Informatyka, studia dzienne, I st. semestr VI

Komputerowe systemy rozpoznawania **2019/2020**

Prowadzący: Marcin Kacprowicz wtorek, 14:00

Data oddania: _____ Ocena: _____

Sebastian Kaźmierski 216795
Bartosz Paluszkiewicz 216856

Zadanie 1: Ekstrakcja cech, miary podobieństwa, klasyfikacja

1. Cel

Celem zadania jest stworzenie aplikacji która wykorzystując metodę Knn będzie wykonywała klasyfikację artykułów w dwóch kategoriach (Places,Topics). Aplikacja ma również implementować moduł ekstrakcji cech który będzie obliczał wartości cech klasyfikowanych artykułów. Po stworzeniu aplikacji przy jej wykorzystaniu zostaną przeprowadzone eksperymenty z różnymi wartościami parametrów metody Knn.

2. Wprowadzenie

2.1. Klasyfikator

Zaimplementowany klasyfikator zbioru tekstów wykorzystuje algorytm k najbliższych sąsiadów (k-nn, k nearest keighbours), który jest jednym z bezparametrowych metod klasyfikacji statystycznej.
W trakcie działania nie tworzy wewnętrznej reprezentacji danych uczących, lecz przechowuje wszystkie wzore uczące i dopiero w momencie pojawienia się wzorca testującego szuka rozwiązania, wyznaczając odległość do wszystkich wzorców.
Bezparametrowość algorytmu objawia się brakiem założeń co do rozkładu podstawowych danych, np. rozkładu jednostajnego.

2.2. Wektor cech

Do ekstrakcji danych wykorzystaliśmy wektor następujących cech:

1. Liczba wszystkich słów kluczowych w pierwszych 10% procentach tekstu
2. Liczba wszystkich słów kluczowych w całym tekście
3. Liczba wszystkich nazw własnych w stosunku do liczby słów w tekście (po stop liście)
4. Średnia długość unikalnych nazw własnych
5. Długość tekstu (liczba wszystkich słów po stop liście)
6. Liczba słów które występują więcej niż raz w stosunku do długości tekstu (po stop liście i po stemizacji)
7. Średnia długość zdań (liczba słów) (po stop liście i po stemizacji)
8. Liczba unikalnych słów w stosunku do długości tekstu (po stop liście)
9. Liczba akapitów w stosunku do długości tekstu (po stop liście)
10. Średnia długość akapitu (liczba słów) (przed stop listą)
11. Liczba wszystkich słów usuniętych przez stop listę w stosunku do długości tekstu po stop liście
12. Liczba słów kluczowych dla WEST_GERMANY
13. Liczba słów kluczowych dla USA
14. Liczba słów kluczowych dla FRANCE
15. Liczba słów kluczowych dla UK
16. Liczba słów kluczowych dla CANADA
17. Liczba słów kluczowych dla JAPAN

* nazwa własna – słowo pisane wielką literą które nie jest na początku zdania lub jest na początku zdania, ale przynajmniej raz występuje w środku
* długość tekstu – liczba wszystkich słów w tekście

2.3. Zastosowane metryki

Do obliczania odległości zastosowaliśmy trzy metryki (2.3.1 - 2.3.3), miarę własną (2.3.6) oraz dwie miary prawdopodobieństwa (2.3.4, 2.3.5). Ze względu na to, że miary podobieństwa służą określaniu "bliskości" wektorów, a nie odległości między nimi, do obliczeń musieliśmy wykorzystać odwrotności uzyskanych wartości.

2.3.1. Metryka Euklidesa

Odległość euklidesową określamy pierwiastek sumy kwadratów różnic wartości cech wektorów A, B i opisujemy wzorem:

$$d(A,B) = \sqrt{\sum_{i=1}^{n}((x_{iA} - x_{iB})^2)} \tag{1}$$

2.3.2. Metryka Czebyszewa

Odległością Czebyszewa określamy największy moduł różnic między wartościami cechami wektorów A, B i określamy wzorem:

$$d(A,B) = \max_i |x_{iA} - x_{iB}| \tag{2}$$

2.3.3. Metryka Uliczna

Odległością w metryce ulicznej (odległością taksówkową, uliczną, Manhattan) nazywamy sumę modułów różnic wartości w każdym wymiarze wektorów A, B i określamy wzorem:

$$d(A,B) = \sum_{i=1}^{n} |x_{iA} - x_{iB}| \tag{3}$$

2.3.4. Minimum - maksimum

Odległość między wketorami A, B określamy jako odwrotność stosunku sumy minimów wartości cech do sumy maksimów wartości cech, czyli stosunek maksimów do minimów wartości cech wektorów.

$$d(A,B) = (\frac{\sum_{i=1}^{n} min(x_{iA}, x_{iB})}{\sum_{i=1}^{n} max(x_{iA}, x_{iB})})^{-}1 = \frac{\sum_{i=1}^{n} max(x_{iA}, x_{iB})}{\sum_{i=1}^{n} min(x_{iA}, x_{iB})} \tag{4}$$

2.3.5. Średnia arytmetyczna - minimum

Odległość między wektorami A, B określamy jako odwrotność stosunku sumy minimalnych wartości cech wektorów i średniej arytmetycznej wartości cech wektorów, czyli stosunek średniej arytmetycznej do sumy minimów wartości cech wektorów.

$$d(A,B) = (\frac{\sum_{i=1}^{n} min(x_{iA}, x_{iB})}{\frac{1}{2}\sum_{i=1}^{n}(x_{iA} + x_{iB})})^{-}1 = \frac{\frac{1}{2}\sum_{i=1}^{n}(x_{iA} + x_{iB})}{\sum_{i=1}^{n} min(x_{iA}, x_{iB})} \tag{5}$$

2.3.6. Miara własna

Odległość obliczoną przy użyciu autorskiej miary określamy jako sumę iloczynów ułożonych rosnąco różnic wartości między cechami wektora i współczynnika zależnego od pozycji tej odległości w posortowanym zbiorze, a także wielkości zbioru.

$$d(A,B) = \sum_{i=0}^{N-1} C_i(1 - \frac{i}{2N}) \tag{6}$$

$$C(A,B) = (\sum_{i=1}^{N} |x_A - x_{iB}|) \uparrow \tag{7}$$

2.4. Miary jakości

2.4.1. Accuracy - dokładność

Dokładnością nazywamy stosunek liczby poprawnych prognoz do wszystkich wykonanych prognoz,

$$accuracy = \frac{correct\ predictions}{total\ predictions} \tag{8}$$

W celu obliczenia kolejnych miar tworzymy macierz błędu, której uproszczoną formą jest poniższa tabela [2]:

	Positive prediction	Negative prediction
Positive class	True Positive (TP)	False Negative (FN)
Negative class	False Positive (FP)	True Negative (TN)

W naszym przypadku musimy zakwalifikować 6, a nie 2 klasy, więc macierz błędu ma w rzeczywistości wymiary 6 x 6.

2.4.2. Precision - precyzja

Precyzja określa stosunek poprawnie zakwalifikowanych obiektów danej klasy do wszystkich prognoz, które wskazują na tę klasę (suma słusznie i niesłusznie zakwalifikowanych obiektów)

$$precision = \frac{\sum_c^C TruePositive_c}{\sum_c^C (TruePositive_c + FalsePositive_c)} \tag{9}$$

2.4.3. Recall - skuteczność

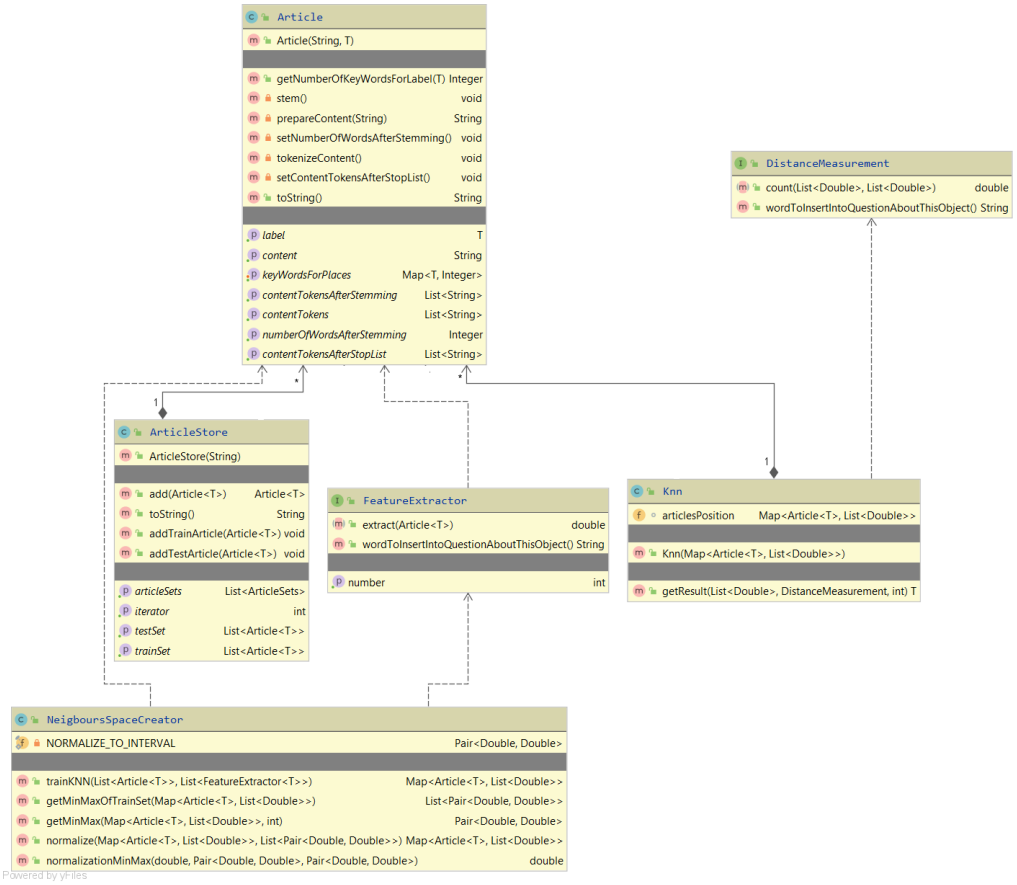
Skuteczność określa stosunek poprawnie zakwalifikowanych obiektów danej klasy do sumy poprawnie zakwalifikowanych i poprawnie odrzuconych obiektów

$$recall = \frac{\sum_c^C TruePositive_c}{\sum_c^C (TruePositive_c + FalseNegative_c)} \tag{10}$$

3. Opis implementacji

Podstawową klasą jest klasa „Article" która przechowuje w sobie treść artykułu oraz informację o etykiecie którą ten artykuł jest oznaczony. Wszystkie artykuły są przechowywane w obiekcie klasy „ArticleStore" która odpowiada za podział artykułów na zbiór uczący i testowy. Podczas tworzenia programu został wykorzystany wzorzec projektowy strategii a świadczą o tym interfejsy „FeatureExtractors" implementowany przez klasy które są odpowiedzialne za obliczanie wartości cech tekstu oraz „DistanceMeasurement" implementowany przez klasy które reprezentują poszczególne miary/metryki. Klasa „NeigboursSpaceCreator" jest odpowiedzialna za przygotowanie zbioru uczącego czyli przechowywanie wartości cech dla artykułów ze zbioru uczącego. Klasa „Knn" na podstawie k najbliższych sąsiadów przydziela elementowi ze zbioru uczącego odpowiednią etykietę. Program został napisany w języku Java w wersji 13.

Rysunek 1. Diagram UML



4. Materiały i metody

Badania zostały przeprowadzone na zbiorze artykułów w języku angielskim które były oznaczone etykietami w dwóch kategoriach w kategorii „places” (west-germany, usa, france,uk ,canada, japan) oraz w kategorii „topics” (earn, acq)

4.1. Wpływ liczby sąsiadów na jakość klasyfikacji

Badanie zostało przeprowadzone z różna ilością sąsiadów 1,3,4,7,11,12,13,19,37,59
Pozostałe parametry były stałe:
Metryka: Metryka Euklidesa
Podział danych: 70% - zbiór uczący (Places), 50% - zbiór uczący (Topics)
Cechy: 1,2,3,4,5,6,7,8,9,10,11

4.2. Wpływ podziału danych na jakość klasyfikacji

Badanie zostało przeprowadzone z różną wielkość zbioru uczącego 30%,40%,50%,60%,70%
Pozostałe parametry były stałe:
Metryka: Metryka Euklidesa
Liczba sąsiadów: 37
Cechy: 1,2,3,4,5,6,7,8,9,10,11

4.3. Wpływ metryki/miary na jakość klasyfikacji

Badanie zostało przeprowadzone z różnymi metrykami/miarami: Średnia arytmetyczna minimum, Metryka Czebyszewa, Metryka Euklidesa, Minimum-maximum, Metryka Uliczna, Nasza Metryka Pozostałe parametry były stałe:
Podział danych: 60% - zbiór uczący (Places), 50% - zbiór uczący (Topics)
Liczba sąsiadów: 37
Cechy: 1,2,3,4,5,6,7,8,9,10,11

4.4. Wpływ cech na jakość klasyfikacji

Badania zostały przeprowadzone w taki sposób, że od podstawowego zbioru cech czyli cech: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 odejmowaliśmy jedną cechę (lub dodawaliśmy w przypadku cech związanych z etykietami) i najbardziej interesujące przypadki zamieściliśmy w wynikach. Pozostałe parametry były stałe:
Metryka: Metryka Euklidesa (Topics), Metryka Uliczna (Places)
Liczba sąsiadów: 37
Podział danych: 60%

4.5. Prównanie naszej metryki i metryki Czebyszewa

Badanie zostało przeprowadzone z różna ilością sąsiadów 1,3,4,7,11,12,13,19,37,59 i przy użyci Naszej Metryki i Metryki Czebyszewa Pozostałe parametry były stałe:
Podział danych: 60%
Cechy: 1,2,3,4,5,6,7,8,9,10,11

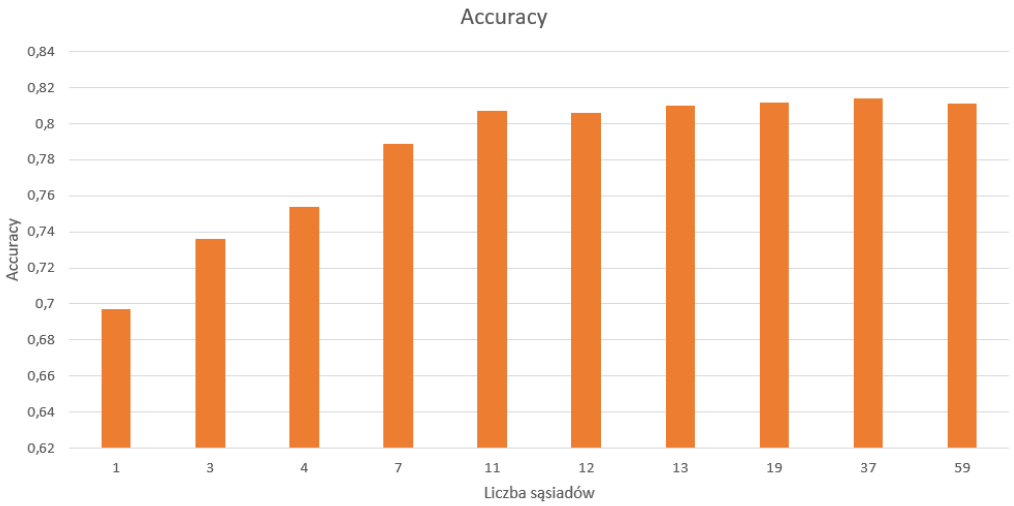
5. Wyniki

5.1. Wyniki klasyfikacji metodą k-NN dla 10 różnych wartości parametru k (Places)

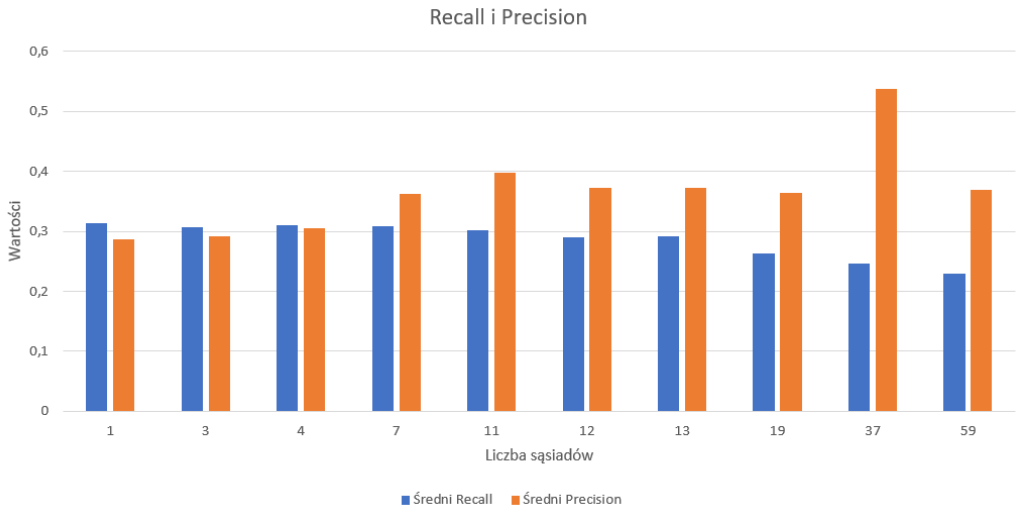
Tabela 1. Wyniki klasyfikacji dla różnej liczby sąsiadów

Liczba sąsiadów	Accuracy	Średni Recall	Średni Precision
1	0,697	0,314	0,287
3	0,736	0,307	0,292
4	0,754	0,311	0,305
7	0,789	0,309	0,363
11	0,807	0,301	0,398
12	0,806	0,290	0,372
13	0,81	0,291	0,373
19	0,812	0,263	0,364
37	0,814	0,246	0,538
59	0,811	0,229	0,369

Rysunek 2. Accuracy w zależności od liczby sąsiadów



Rysunek 3. Recall i Precision w zależności od liczby sąsiadów

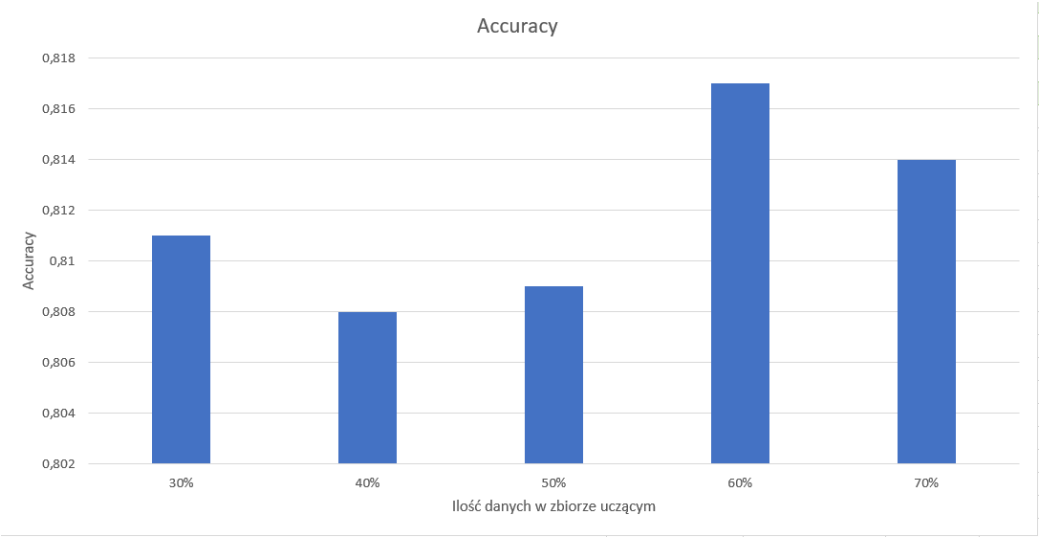


5.2. Wyniki klasyfikacji metodą k-NN dla 5 różnych podziałów na zbiór uczący i testowy (Places)

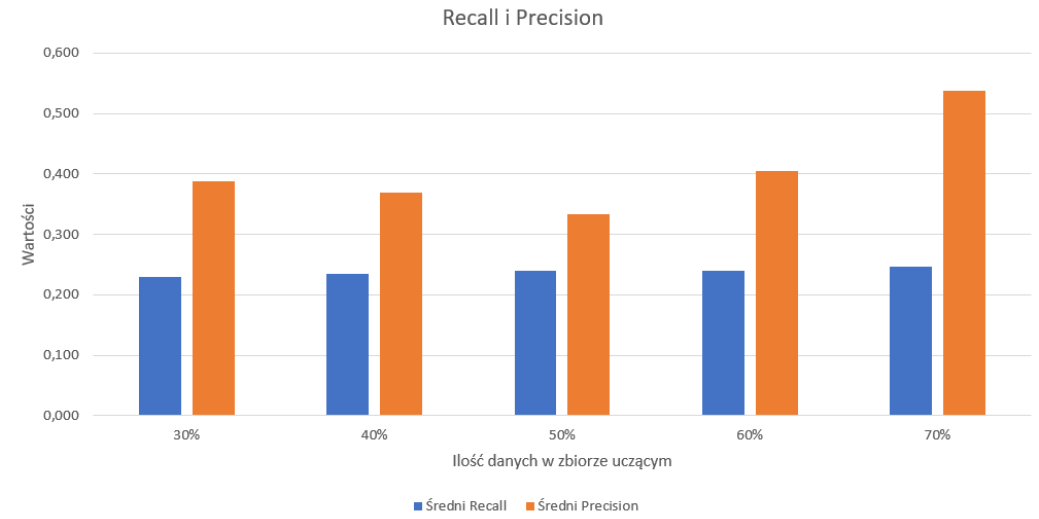
Tabela 2. Wyniki klasyfikacji dla różnych podziałów na zbiór uczący i testowy

Ilość danych w zbiorze uczący	**Accuracy**	**Średni Recall**	**Średni Precision**
30%	0,811	0,229	0,387
40%	0,808	0,235	0,369
50%	0,809	0,240	0,334
60%	0,817	0,239	0,404
70%	0,814	0,246	0,538

Rysunek 4. Accuracy w zależności od podziału na zbiór uczący i testowy



Rysunek 5. Recall i Precision w zależności od podziału na zbiór uczący i testowy

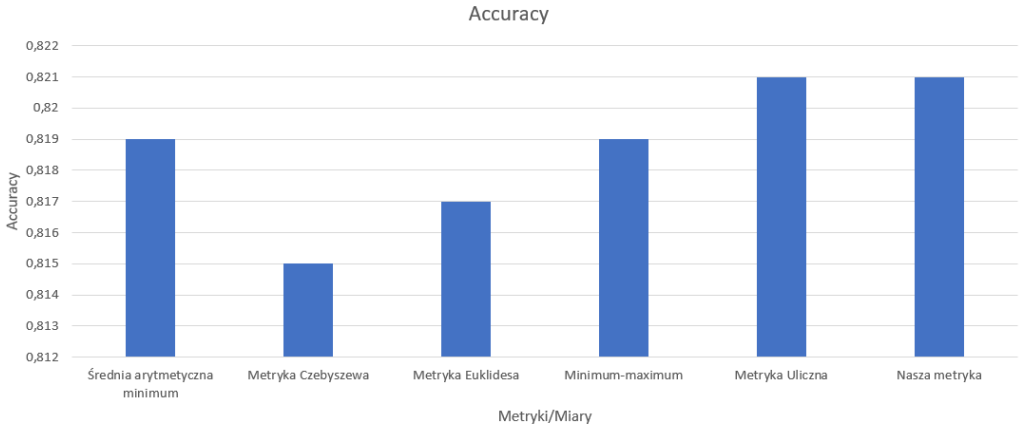


5.3. Wyniki klasyfikacji metodą k-NN dla 6 różnych metryk/miar (Places)

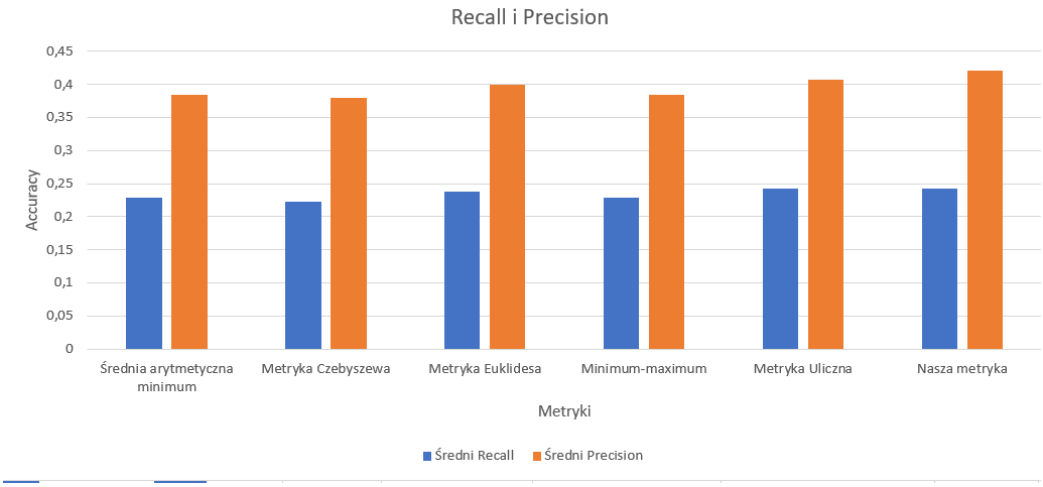
Tabela 3. Wyniki klasyfikacji dla różnych metryk/miar

Metryka	Accuracy	Średni Recall	Średni Precision
Średnia arytmetyczna minimum	0,819	0,229	0,384
Metryka Czebyszewa	0,815	0,223	0,379
Metryka Euklidesa	0,817	0,238	0,399
Minimum-maximum	0,819	0,229	0,384
Metryka Uliczna	0,821	0,242	0,407
Nasza metryka	0,821	0,242	0,42

Rysunek 6. Accuracy w zależności od miary/metryki



Rysunek 7. Recall i Precision w zależności od miary/metryki



5.4. Wyniki klasyfikacji metodą k-NN dla różnych cech (Places)

Podstawowe cechy to 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11

Tabela 4. Wyniki klasyfikacji dla 4 podzbiorów cech

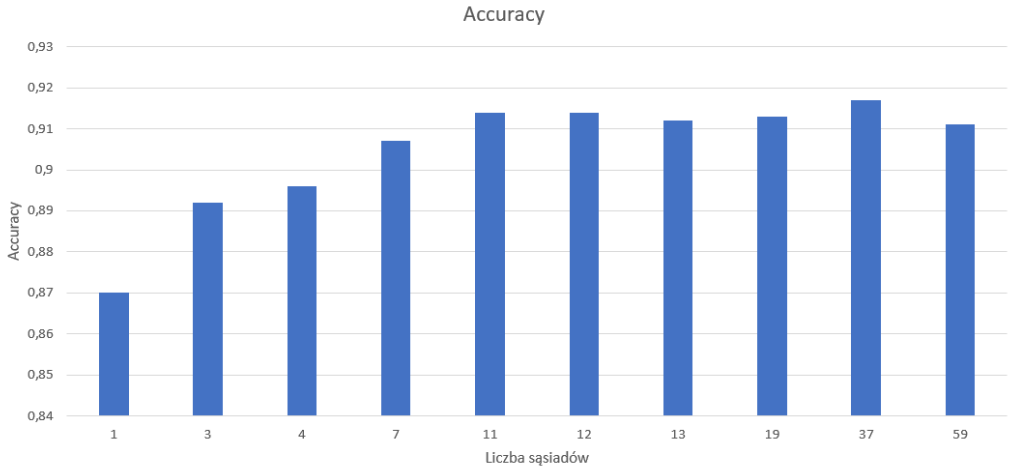
Różnice pomiędzy cechami podstawowymi	**Accuracy**	**Średni Recall**	**Średni Precision**
Cechy podstawowe	0,817	0,238	0,399
Dodatkowe cechy: 12 13 14 15 16 17	0,875	0,492	0,892
Brak cechy 8	0,805	0,179	0,25
Brak cechy 6	0,811	0,22	0,351
Brak cechy 9	0,819	0,241	0,417

Tabela 5. Wyniki klasyfikacji dla różnej liczby sąsiadów (Topics)

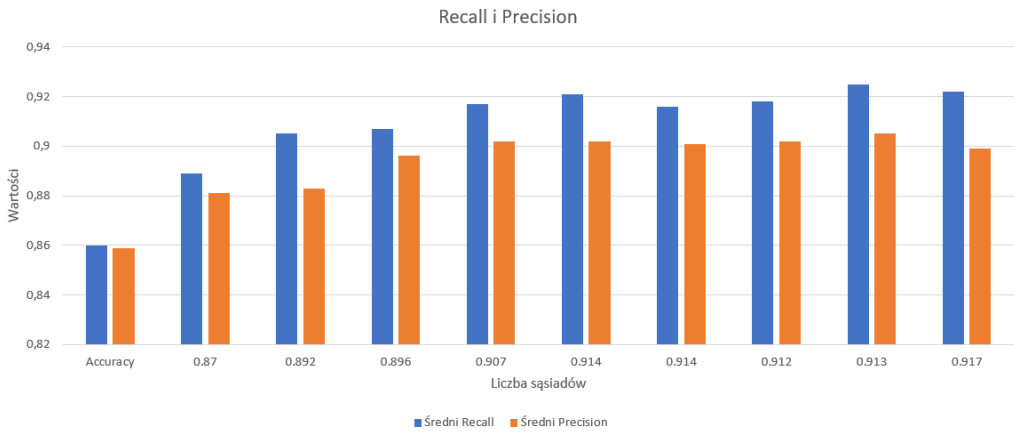
Liczba sąsiadów	Accuracy	Średni Recall	Średni Precision
1	0,87	0,86	0,859
3	0,892	0,889	0,881
4	0,896	0,905	0,883
7	0,907	0,907	0,896
11	0,914	0,917	0,902
12	0,914	0,921	0,902
13	0,912	0,916	0,901
19	0,913	0,918	0,902
37	0,917	0,925	0,905
59	0,911	0,922	0,899

5.5. Wyniki klasyfikacji metodą k-NN dla 10 różnych wartości parametru k

Rysunek 8. Accuracy w zależności od liczby sąsiadów



Rysunek 9. Recall i Precision w zależności od liczby sąsiadów

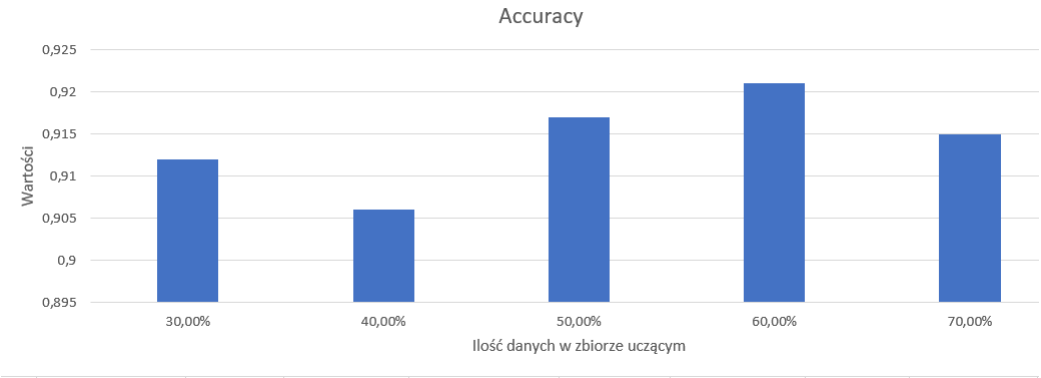


5.6. Wyniki klasyfikacji metodą k-NN dla 5 różnych podziałów na zbiór uczący i testowy (Topics)

Tabela 6. Wyniki klasyfikacji dla różnych podziałów na zbiór uczący i testowy

Podział danych	Accuracy	Średni Recall	Średni Precision
30,00%	0,912	0,921	0,901
40,00%	0,906	0,921	0,893
50,00%	0,917	0,925	0,905
60,00%	0,921	0,926	0,909
70,00%	0,915	0,925	0,902

Rysunek 10. Accuracy w zależności od podziału na zbiór uczący i testowy



Rysunek 11. Recall i Precision w zależności od podziału na zbiór uczący i testowy

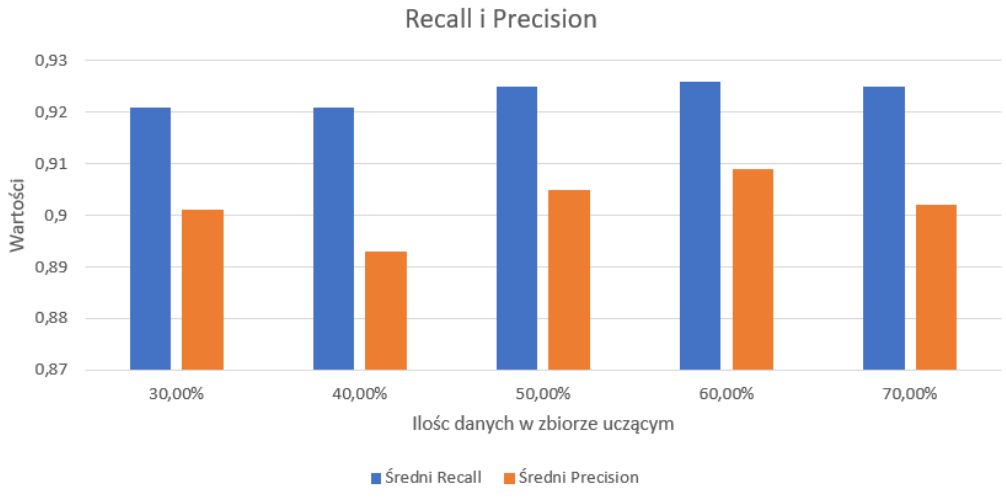
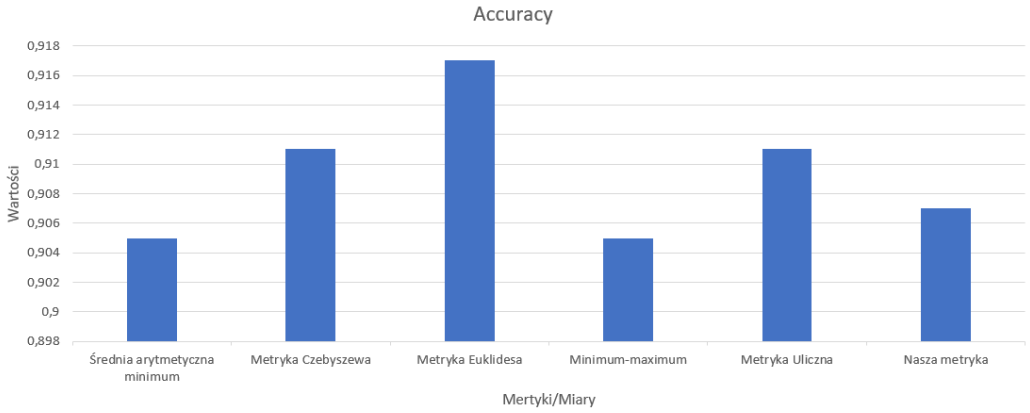


Tabela 7. Wyniki klasyfikacji dla różnych metryk/miar

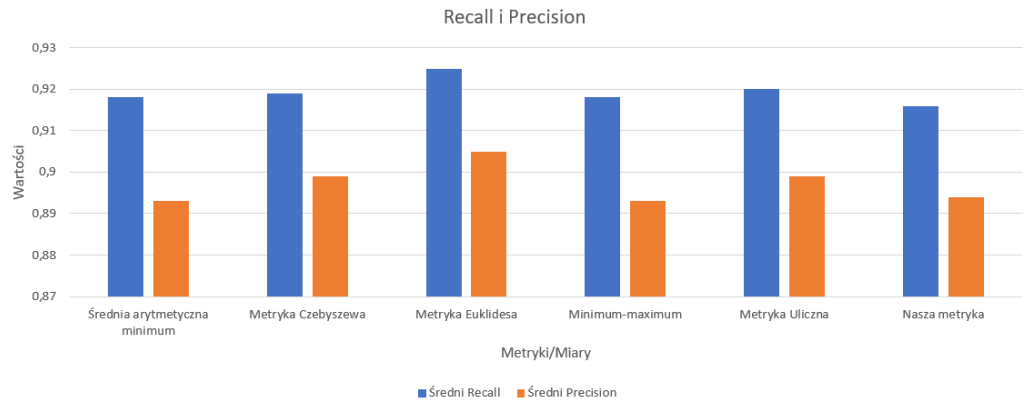
Metryka	Accuracy	Średni Recall	Średni Precision
Średnia arytmetyczna minimum	0,905	0,918	0,893
Metryka Czebyszewa	0,911	0,919	0,899
Metryka Euklidesa	0,917	0,925	0,905
Minimum-maximum	0,905	0,918	0,893
Metryka Uliczna	0,911	0,92	0,899
Nasza metryka	0,907	0,916	0,894

5.7. Wyniki klasyfikacji metodą k-NN dla 6 różnych metryk/miar (Topics)

Rysunek 12. Accuracy w zależności od miary/metryki



Rysunek 13. Recall i Precision w zależności od miary/metryki



5.8. Wyniki klasyfikacji metodą k-NN dla różnych cech (Topics)

Podstawowe cechy to 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11

Tabela 8. Wyniki klasyfikacji dla 4 podzbiorów cech

Różnice pomiędzy cechami podstawowymi	**Accuracy**	**Średni Recall**	**Średni Precision**
Cechy podstawowe	0,921	0,926	0,909
Dodatkowe cechy: 12 13	0,938	0,944	0,926
Brak cechy 11	0,894	0,899	0,88
Brak cechy 3	0,871	0,888	0,861
Brak cechy 10	0,925	0,928	0,913

5.9. Porównanie naszej metryki z metryką Czebyszewa (Places)

Tabela 9. Wyniki klasyfikacji dla 4 podzbiorów cech

Liczba sąsiadów	**Accuracy Nasza Metryka**	**Accuracy Metryka Czebyszewa**
1	0,701	0,694
3	0,751	0,735
4	0,763	0,749
7	0,802	0,79
11	0,816	0,803
12	0,816	0,804
13	0,816	0,806
19	0,819	0,813
37	0,821	0,815
59	0,817	0,814

6. Dyskusja

Liczba sąsiadów

W poniższej dyskusji odwołujemy się do Rysunków 2 i 8 oraz Tabel 1 i 5

Po wyznaczeniu klasyfikacji metodą Knn dla różnej liczby sąsiadów otrzymaliśmy podobne wyniki w przypadku klasyfikacji w kategorii „places” (6 etykiet) i „topic” (2 etykiety). W obu przypadkach można zauważyć znaczy wzrost dokładności klasyfikacji przy zwiększaniu liczby sąsiadów od 1 do 11. Wynika z tego, że zbyt mała liczba sąsiadów znacznie obniża accuracy klasyfikacji. Powyżej 11 sąsiadów accuracy w przypadku kategorii „topic” się stabilizuje i waha wokół wartości osiągniętej w przypadku 11 sąsiadów. W przypadku kategorii „places” accuracy bardzo powoli ale jednak rośnie aż do 37 sąsiadów. Wynika z tego, że liczba sąsiadów przy której metoda Knn osiąga największe accuracy rośnie wraz z liczbą etykiet na które dzielimy dany zbiór. Wynika z tego również, że zbyt duża liczba sąsiadów zwiększa tylko czas obliczeń a nie poprawia otrzymanego wyniku.

Podział danych na zbiór uczący i testowy

Poniżej odwołujemy się do Rysunków 4 i 10 oraz Tabel 2 i 6

W oby kategoriach „places" i „topic" otrzymaliśmy podobne wyniki. Różnica pomiędzy najlepszym i najgorszym wynikiem nie przekraczała 2%. Najlepszy wynik otrzymaliśmy dla podziału 60% danych na zbiór uczący oraz 4% danych na zbiór testowy. Natomiast najgorszy wynik otrzymaliśmy w zestawie który jest lustrzanym odbiciem poprzedniego czyli 40% na zbiór uczący i 60% na zbiór testowy. Wynika z tego, że podział na zbiór uczący i testowy nie ma dużego wpływu na otrzymane wyniki. Wynika z tego również, że to który podział jest najlepszy a który najgorszy jest niezależne od kategorii i ilości etykiet według których chcemy klasyfikować nasze dane. Dodatkowo wynika z tego że podział 60% na zbiór uczący i 40% na zbiór testowy jest dużo lepszym podziałem niż podział 40% na zbiór uczący i 60% na zbiór testowy.

Miara/Metryka

Poniżej odwołujemy się do Rysunków 6 i 12 oraz Tabel 3 i 7

W obu kategoriach „places" i „topic" otrzymaliśmy różne wyniki. Różnica pomiędzy najlepszym a najgorszym wynikiem znowu nie przekracza 2%. W przypadku kategorii „places" dwie najlepsze metryki to „Metryka Uliczna" i „Nasza Metryka" a dwie najgorsze to „Metryka Euklidesa" i „Metryka Czebyszewa". W przypadku kategorii „topic" dwie najlepsze metryki to „Metryka Czebyszewa" i „Metryka Euklidesa" a dwie najgorsze to „Średnia arytmetyczna minimum" i „Minimum maximum". Wynika z tego, że metryka nie ma dużego wpływu na wynik. Wynika z tego również, że to która metryka jest najlepsza zależny od kategorii oraz ilości etykiet na które dzielimy nasz zbiór.

Różne cechy.

Poniżej odwołujemy się do Tabel 4 i 8

W przypadku kategorii „places" znaczną przewagę nad innymi mają cechy z słowami kluczowymi dla poszczególnych etykiet (jednak jest ich 6 i muszą występować wszystkie razem). Wyniki dla pozostały cech bardzo mało się różnią. Jednak za dwie najlepsze to cechy 8 i 6 a za najgorszą ponieważ jej obecność pogarsza wynik cechę 9 W przypadku kategorii „topic" cech ze słowami kluczowymi dla poszczególny etykiet nie mają już tak dużego wpływu jak poprzednio. Dwie najlepsze cechy w tym przypadku to 11 i 3 a najgorsza jest cecha 10.. Wynika z tego, że jakość cechy bardzo mocno zależy od kategorii oraz etykiet na które dzielimy nasz zbiór.

7. Wnioski

1. Znaczny wzrost accuracy występuje wraz ze wzrostem liczby sąsiadów ale tylko to pewnego momentu.
2. Zbyt duża liczba sąsiadów od pewnego momentu nie zwiększa już accuracy
3. Optymalna liczba sąsiadów jest niezależna od liczby etykiet na które dzielimy dany zbiór

4. Podział danych na zbiór uczący i testowy nie ma dużego wpływu na wyniki.
5. To który podział jest najlepszy a który najgorszy jest nie zależne od kategorii i ilości etykiet na które dzielimy nasz zbiór
6. Najlepszy jest podział 60% na zbiór uczący i 40% na zbiór testowy a najgorszy 40% na zbiór testowy i 60% na zbiór uczący.
7. Metryka nie ma dużego wpływu na accuracy otrzymanego wyniku
8. To która metryka jest najlepsza silnie zależy od kategorii i tego na jakie etykiety dzielimy nasz zbiór.
9. Cechy mają duży wpływ na otrzymane wyniki.
10. To które cechy są najlepsze silnie zależy od kategorii i tego na jakie etykiety dzielimy nasz zbiór.

Literatura

[1] Adam Niewiadomski "Materiały, przykłady i ćwiczenia do przedmiotu Komputerowe Systemy Rozpoznawania 1"
[2] https://machinelearningmastery.com/
precision-recall-and-f-measure-for-imbalanced-classification/
[3] https://machinelearningmastery.com/confusion-matrix-machine-learning/
[4] http://home.agh.edu.pl/horzyk/lectures/miw/KNN.pdf