Komputerowe systemy rozpoznawania
2024/2025
Projekt 1. Klasyfikacja dokumentów tekstowych

Dominik Gałkowski, 247659
Jan Śladowski, 247806
Prowadzący: dr inż. Marcin Kacprowicz

7 kwietnia 2025

1 Cel projektu

Celem projektu jest przygotowanie aplikacji, która będzie dokonywała klasyfikacji zbioru dokumentów tekstowych metodą k-NN. Jej zadaniem będzie przydzielenie obiektu do odpowiedniej klasy. W trakcie działania programu konieczne będzie dokonanie ekstrakcji wektorów cech z artykułów dostępnych pod linkiem: `https://archive.ics.uci.edu/dataset/137/reuters+21578+text+categorization+collection`.

2 Klasyfikacja nadzorowana metodą k-NN. Ekstrakcja cech, wektory cech

Metoda k-NN (k-Nearest Neighbors) jest algorytmem leniwym, co oznacza, że nie tworzy wewnętrznej reprezentacji danych uczących, tylko przechowuje wszystkie wzorce uczące. Dopiero po pojawieniu się wzorca testowego, dla którego wyznaczana jest odległość względem wszystkich wzorców uczących, algorytm poszukuje rozwiązania. [1]. Algorytm k-NN wymaga dwóch kluczowych parametrów, metryki, za pomocą, której wyznacza odległości obiektu testującego od wszystkich wzorców uczących oraz liczby sąsiadów k, czyli elementów do których badany element ma najbliżej. Decyzja klasyfikacyjna opiera się na najczęstszej klasie wśród k najbliższych sąsiadów. W przypadku naszego projektu odległość pomiędzy obiektami oznacza skalę podobieństwa tekstów.

W projekcie ekstrakcja cech charakterystycznych tekstu jest dokonywana poprzez stworzenie wektora cech, opisanego na podstawie następujących cech:

1. Długość tekstu - cecha ta oznacza liczbę słów, z których składa się dany

artykuł, co pozwala na porównanie długości różnych tekstów.

$$len = \sum_{i=0}^{n} x_i \tag{1}$$

gdzie x = liczba liter ≥ 3, n = liczba słów w tekście.

2. Dominująca waluta - cecha ta reprezentowana jest poprzez nazwę waluty, ze zbioru walut kluczowych, która pojawia się najczęściej w badanym artykule. Na przykład w przypadku, gdy w badanym tekście pojawi się dwukrotnie słowo "U.S. Dollar"i tylko raz "Japanese Yen"to dla tej cechy zostanie zwrócona wartość tekstowa "U.S. Dollar".

$$w = \arg\max_{w \in W} f(w) \tag{2}$$

gdzie W - zbiór walut kluczowych, $f(w)$ - liczba wystąpień waluty w w tekście.

3. Nazwy miejsca - cecha ta jest reprezentacją tekstową wszystkich miejsc, np. nazw miast lub regionów pojawiających się ze zbioru miejsc kluczowych. Przykład: "AMR Corp will hold a press conference this morning in New York at 0900 EST, a company spokesman said."wynikiem dla tego cytatu będzie zbiór $M' = \{\text{New York}\}$.

$$M' = x \in M \wedge x \in T \tag{3}$$

gdzie M - zbiór miejsc kluczowych, T - zbiór słów znajdujących się w tekście, x = liczba liter ≥ 3.

4. Liczba unikalnych słów - cecha oznaczająca wystąpienia słów unikalnych, czyli takich, które nie pojawiają się więcej niż jeden raz w tekście. Przykład: "AMR Corp will hold a press conference this morning in New York at 0900 EST, a company spokesman said. And the next week also in New York", słowa "New York"nie zostaną zliczone.

$$uk = \mid x : x \in T \wedge f(x) = 1 \mid \tag{4}$$

gdzie T - zbiór słów znajdujących się w tekście, $f(x)$ - funkcja zwracająca liczbę wystąpień słowa x w tekście, x = liczba liter ≥ 3.

5. Średnia długość słowa - cecha opisująca średnią długość słów w badanym tekście.

$$al = \frac{\sum_{i=0}^{m} a_i}{\sum_{i=0}^{n} x_i} \tag{5}$$

gdzie a_i - litera, x - liczba liter ≥ 3, n = liczba słów w tekście, m = liczba liter w tekście.

6. Liczba słów kluczowych w pierwszych 3 zdaniach - cecha ta oznacza bezwzględną liczbę wystąpień słów ze zbioru słów kluczowych w pewnym fragmencie tekstu (pierwsze 3 zdania).

$$fw = \mid x : x \in K \wedge x \in T_{\mathrm{y}} \mid \tag{6}$$

gdzie K - zbiór słów kluczowych, T_y - zbiór słów znajdujący się w pierwszych trzech zdaniach tekstu, x = liczba liter ≥ 3.

7. Liczba słów zaczynających się wielką literą - cecha ta oznaczą liczbę wystąpień słów zaczynających się wielką literą, nie uzwlędniając przy tym słów rozpoczynających nowe zdanie.

$$bw = \sum_{i=0}^{n} x_i \tag{7}$$

gdzie x = słowo zaczynające się wielką literą, n = liczba słów zaczynających się wielka literą w tekście

8. Pierwsze kluczowe słowo w tekście - cecha opisująca pierwsze znalezione słowo znajdujące się w zbiorze słów kluczowych. Przykład: "AMR Corp will hold a press conference this morning in New York at 0900 EST, a company spokesman said."wynikiem dla tego cytatu będzie x_{first} = New York.

$$x_{first} = \min\{x : x \in K \wedge x \in T\} \tag{8}$$

gdzie K - zbiór słów kluczowych, T - zbiór słów znajdujących się w tekście, x = liczba liter ≥ 3.

9. Liczba słów kluczowych - cecha ta oznacza bezwzględną liczbę wystąpień słów ze zbioru słów kluczowych.

$$kw = \mid x : x \in K \wedge x \in T \mid \tag{9}$$

gdzie K - zbiór słów kluczowych, T - zbiór słów znajdujących się w tekście, x = liczba liter ≥ 3.

10. Względna liczba słów kluczowych - cecha która reprezentuje stosunek słów kluczowych do długości całego tekstu.

$$rw = \frac{\mid x : x \in K \wedge x \in T \mid}{\sum_{i=0}^{n} x_i} \tag{10}$$

gdzie K - zbiór słów kluczowych, x = liczba liter ≥ 3, T - zbiór słów znajdujących się w tekście, n = liczba słów w tekście

11. Nazwiska - cecha ta jest reprezentacją tekstową wszystkich nazwisk pojawiających się ze zbioru nazwisk kluczowych. Przykład: "Wallis was quoted as saying the Reagan Administration wants Japanese cooperation so the

White House can ensure any U.S."wynikiem dla tego cytatu będzie zbiór $N' = \{\text{Reagan}\}$.

$$N' = x \in N \wedge x \in T \tag{11}$$

gdzie N - zbiór nazwisk kluczowych, T - zbiór słów znajdujących się w tekście, x = liczba liter ≥ 3.

Wektor cech będzie miał postać:

$$v = [c1, c2, c3, c4, c5, c6, c7, c8, c9, c10, c11] \tag{12}$$

3 Miary jakości klasyfikacji

W celu określenia jakości przeprowadzonej klasyfikacji należy skorzystać z czterech miar jakości. W trakcie omawiania tej sekcji będziemy się posługiwać symbolami, które będą oznaczać klasy, do których można przypisać dany tekst (J - Japonia, F - Francja, W - Niemcy Zachodnie, C - Kanada, U - USA, UK - Wielka Brytania).

3.1 Dokładność (Accuracy)

Dokładność to miara, która określa jaka część obiektów, ze wszystkich zaklasyfikowanych, została zaklasyfikowana poprawnie. Dokładność jest obliczana dla wszystkich klas jednocześnie i przyjmuje wartości z zakresu $[0, 1]$. Wyższa wartość dokładności oznacza, że ogólny procent poprawnie sklasyfikowanych obiektów jest większy, co sugeruje, że skuteczność klasyfikatora jest większa.

$$ACC = \frac{TP}{TP + N} \tag{13}$$

gdzie ACC - accuracy, TP - liczba wszystkich poprawnie sklasyfikowanych tekstów, N - liczba niepoprawnie sklasyfikowanych tekstów.

3.2 Precyzja (Precision)

Dzięki precyzji dowiadujemy się, ile wśród obiektów sklasyfikowanych do danej klasy jest rzeczywiście tej klasy. Precyzja jest obliczana dla wszystkich klas oddzielnie i przyjmuje wartości z zakresu $[0, 1]$. Im wyższy współczynnik precyzji, tym mniej błędnych klasyfikacji do danej klasy.

$$PPV_x = \frac{TP_x}{TP_x + N_x} \tag{14}$$

gdzie PPV_x - precision dla danej klasy x, TP_x - liczba poprawnie sklasyfikowanych tekstów klasy x, N_x - liczba niepoprawnie sklasyfikowanych tekstów do klasy x, $x \in \{\text{C, J, U, F, W, UK}\}$.

3.3 Czułość (Recall)

Czułość opisuje jaki jest udział poprawnie sklasyfikowanych obiektów wśród wszystkich obiektów tej klasy. Czułość jest obliczana dla wszystkich klas oddzielnie i przyjmuje wartości z zakresu $[0, 1]$. Wyższa wartość czułości oznacza, że klasyfikator skuteczniej wykrywa wszystkie przypadki danej klasy, co oznacza zmniejszenie liczby pominiętych istotnych obiektów.

$$TPR_x = \frac{TP_x}{TP_x + NF_x} \tag{15}$$

gdzie TPR_x - recall dla danej klasy x, TP_x - liczba poprawnie sklasyfikowanych tekstów klasy x, NF_x - liczba tekstów klasy x, które zostały przypisane do innej klasy, $x \in \{\text{C, J, U, F, W, UK}\}$.

3.4 F1

F1 to średnia harmoniczna pomiędzy precyzją a czułością, pozwalająca ocenić równowagę między nimi. F1 jest obliczana dla wszystkich klas oddzielnie i przyjmuje wartości z zakresu $[0, 1]$. Im wyższa wartość miary F1, tym lepsza równowaga pomiędzy precyzją, a czułością

$$F1_x = \frac{2 \times PPV_x \times TPR_x}{PPV_x + TPR_x} \tag{16}$$

gdzie $F1_x$ - miara F1 dla danej klasy x, $x \in \{\text{C, J, U, F, W, UK}\}$.

3.5 Przykład z wykorzystaniem miar jakości klasyfikacji

Mamy trzy zbiory, na ich podstawie obliczymy accuracy oraz precision, recall i F1 dla tekstów przypisanych do klasy Japonii:

1. Zbiór tekstów przypisanych jako Japonia {J, J, J, F, U}.
2. Zbiór tekstów przypisanych jako Francja {F, F, F, J}.
3. Zbiór tekstów przypisanych jako USA {U, U, F, F}.

-	TP_X	N_X	NF_X
Japonia (J)	$TP_J = 3$	$N_J = 2$	$NF_J = 1$
Francja (F)	$TP_F = 3$	$N_F = 1$	$NF_F = 3$
USA (U)	$TP_U = 2$	$N_U = 2$	$NF_U = 1$

Tabela 1: Wartości dla klasyfikacji tekstów

gdzie TP_x - liczba poprawnie sklasyfikowanych tekstów klasy x, N_x - liczba niepoprawnie sklasyfikowanych tekstów do klasy x, NF_x - liczba tekstów klasy x, które zostały przypisane do innej klasy, $x \in \{\text{C, J, U, F, W, UK}\}$.

- $TP = 3 + 3 + 2 = 8$ (suma wszystkich poprawnie sklasyfikowanych tekstów),
- $N = 2 + 1 + 2 = 5$ (suma wszystkich tekstów przypisanych do niewłaściwej klasy).

$$ACC = \frac{TP}{TP + N} = \frac{8}{13} \approx 0.62$$

- $TP_J = 3$ (Liczba tekstów poprawnie sklasyfikowanych do Japonii),
- $N_J = 2$ (Liczba tekstów niepoprawnie przypisanych do Japonii).

$$PPV_J = \frac{TP_J}{TP_J + N_J} = \frac{3}{5} = 0.6$$

- $TP_J = 3$ (Liczba tekstów poprawnie przypisanych do Japonii),
- $NF_J = 1$ (Liczba tekstów klasy Japonia, które zostały błędnie przypisane do innej klasy).

$$TPR_J = \frac{TP_J}{TP_J + NF_J} = \frac{3}{4} = 0.75$$

$$F1_J = \frac{2 \times PPV_J \times TPR_J}{PPV_J + TPR_J} = \frac{0.9}{1.35} \approx 0.67$$

4 Metryki i miary podobieństwa tekstów w klasyfikacji

Metoda klasyfikacji k-NN polega na znajdowaniu k najbliższych sąsiadów, kluczową rolę w tym procesie odgrywają metryki oraz miary, które są wykorzystywane do ustalenia stopnia zgodności pomiędzy obiektami. Metryki umożliwiają obliczenie odległości między wektorami liczbowymi. Natomiast w przypadku cech tekstowych, zanim będzie można obliczyć ich podobieństwo, należy dokonać ich transformacji na wartości liczbowe. Umożliwiają to miary, które określają podobieństwo między ciągami znaków.

4.1 Metryki

1. Metryka euklidesowa - w celu obliczenia odległości $\rho_E(v1, v2)$ między dwoma wektorami $v1, v2$ należy obliczyć pierwiastek kwadratowy z sumy kwadratów różnic ich składowych zgodnie ze wzorem:

$$\rho_E(v1, v2) = \sqrt{\sum_{i=1}^{n} (v1_i - v2_i)^2} \tag{17}$$

   gdzie n - liczba cech w wektorach.

2. Metryka uliczna - w celu obliczenia odległości $\rho_M(v1, v2)$ między dwoma wektorami $v1, v2$ należy obliczyć sumę wartości bezwzględnych różnic cech zgodnie ze wzorem:

$$\rho_M(v1, v2) = \sum_{i=1}^{n} |v1_i \underset{m}{-} v2_i| \tag{18}$$

gdzie n - liczba cech w wektorach.

3. Metryka Czebyszewa - w celu obliczenia odległości $\rho_C(v1, v2)$ między dwoma wektorami $v1, v2$ należy obliczyć maksymalną wartość bezwzględnych różnic cech zgodnie ze wzorem:

$$\rho_C(v1, v2) = \max_{i=1,\ldots,n} |v1_i \underset{m}{-} v2_i| \tag{19}$$

gdzie n - liczba cech w wektorach.

gdzie

$$v1_i \underset{m}{-} v2_i = \begin{cases} v1_i - v2_i & \text{dla } v1_i, v2_i \in \mathbb{R} \\ 1 - sim_w(v1_i, v2_i) & \text{jeśli } v1_i, v2_i \text{ są tekstami} \\ 1 - sim_z(v1_i, v2_i) & \text{jeśli } v1_i, v2_i \text{ są zbiorami tekstów} \end{cases}$$

gdzie $v1_i$, $v2_i$ - i-ta składowa wektorów cech $v1$ oraz $v2$,
sim_w - podobieństwo tekstów obliczone uogólnioną miarę n-gramów z ograniczeniami (pkt 4.2),
sim_z - podobieństwo zbiorów wyrazów obliczone miarą podobieństwa zdań (pkt 4.3).
W celu poprawnego przeprowadzenia obliczeń dla metryk należy uprzednio przeprowadzić normalizację cech wektórów, tak aby żadna z cech nie była dominująca. Wektory zostaną znormalizowane za pomocą metody min-max scaling do zakresu [0, 1].
Załóżmy, że mamy dwa wektory cech:

1. $v1 = (1, 2, 30)$
2. $v2 = (4, 6, 3)$

$$c_{\min} = 1, \quad c_{\max} = 30$$

gdzie c to cecha składowa wektora $v1$ lub $v2$.
Aby otrzymać znormalizowane wartości należy skorzystać z wzoru:

$$c' = \frac{c - c_{min}}{c_{max} - c_{min}} \tag{20}$$

Znormalizowana postać wektorów:

1. $v1' = (0.000, 0.034, 1.000)$

2. $v2' = (0.103, 0.172, 0.069)$

Z wykorzystaniem powyższych wektorów otrzymujemy:

$$\rho_E(v1, v2) = \sqrt{(0.000 - 0.103)^2 + (0.034 - 0.172)^2 + (1.000 - 0.069)^2} =$$
$$= \sqrt{0.946} \approx 0.947$$
$$\rho_M(v1, v2) = |0.000 - 0.103| + |0.034 - 0.172| + |1.000 - 0.069| = 1.172$$
$$\rho_C(v1, v2) = \max(|0.000 - 0.103|, |0.034 - 0.172|, |1.000 - 0.069|) =$$
$$= \max(0.103, 0.138, 0.931) = 0.931$$

Metryka euklidesowa, uliczna oraz Czebyszewa przyjmują wartości z zakresu $[0, \infty)$. Im otrzymana wartość jest mniejsza, tym oba wektory cech są do siebie bardziej podobne.

4.2 Uogólniona miara n-gramów z ograniczeniami

Wykorzystując uogólnioną miarę n-gramów z ograniczeniami możemy pewną liczbą wyrazić podobieństwo dwóch łańcuchów znaków. Ta miara przyjmuje wartości z zakresu $[0, 1]$, przy czym wartości wyższe oznaczają większe podobieństwo po między badanymi łańcuchami znaków. Krańcowe wartości oznaczają: 0 – różne łańcuchy znaków, 1 - identyczne łańcuchy znaków. Przekształcenie cech tekstowych na wartości numeryczne umożliwia obliczenie ich wpływu na odległości między wektorami. Odległość pomiędzy dwoma łańcuchami znaków możemy określić poprzez:

$$d = 1 - sim_w(s1, s2) \tag{21}$$

gdzie $sim_w(s1, s2)$ oznacza uogólnioną miarę n-gramów z ograniczeniami

$$\mu_N(s1, s2) = f(N, n_1, n_2) \sum_{i=n_1}^{n_2} \sum_{j=1}^{N(s1)-i+1} h(i, j) \tag{22}$$

gdzie $s1, s2$ - cechy, które przyjmują wartości tekstowe;

$$f(N, n_1, n_2) = \frac{2}{(N - n_1 + 1)(N - n_1 + 2) - (N - n_2 + 1)(N - n_2)} \tag{23}$$

wyraża odwrotność liczby możliwych podciagów o długosciach od n_1 do n_2
$1 \leq n_1 \leq n_2 \leq N$;
$h(i, j) = 1$ jeśli i-elementowy podciag w słowie s1 zaczynajacy sie od j-tej pozycji w słowie s_1 pojawia sie przynajmniej raz w słowie s_2 (inaczej $h(i, j) = 0$);
$N(s1), N(s2)$ - oznaczają liczbę liter w słowach $v1$ i $v2$;
$N = \max\{N(v1), N(v2)\}$.

Załóżmy, że mamy dwa wektory cech (wektory powinny być znomalizowane, ale na potrzeby tego przykładu zostało to pominięte):

1. $v1 = (1, KARTON, 3)$
2. $v2 = (4, KARNISZ, 3)$

Traktując drugą cechę jako łańcuchy znaków, mamy:

$$s_1 = \{K, A, R, T, O, N\}, \quad s_2 = \{K, A, R, N, I, S, Z\}$$

czyli:

$$N(s_1) = 6, N(s_2) = 7, N = \max\{N(s_1), N(s_2)\} = 7$$

Obliczając podobieństwo przyjmujemy $n_1 = 2$ oraz $n_2 = 3$

$$\mu_N(s_1, s_2) = \frac{2}{(7-2+1)(7-2+2) - (7-3+1)(7-3)} \sum_{i=2}^{3} \sum_{j=1}^{6-i+1} h(i,j) =$$

$$= \frac{2+1}{11} \approx 0.27.$$

ponieważ w s_2 występują poniższe podciągi z s_1
2 - 2-elementowe KA, AR;
1 - 3-elementowy KAR;

Wówczas odległość euklidesowa pomiędzy wektorami wynosi:

$$\rho_E(v1, v2) = \sqrt{(1-4)^2 + (1-0.27)^2 + (3-3)^2} = \sqrt{9 + 0.73 + 0} =$$

$$= \sqrt{9.73} \approx 3.12$$

4.3 Miara podobieństwa zdań

Wykorzystując uogólnioną miarę podobieństwa zdań, traktowanych jako zbiory (a nie ciągi) wyrazów, możemy pewną liczbą wyrazić stopień podobieństwa pomiędzy dwoma zdaniami. Miara ta przyjmuje wartości z zakresu $[0, 1]$, gdzie wyższe wartości oznaczają większe podobieństwo między porównywanymi zdaniami. Krańcowe wartości interpretujemy następująco: 0 – zdania zupełnie różne, 1 – zdania identyczne pod względem zestawu użytych słów. Przekształcenie cech tekstowych na wartości numeryczne umożliwia analizę ich wpływu na odległości w przestrzeni wektorowej. Odległość pomiędzy dwoma zdaniami możemy określić za pomocą następującej formuły:

$$d = 1 - sim_z(z1, z2) \tag{24}$$

gdzie $sim_z(z1, z2)$ oznacza miarę podobieństwa zdań

$$\mu_{NZ}(z_1, z_2) = \frac{1}{N} \sum_{i=1}^{N(z_1)} \max_{j=1,\ldots,N(z_1)} \mu_N(s_{1i}, s_{2j}) \tag{25}$$

gdzie:
s_{1i} – i-ty wyraz w zdaniu z_1;
s_{2j} – j-ty wyraz w zdaniu z_2;
$\mu_N(s_{1i}, s_{2j})$ – wartość funkcji (22) dla (s_{1i}, s_{2j});
$N(z_1), N(z_2)$ – liczba słów w zdaniach z_1, z_2;
$N = \max\{N(z_1), N(z_2)\}$.

Załóżmy, że mamy dwa wektory cech (wektory powinny być znormalizowane, ale na potrzeby tego przykładu zostało to pominięte):

1. $v_1 = (1,\ \text{kot je}, 3)$
2. $v_2 = (4,\ \text{kot pije wodę}, 3)$

Traktując drugą cechę jako zdania złożone ze zbiorów wyrazów, mamy:

$$z_1 = \{\text{kot}, \text{je}\}, \quad z_2 = \{\text{kot}, \text{pije}, \text{wodę}\}$$

czyli:

$$N(z_1) = 2, \quad N(z_2) = 3, \quad N = \max\{N(z_1), N(z_2)\} = 3$$

Wartość podobieństwa zdań:

$$\mu_{NZ}(z_1, z_2) = \frac{1}{3}\sum_{i=1}^{3} \max_{j=1,2} \mu_N(s_{1i}, s_{2j}) = \frac{1 + 0.2 + 0}{3} = 0.4$$

gdzie

$$\max\{\mu_N(\text{kot}, \text{kot}), \mu_N(\text{je}, \text{kot})\} = 1.0$$
$$\max\{\mu_N(\text{kot}, \text{pije}), \mu_N(\text{je}, \text{pije})\} = 0.2$$
$$\max\{\mu_N(\text{kot}, \text{wodę}), \mu_N(\text{je}, \text{wodę})\} = 0$$

Wówczas odległość euklidesowa pomiędzy wektorami wynosi:

$$\rho_E(v_1, v_2) = \sqrt{(1-4)^2 + (1-0.4)^2 + (3-3)^2} = \sqrt{9 + 0.36 + 0} = \sqrt{9.25} \approx 3.06$$

5 Wyniki klasyfikacji dla różnych parametrów wejściowych

W niniejszej sekcji przeprowadzono eksperymenty polegające na przeprowadzeniu klasyfikacji tekstów dla ograniczonego zbioru składającego się z 1277 artykułów. Celem tego badania było przeprowadzenie analizy wpływu parametrów wejściowych algorytmu k-najbliższych sąsiadów na skuteczność klasyfikatora opisaną jako dokładność (accuracy).
Badanie przeprowadzono w oparciu o różne warianty konfiguracji parametów wejściowych, zgodnie z punktami 3–8 zawartymi w opisie Projektu 1. Uwzględniono m.in. różne wartości parametru k, różny podział zbioru pomiędzy uczący, a testowy oraz różne metryki. Eksperymenty polegają na tym, że w każdym z nich będzie zmieniany tylko jeden parametr, którego wpływ będzie aktualnie badany.

5.1 Różne wartości parametru k

W poniższej tabeli przedstawiono wpływ różnych wartości parametru k na dokładność klasyfikacji tekstów. Eksperymenty przeprowadzono przy stałych pozostałych ustawieniach: podziale zbioru danych na 60% zbioru uczącego i 40% zbioru testowego, metryce euklidesowej oraz zakresie długości n-gramów od $n_1 = 2$ do $n_2 = 4$.

Nr	**Parametr k**	**Accuracy**
1	2	0.8611
2	3	0.8885
3	5	0.8885
4	10	0.8924
5	20	0.8826
6	50	0.8532

Tabela 2: Wyniki miary Accuracy dla różnych parametrów k

Na podstawie wyników można zaobserwować, że miara Accuracy rośnie wraz ze wzrostem parametru k do wartości $k = 10$, osiągając maksimum równe 0,8924. Dalsze zwiększanie wartości k prowadzi jednak do spadku dokładności klasyfikatora. Może to świadczyć o nadmiernym uogólnieniu modelu przy zbyt dużej liczbie sąsiadów, co skutkuje mniejszą precyzją klasyfikacji.

5.2 Różny podział zbioru pomiędzy uczący, a testowy

W poniższej tabeli przedstawiono wpływ proporcji podziału zbioru na dokładność klasyfikacji tekstów. Eksperymenty przeprowadzono przy stałych pozostałych ustawieniach: $k = 10$, metryce euklidesowej oraz zakresie długości n-gramów od $n_1 = 2$ do $n_2 = 4$. Wartości w tabeli pod kolumną "Podział zbioru"opisują ile jest artykułów w zbiorze uczącym względem liczby wszystkich artykułów.

Nr	**Podział zbioru [%]**	**Accuracy**
1	20	0.8787
2	40	0.8827
3	60	0.8924
4	80	0.9063

Tabela 3: Wyniki miary Accuracy dla różnego podziału zbiorów

Wyniki wskazują na istotny wpływ proporcji podziału zbioru na dokładność klasyfikacji. W miarę zwiększania udziału zbioru uczącego w całości danych, skuteczność klasyfikatora również rośnie, osiągając najwyższą wartość przy podziale 80% danych do zbioru uczącego. Zwiększenie tego udziału powoduje, że model ma więcej danych do nauki, co przekłada się na lepsze dopasowanie do

wzorców w danych. Z kolei mniejsze proporcje danych uczących prowadzą do gorszej efektywności, co może być wynikiem niedostatecznej liczby próbek w procesie treningu.

5.3 Różne metryki

W poniższej tabeli przedstawiono wpływ różnych metryk na dokładność klasyfikacji tekstów. Eksperymenty przeprowadzono przy stałych pozostałych ustawieniach: podziale zbioru danych na 60% zbioru uczącego i 40% zbioru testowego $k = 10$ oraz zakresie długości n-gramów od $n_1 = 2$ do $n_2 = 4$.

Nr	Metryka	Accuracy
1	Metryka euklidesowa	0.8924
2	Metryka uliczna	0.8767
3	Metryka Czybyszewa	0.8102

Tabela 4: Wyniki miary Accuracy dla różnych metryk

Wyniki wskazują na najlepszą skuteczność klasyfikatora przy zastosowaniu metryki euklidesowej, uzyskując miarę Accuracy równą 0,8864. Metryka uliczna, choć również dała zadowalający wynik (Accuracy = 0,8767), okazała się nieco mniej efektywna w porównaniu do metryki euklidesowej. Z kolei metryka Czybyszewa zwróciła najgorszy wynik.

Wstępne wyniki miary Accuracy dla próbnych klasyfikacji na ograniczonym zbiorze tekstów (podać parametry i kryteria wyboru wg punktów 3.-8. z opisu Projektu 1.). **Sekcja uzupełniona jako efekt zadania Tydzień 05 wg Harmonogramu Zajęć na WIKAMP KSR.**

6 Dyskusja, wnioski, sprawozdanie końcowe

Wyniki kolejnych eksperymentów wg punktów 2.-8. opisu projektu 1. Każdorazowo podane parametry, dla których przeprowadzana eksperyment. Wykresy (np. słupowe) i tabele wyników obowiązkowe, dokładnie opisane w „captions" (tytułach), konieczny opis osi i jednostek wykresów oraz kolumn i wierszy tabel.

Ewentualne wyniki realizacji punktu 9. opisu Projektu 1., czyli „na ocenę 5.0" i ich porównanie do wyników z części obowiązkowej.Dokładne interpretacje uzyskanych wyników w zależności od parametrów klasyfikacji opisanych w punktach 3.-8 opisu Projektu 1. Omówić i wyjaśnić napotkane problemy (jeśli były). Każdy wniosek/problem powinien mieć poparcie w przeprowadzonych eksperymentach (odwołania do konkretnych wyników: wykresów, tabel).
Dla końcowej oceny jest to najważniejsza sekcja sprawozdania, gdyż prezentuje poziom zrozumienia rozwiązywanego problemu.

** Możliwości kontynuacji prac w obszarze systemów rozpoznawania, zwłaszcza w kontekście pracy inżynierskiej, magisterskiej, naukowej, itp. **

Sekcja uzupełniona jako efekt zadań Tydzień 05 i Tydzień 06 wg Harmonogramu Zajęć na WIKAMP KSR.

7 Braki w realizacji projektu 1.

Wymienić wg opisu Projektu 1. wszystkie niezrealizowane obowiązkowe elementy projektu, ewentualnie podać merytoryczne (ale nie czasowe) przyczyny tych braków.

Literatura

[1] Metoda k-NN `https://home.agh.edu.pl/~horzyk/lectures/miw/KNN.pdf` [dostęp: 28.03.2025r.]

[2] Wikipedia, Tablica pomyłek, `https://pl.wikipedia.org/wiki/Tablica_pomy%C5%82ek`. [dostęp: 28.03.2025r.]

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

Literatura zawiera wyłącznie źródła recenzowane i/lub o potwierdzonej wiarygodności, możliwe do weryfikacji i cytowane w sprawozdaniu.