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

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

30 marca 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}.

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

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

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

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

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 $v1_i$, $v2_i$ - i-ta składowa wektorów cech $v1$ oraz $v2$, 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:

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

   gdzie $v1_i$, $v2_i$ - i-ta składowa wektorów cech $v1$ oraz $v2$, 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:

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

   gdzie $v1_i$, $v2_i$ - i-ta składowa wektorów cech $v1$ oraz $v2$, n - liczba cech w wektorach.

Załóżmy, że mamy dwa wektory cech:

1. $v1 = (1, 2, 3)$

2. $v2 = (4, 6, 3)$

$$\rho_E(v1, v2) = \sqrt{(1-4)^2 + (2-6)^2 + (3-3)^2} = \sqrt{9+16+0} = \sqrt{25} = 5$$

$$\rho_M(v1, v2) = |1-4| + |2-6| + |3-3| = 3 + 4 + 0 = 7$$

$$\rho_C(v1, v2) = \max(|1-4|, |2-6|, |3-3|) = \max(3, 4, 0) = 4$$

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 Miara Jaccarda

Wykorzystując miarę Jaccarda możemy pewną liczbą wyrazić podobieństwo dwóch łańcuchów znaków. Miara Jaccarda przyjmuje wartości z zakresu $[0, 1]$, przy czym wartości wyższe oznaczają większe podobnieństwo dwóch: 0 oznacza brak wspólnych elementów, 1 oznacza identyczne zbiory znaków. Przekształcenie cech tekstowych na wartości numeryczne umożliwia obliczenie ich wpływu na odległości między wektorami. Odległość pomiędzy dwoma łańcuchami znaków możemy określić poprzez:

$$d = 1 - sim(A, B) \tag{20}$$

gdzie $sim(A, B)$ oznacza miarę Jaccarda

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{21}$$

gdzie A oraz B to zbiory składające się z liter cechy tekstowej odpowiedniej dla wektora $v1$ oraz $v2$,
$|A \cap B|$ to **moc części wspólnej** dwóch zbiorów A i B, czyli liczba elementów (w tym przypadku liter), które znajdują się **w obu zbiorach jednocześnie**,
$|A \cup B|$ to **moc sumy** zbiorów A i B, czyli liczba elementów (liter), które znajdują się **w co najmniej jednym z tych zbiorów**. Przy tym, zbiór sumy musi zawierać tylko **unikalne** elementy, czyli powtarzające się litery liczymy tylko raz.

Załóżmy, że mamy dwa wektory cech:

1. $v1 = (1, ABC, 3)$
2. $v2 = (4, CED, 3)$

$$A = \{A, B, C\}, \quad B = \{C, E, D\}$$

$$J(A, B) = \frac{\{\text{C}\}}{\{\text{A,B,C,E,D}\}} = \frac{1}{5} = 0.2.$$

$$\rho_E(v1, v2) = \sqrt{(1-4)^2 + (1-0.2)^2 + (3-3)^2} = \sqrt{9 + 0.64 + 0} = \sqrt{25} = 3.10$$

5 Wyniki klasyfikacji dla różnych parametrów wejściowych

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.