
15	81,6	23,9	45
20	81,5	23,9	52,5

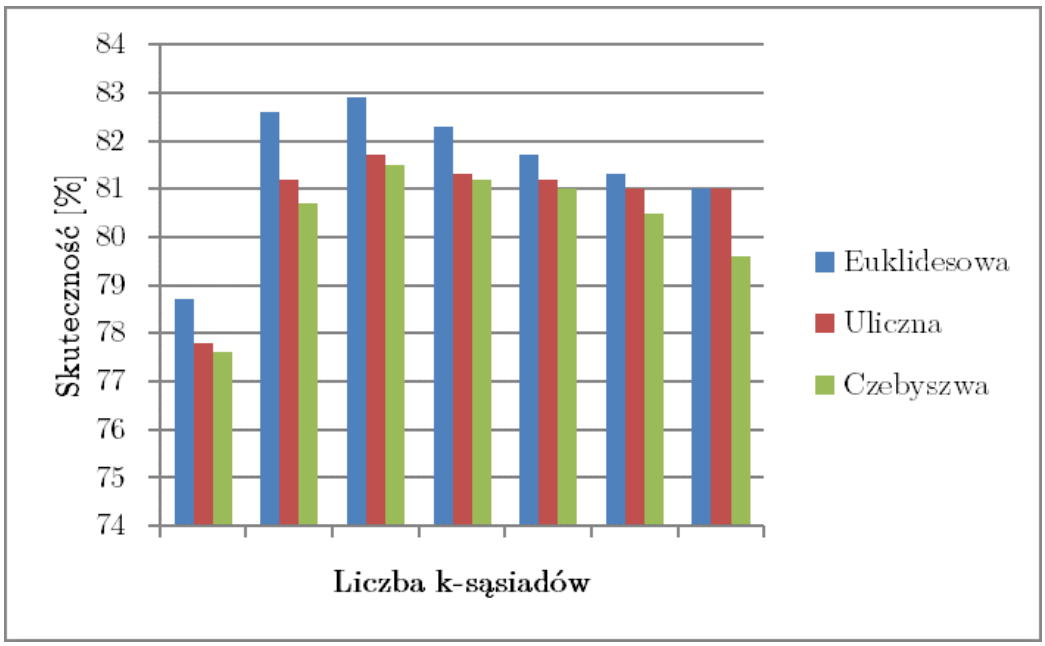
Tabela 17. Skuteczność klasyfikacji dla podziału: zbiór treningowy 60%, zbiór testowy 40%

k	**places** [%]	**topics** [%]	**medium** [%]
2	79,1	43,5	80
3	82,4	39,1	60
5	81,6	26,1	40
7	81,7	30,4	50
10	81,9	26,1	55
15	82,2	26,1	60
20	82,1	26,1	65

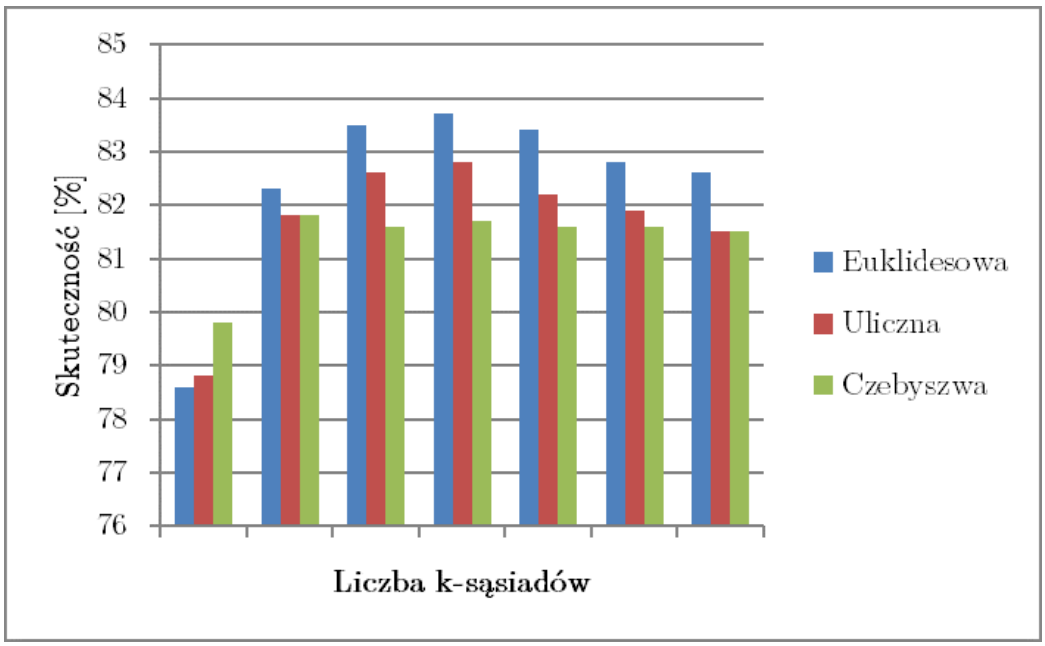
Tabela 18. Skuteczność klasyfikacji dla podziału: zbiór treningowy 80%, zbiór testowy 20%

5.4. Inverse document frequency - podsumowanie wykresami

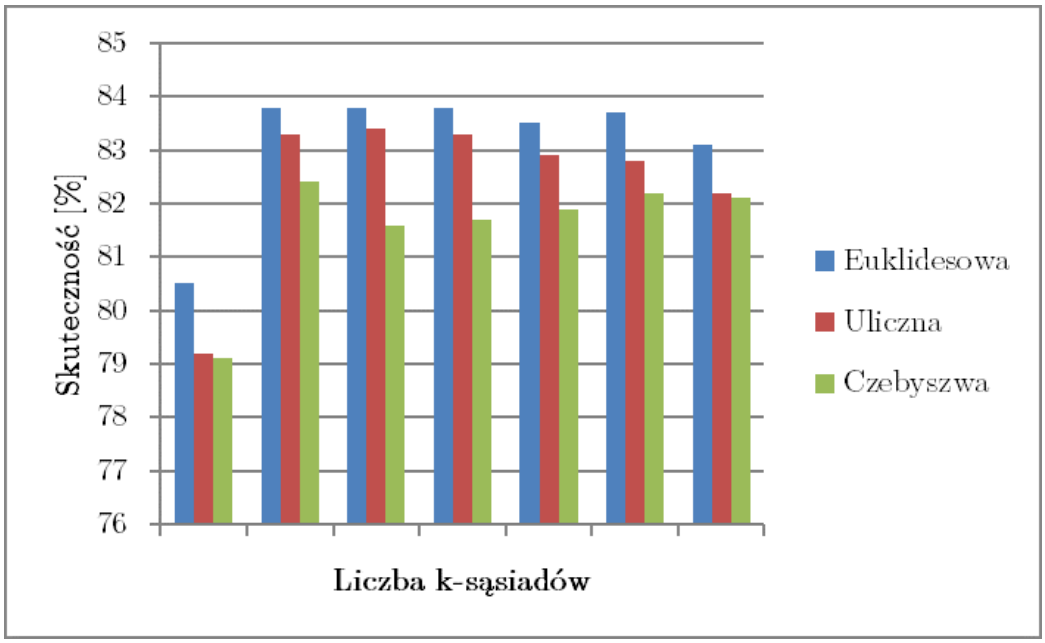
5.4.1. Places



Rysunek 15. Dane z tabel 1-9 dla kategorii places, zbiór treningowy 20%, zbiór testowy 80%

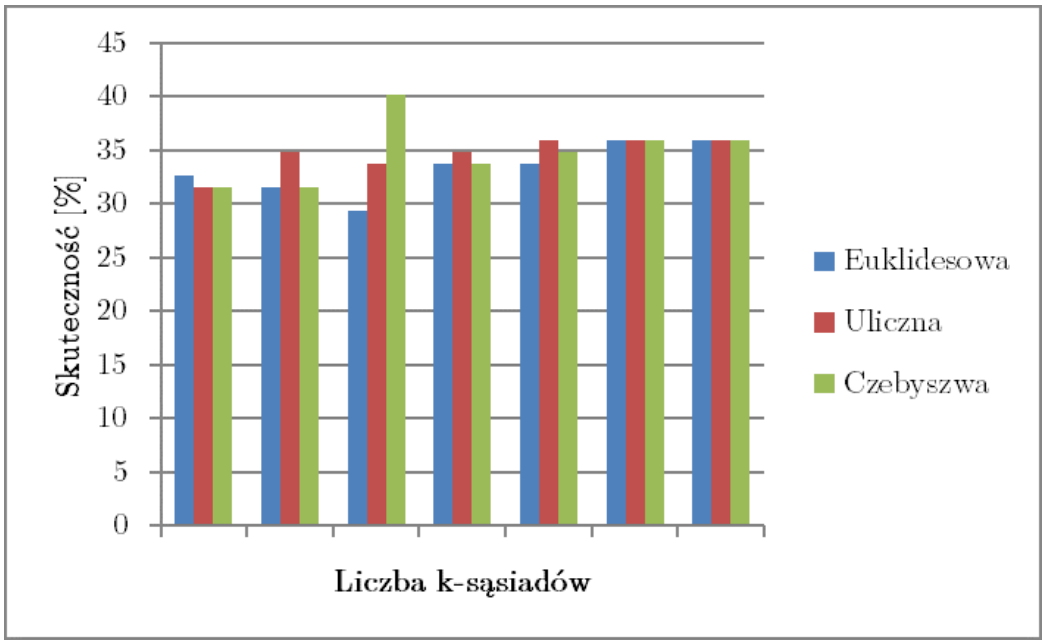


Rysunek 16. Dane z tabel 1-9 dla kategorii places, zbiór treningowy 60%, zbiór testowy 40%

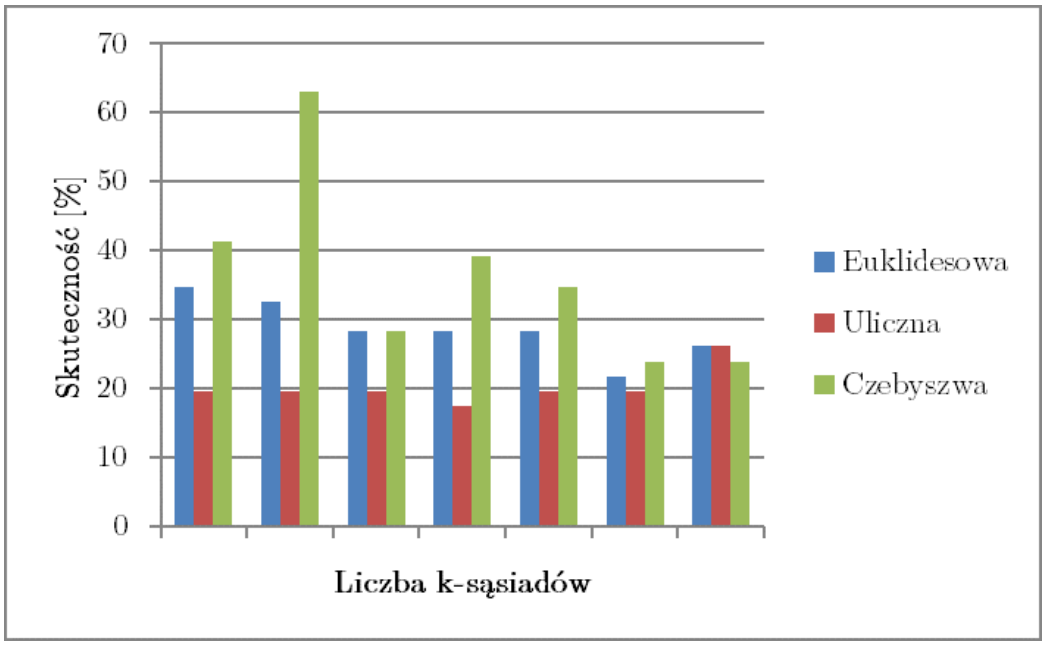


Rysunek 17. Dane z tabel 1-9 dla kategorii places, zbiór treningowy 80%, zbiór testowy 20%

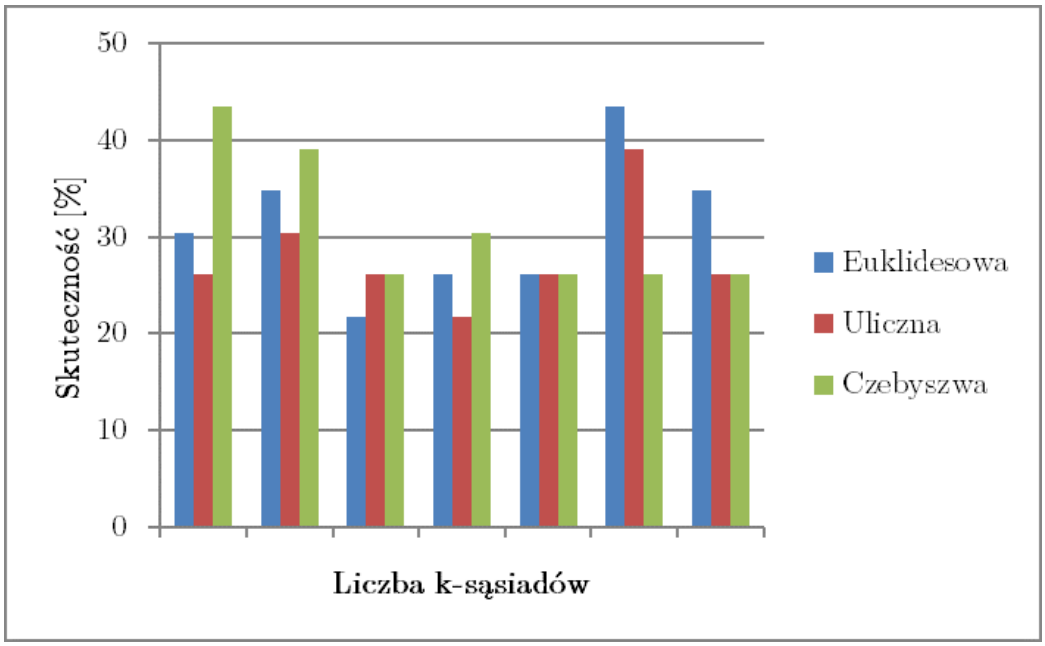
5.4.2. Topics



Rysunek 18. Dane z tabel 1-9 dla kategorii topics, zbiór treningowy 20%, zbiór testowy 80%

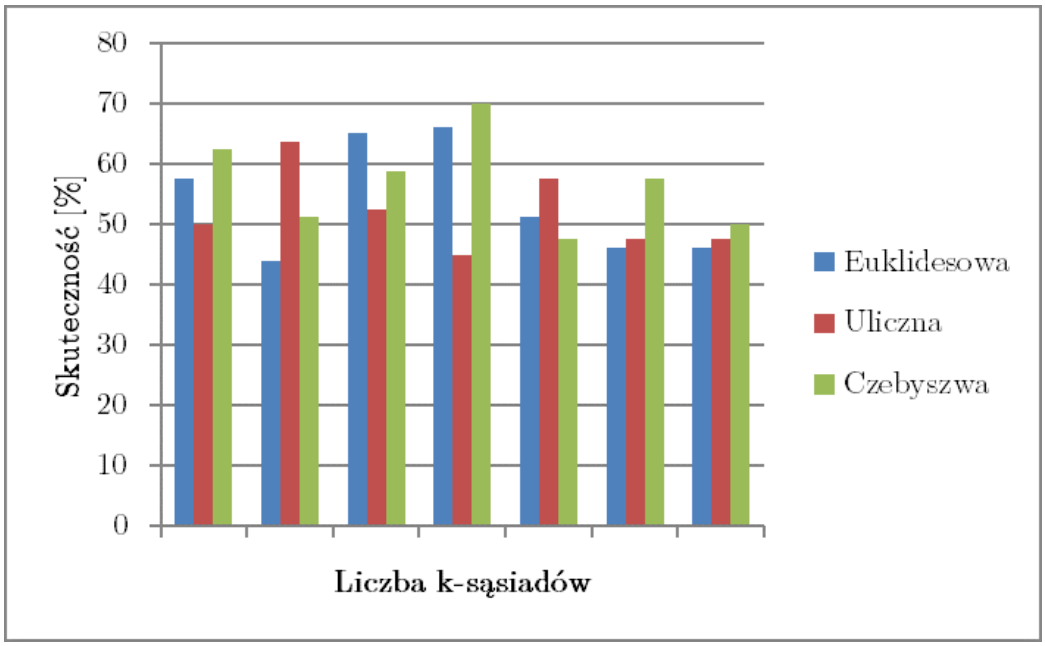


Rysunek 19. Dane z tabel 1-9 dla kategorii topics, zbiór treningowy 60%, zbiór testowy 40%

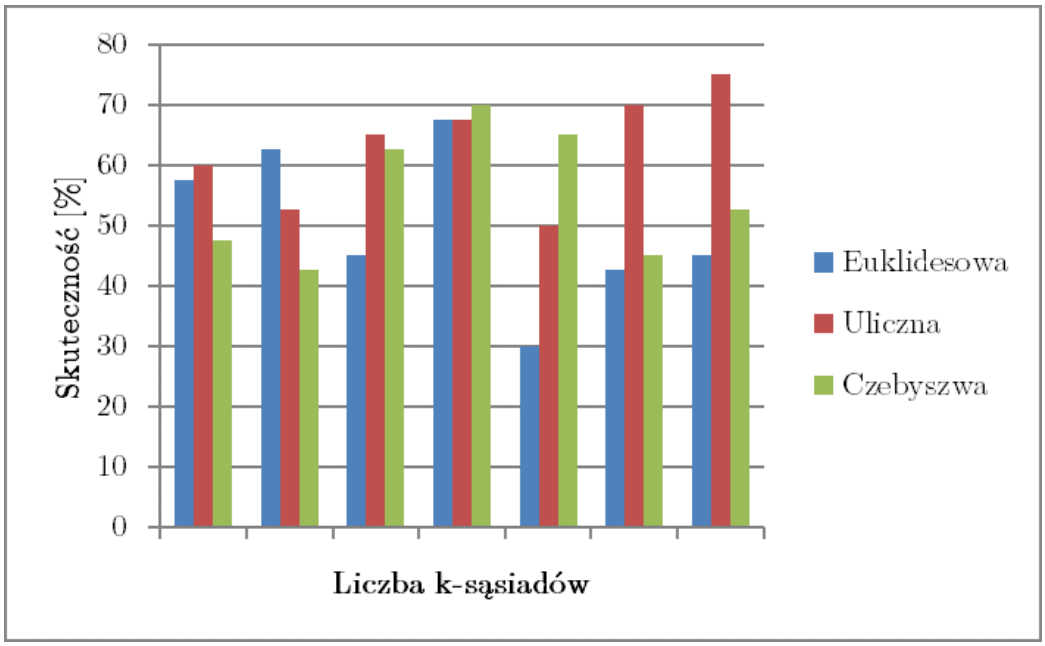


Rysunek 20. Dane z tabel 1-9 dla kategorii topics, zbiór treningowy 80%, zbiór testowy 20%

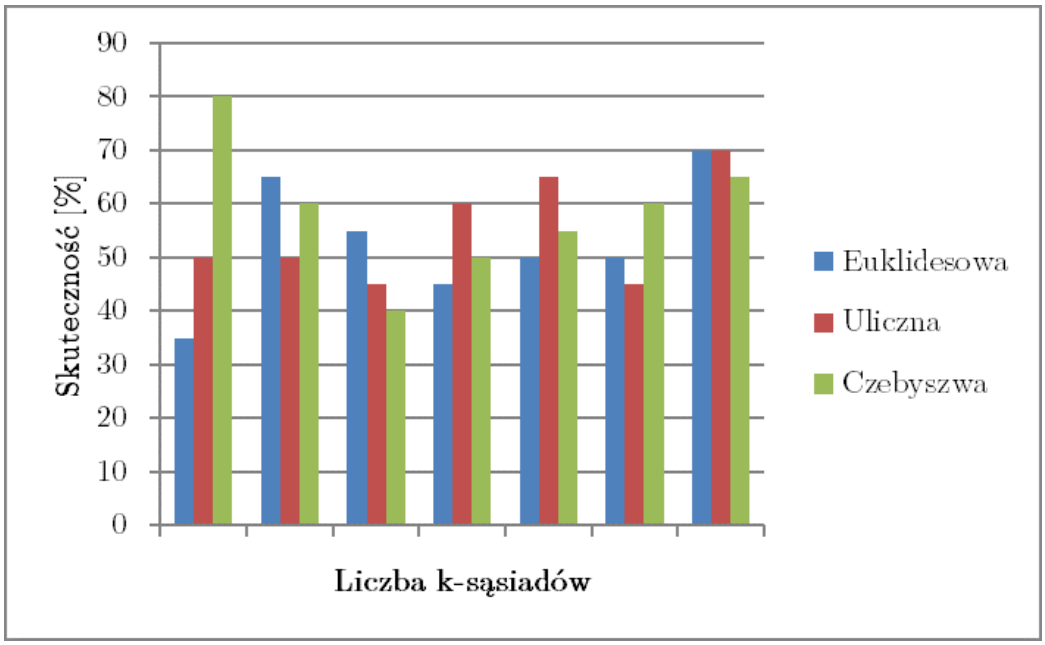
5.4.3. Medium



Rysunek 21. Dane z tabel 1-9 dla kategorii medium, zbiór treningowy 20%, zbiór testowy 80%



Rysunek 22. Dane z tabel 1-9 dla kategorii medium, zbiór treningowy 60%, zbiór testowy 40%



Rysunek 23. Dane z tabel 1-9 dla kategorii medium, zbiór treningowy 80%, zbiór testowy 20%

5.5. Najlepsze wyniki

k	data	metryka	skuteczność [%]	metoda
3	medium	uliczna	90	TF
7	places	euklidesowa	87	TF
3	places	euklidesowa	83,8	IDF
5	places	Czebyszewa	83,8	IDF
5	places	uliczna	83,4	IDF

Tabela 19. Największa otrzymana przez nas skuteczność

6. Dyskusja

Jako cel przeprowadzonych eksperymentów obraliśmy ustalenie jak dobrane parametry naszego algorytmu będą wpływały na skuteczność. Rozważania należy zacząć od wyboru metody ekstrakcji cech. Według naszych spostrzeżeń obie metody sprawdzały się równie dobrze, a drobne różnica w otrzymanej skuteczności wynikały z doboru stosunku zbioru treningowego do zbioru testowego. Wszystkie przetestowane przez nas metryki wykazały się wymierną poprawnością, jednakże fluktuacje wyników między poszczególnymi iteracjami wynikały bezpośrednio z doboru liczby sąsiadów (K) i danego zbioru danych na którym odbywał się eksperyment.

Dla tagu 'place' bezapelacyjnym zwycięzcą okazała się metryka Euklidesowa zawsze osiągająca wyższe rezultaty niż inne metryki. Jej dominacja

była niezależne od doboru liczby sąsiadów. Niedotrenowanie lub przetrenowanie klasyfikatora skutkowało w mniejszej skuteczności. Najlepsze wyniki otrzymaliśmy dla K oscylującego wokół 5.

Dla tagu 'topic' wyniki okazały się być zróżnicowane i niezależne od jakiekolwiek zaimplementowanej przez nas metryki. Tutaj największy wpływ okazała się mieć stosunek zbioru treningowego do testowego. Zbyt mały zbiór treningowy (20%) powoduje spadek skuteczności dopasowania dla wszystkich metryk. Dla liczby sąsiadów większej niż 2 wyniki oscylowały wokół podobnych wartości.

Dla tagu 'medium', który został zaczerpnięty z przygotowanego przez nas zbioru artykułów, średnie najlepsze wyniki (powyżej 60% skuteczności) otrzymywaliśmy z zbiorem treningowym na poziomie 60%. Podczas eksperymentu najlepiej poradziła sobie metryka Uliczna, jednakże pozostałe metryki wykazywały się różną skutecznością, zależną od doboru liczby sąsiadów. Widoczna byłą rosnąca skuteczność, dla co raz to większego K.

7. Wnioski

— Metoda K najbliższych sąsiadów wskazuje dużą (ponad 80%) skuteczność przy klasyfikacji tekstów.
— Dla dużego zbioru danych (7000 artykułów), liczba sąsiadów oscylująca wokół 5 zawsze dawała lepsze rezultaty.
— Największą skuteczność otrzymaliśmy dla zbioru treningowego na poziomie 60%. Zbiory 20% i 80% okazały się niezadawalające.
— Metryka jako parametr naszego algorytmu miała znaczący mały wpływ na skuteczność. Wpływ ten był jednak zauważalny i metryka Euklidesowa najczęściej okazywała się zwycięzcą.

Literatura

[1] Methods for the linguistic summarization of data - aplications of fuzzy sets and their extensions, Adam Niewiadomski, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2008
[2] https://github.com/hklemp/dotnet-stop-words
[3] http://snowball.tartarus.org/algorithms/english/stemmer.html
[4] https://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection
[5] https://github.com/jurczewski/KSR/blob/master/Zad1/Zad1/Data/ours.sgm