Informatyka, studia dzienne, I st. semestr VI

Komputerowe systemy rozpoznawania **2019/2020**
Prowadzący: dr inż. Marcin Kacprowicz poniedziałek, 12:00

Data oddania: _____ Ocena: _____

Radosław Grela 216769
Jakub Wąchała 216914

Zadanie 1: ekstrakcja cech, miary podobieństwa, klasyfikacja*

1. Cel

Celem naszego zadania było stworzenie aplikacji do klasyfikacji tekstów za pomocą metody k-NN (k najbliższych sąsiadów) oraz różnych metryk i miar podobieństwa, a następnie porównać kategorie z tymi wygenerowanymi przez aplikację.

2. Wprowadzenie

Głównym zagadnieniem projektowym, z którym mieliśmy do czynienia w ramach zadania 1 była klasyfikacja statystyczna tekstów na podstawie wektora wyekstrahowanych cech. Do przeprowadzenie eksperymentu zaimplementowaliśmy algorytm *k-najbliższych sąsiadów*.

Algorytm k-najbliższych sąsiadów *(k-NN - k-nearest neighbors)* to jeden z algorytmów zaliczanych do grupy algorytmów leniwych. Jest to taka grupa algorytmów, która szuka rozwiązania dopiero, gdy pojawia się wzorzec testujący. Przechowuje wzorce uczące, a dopiero później wyznacza się odległość wzorca testowego względem wzorców treningowych. [1].

Algorytm ten działa w taki sposób, że dla każdego wzorca testowego obliczana jest odległość za pomocą wybranej wetryki względem wzorców treningowych, a następnie wybierana jest k najbliższych wzorców treningowych.

* Github: `https://github.com/Bonniu/KSR`

Wynik wyznaczony jest jako najczęstszy element wśród nich. W naszym zadaniu odległość ta jest równa skali podobieństwa tekstów.

2.1. Ekstrakcja cech

Do ekstrakcji cech charakterystycznych tekstu utworzyliśmy wektor cech, który opisuje tekst za pomocą 10 cech. Liczba słów zawsze jest liczona po zastosowaniu stop-listy oraz stemizacji, bez znaków przestankowych.

- C_1 - Stosunek słów kluczowych do wszystkich słów w pierwszych 10% tekstu. Obliczona jest za pomocą wzoru:

$$C_1 = s_{k10}/s_{10} \tag{1}$$

  gdzie s_{k10} to liczba słów kluczowych, a s_{10} to liczba wszystkich słów w pierwszych 10% tekstu. Przed normalizacją cecha C_1 zawierała się w wartościach $\in [0, 1]$.
- C_2 - Stosunek słów kluczowych do wszystkich słów w ostatnich 10% tekstu. Obliczona jest za pomocą wzoru:

$$C_2 = s_{k90}/s_{90} \tag{2}$$

  gdzie s_{k90} to liczba słów kluczowych, a s_{90} to liczba wszystkich słów w ostatnich 10% tekstu. Przed normalizacją cecha C_2 zawierała się w wartościach $\in [0, 0.5]$.
- C_3 - Stosunek słów kluczowych do wszystkich słów w dokumencie. Obliczona jest za pomocą wzoru:

$$C_3 = s_k/s \tag{3}$$

  gdzie s_k to liczba słów kluczowych, a s to liczba wszystkich słów w dokumencie. Przed normalizacją cecha C_3 zawierała się w wartościach $\in [0, 0.155]$.
- C_4 - Stosunek słów kluczowych, których ilość liter $\in$ (0,4] do wszystkich słów w dokumencie. Obliczona jest za pomocą wzoru:

$$C_4 = s_k/s \tag{4}$$

  gdzie s_k to liczba słów kluczowych, których ilość liter $\in$ (0,4], a s to liczba wszystkich słów w dokumencie. Przed normalizacją cecha C_4 zawierała się w wartościach $\in [0, 0.075]$.
- C_5 - Stosunek słów kluczowych, których ilość liter jest $\geq$8 do wszystkich słów w dokumencie. Obliczona jest za pomocą wzoru:

$$C_5 = s_k/s \tag{5}$$

  gdzie s_k to liczba słów kluczowych, a s to liczba wszystkich słów w dokumencie. Przed normalizacją cecha C_5 zawierała się w wartościach $\in [0, 0.1]$.
- C_6 - Stosunek linii do ilości akapitów. Obliczona jest za pomocą wzoru:

$$C_6 = l/a \tag{6}$$

  gdzie l to liczba linii, a a to liczba akapitów. Przed normalizacją cecha C_6 zawierała się w wartościach $\in [1, 14]$.

- C_7 - Stosunek słów, których ilość liter jest większa niż 6 do wszystkich słów. Obliczona jest za pomocą wzoru:

$$C_7 = s_6/s \tag{7}$$

gdzie s_6 to liczba słów których ilość liter jest większa niż 6, a s to liczba wszystkich słów w dokumencie. Przed normalizacją cecha C_7 zawierała się w wartościach $\in [0, 0.591]$.
- C_8 - Stosunek słów kluczowych, których ilość liter jest $\leq$6 do wszystkich słów w dokumencie. Obliczona jest za pomocą wzoru:

$$C_8 = s_{6m}/s \tag{8}$$

gdzie s_{6m} to liczba słów kluczowych, których ilość liter jest $\leq$6, a s to liczba wszystkich słów w dokumencie. Przed normalizacją cecha C_8 zawierała się w wartościach $\in [0.409, 1]$.
- C_9 - Ilość słów unikalnych. Jest to liczba słów, które wystąpiły w tekście co najmniej raz. Przykładowo, dla zdania *„Być albo nie być"* ilość słów unikalnych jest równa 3 (*być, albo, nie*). Przed normalizacją cecha C_9 przyjmuje wartości $\in [1, 420]$.
- C_{10} - Ilość słów, których ilość liter $\in$ [5,8]. Pseudokod obliczający wartość cechy C_{10}:
  — *C_{10}=0*
  — *Dla każdego słowa w artykule:*
    — *Jeżeli długość słowa>=5 i długość słowa <=8:*
      — *C_{10}++;*
  — *Zwróć C_{10}*

  Przed normalizacją cecha C_{10} zawierała się w wartościach $\in [1, 574]$.

- W przypadku cech mniej intuicyjnych (a najlepiej wszystkich) - mile widziany przykład jak liczyć (może być jeden tekst dla wszystkich cech pod ich opisem). - Jakie znaczenie ma ta cecha tekstu dla jego rozpoznania? Czy np. im tekst dłuższy, tym bardziej związany z etykietą USA lub CANADA? (istotne!)

2.2. Metryki i miary

Do liczenia odległości pomiędzy artykułami oraz obliczenia miary podobieństwa używaliśmy 3 metryk i 2 miar.

1. Metryka Euklidesowa - aby obliczyć odległość $d_e(x, y)$ między artykułami x i y należy obliczyć pierwiastek kwadratowy z sumy kwadratów różnic wartości współrzędnych wektora o tych samych indeksach. Wzór jest następujący:

$$d_e(x, y) = \sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2} \tag{9}$$

2. Metryka Manhattan - odległość $d_m(x, y)$ jest równa sumie wartości bezwzględnych z różnic wartości współrzędnych wektora o tych samych indeksach:

$$d_m(x, y) = \sum_{n=1}^{N} |x_n - y_n| \tag{10}$$

3. Metryka Czebyszewa - odległość $d_c(x, y)$ w tej metryce jest równa maksymalnej wartości bezwzględnych różnic współrzędnych punktów x oraz y, zgodnie ze wzorem:
$$d_c(x, y) = \max_i |x_i - y_i| \tag{11}$$

4. Miara *Term Frequency Matrix*, czyli po polsku „macierz częstości występowania terminów" podaje wartość podobieństwa dokumentów d_1 i d_2 ze względu na wybrany zbiór terminów, np. słów kluczowych. Ustawiamy macierz słów kluczowych i dokumentów:
$$\begin{array}{c} \\ d_1 \\ d_2 \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & \dots & t_n \end{array} \\ \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \end{bmatrix} \end{array} \tag{12}$$
Następnie podobieństwo otrzymanych wektorów jest obliczone przy pomocy amplitudy kosinusowej:
$$r_{ca}(V_1, V_2) = \frac{|\sum_{i=1}^{n} a_{1i} * a_{2i}|}{\sqrt{\sum_{i=1}^{n} a_{1i}^2 * \sum_{i=1}^{n} a_{2i}^2}} \tag{13}$$

5. Miara *n-gramów*

3. Opis implementacji

Należy tu zamieścić krótki i zwięzły opis zaprojektowanych klas oraz powiązań między nimi. Powinien się tu również znaleźć diagram UML (diagram klas) prezentujący najistotniejsze elementy stworzonej aplikacji. Należy także podać, w jakim języku programowania została stworzona aplikacja.

4. Materiały i metody

Wykonana przez nas klasyfikacja została wykonana za pomocą wszystkich trzech metryk. Każdy przypadek testowy był klasyfikowany dla dziesięciu różnych wartości k najbliższych sąsiadów: 2, 3, 4, 5, 7, 10, 13, 15, 20, 25.

Klasyfikacji dokonywaliśmy tylko na tych tekstach, które miały jedną z etykiet: *west-germany, usa, france, uk, canada, japan* i były to ich jedyne etykiety.

Dokonaliśmy pięciu różnych podziałów na dane testowe oraz treningowe:
— 10% dane treningowe, 90% dane testowe
— 30% dane treningowe, 70% dane testowe
— 50% dane treningowe, 50% dane testowe
— 70% dane treningowe, 30% dane testowe
— 85% dane treningowe, 15% dane testowe

5. Wyniki

5.1. Badanie podziału na dane treningowe i testowe

k	Accuracy [%]
2	66,19
3	72,61
4	73,81
5	75,87
7	77,56
10	77,95
13	78,71
15	78,98
20	79,07
25	79,41

Tabela 1. Skuteczność klasyfikacji dla metryki Euklidesowej dla podziału zbioru 30% treningowe/70% testowe.

k	Accuracy [%]
2	67,14
3	72,97
4	74,25
5	75,79
7	77,82
10	78,74
13	79,26
15	79,35
20	79,74
25	79,81

Tabela 2. Skuteczność klasyfikacji dla metryki Euklidesowej dla podziału zbioru 50% treningowe/50% testowe.

k	Accuracy [%]
2	67,62
3	73,95
4	75,52
5	77,64
7	79,86
10	80,77
13	81,11
15	81,38
20	81,38
25	81,48

Tabela 3. Skuteczność klasyfikacji dla metryki Euklidesowej dla podziału zbioru 70% treningowe/30% testowe.

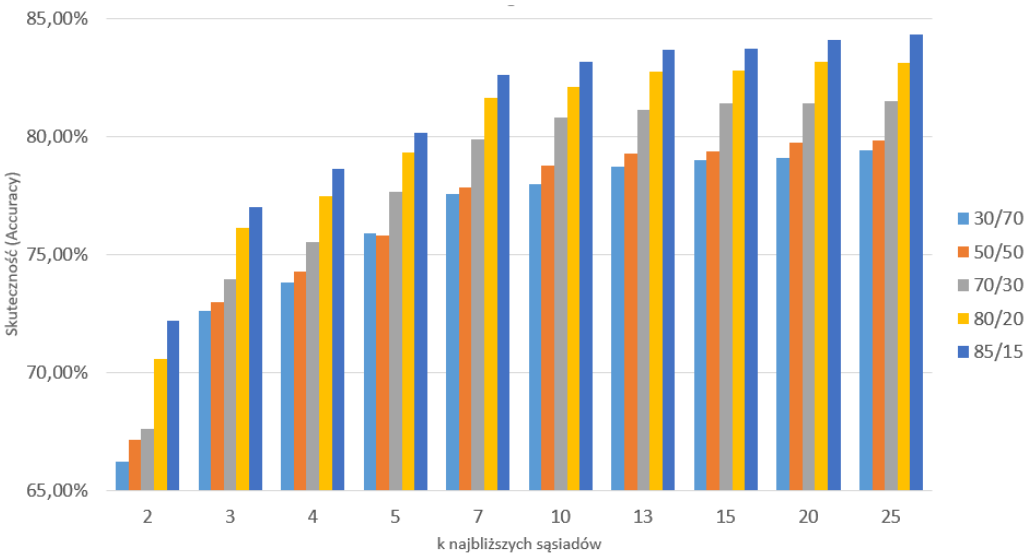
k	Accuracy [%]
2	70,56
3	76,11
4	77,47
5	79,31
7	81,62
10	82,10
13	82,73
15	82,80
20	83,17
25	83,09

Tabela 4. Skuteczność klasyfikacji dla metryki Euklidesowej dla podziału zbioru 80% treningowe/20% testowe.

k	Accuracy [%]
2	72,16
3	76,99
4	78,60
5	80,16
7	82,59
10	83,13
13	83,67
15	83,72
20	84,06
25	84,30

Tabela 5. Skuteczność klasyfikacji dla metryki Euklidesowej dla podziału zbioru 85% treningowe/15% testowe.

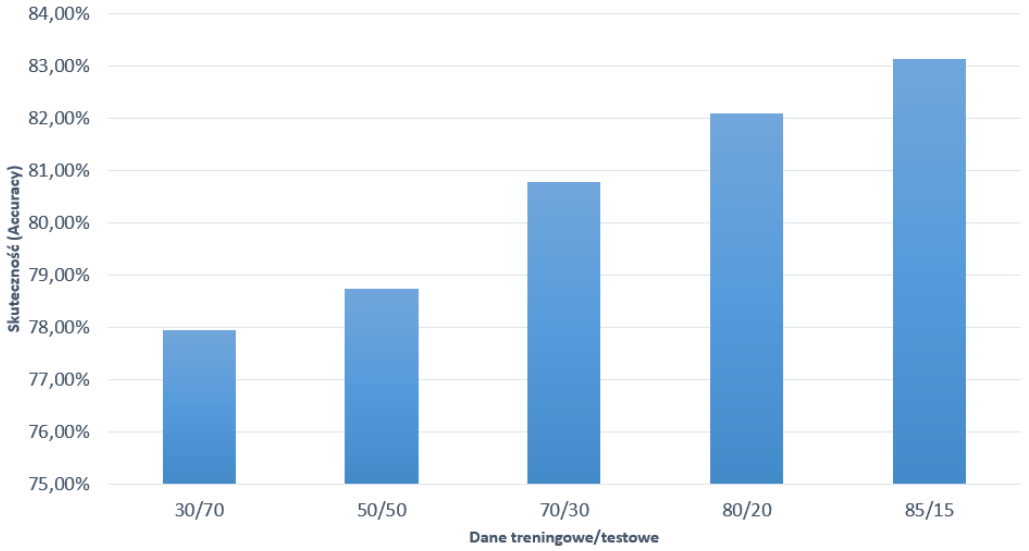
6. Dyskusja



Rysunek 1. Dane z tabel 1-5 zebrane na wykresie.

Dane treningowe/testowe	**Accuracy [%]**
30/70	77,95
50/50	78,74
70/30	80,77
80/20	82,10
85/15	83,13

Tabela 6. Zależność Accuracy od pięciu wartości proporcji podziału zbioru dla k=10.



Rysunek 2. Wykres przedstawiający zależność Accuracy od pięciu wartości proporcji podziału zbioru dla k=10.

Sekcja ta powinna zawierać dokładną interpretację uzyskanych wyników

eksperymentów wraz ze szczegółowymi wnioskami z nich płynącymi. Najcenniejsze są, rzecz jasna, wnioski o charakterze uniwersalnym, które mogą być istotne przy innych, podobnych zadaniach. Należy również omówić i wyjaśnić wszystkie napotakane problemy (jeśli takie były). Każdy wniosek powinien mieć poparcie we wcześniej przeprowadzonych eksperymentach (odwołania do konkretnych wyników). Jest to jedna z najważniejszych sekcji tego sprawozdania, gdyż prezentuje poziom zrozumienia badanego problemu.

7. Wnioski

W tej, przedostatniej, sekcji należy zamieścić podsumowanie najważniejszych wniosków z sekcji poprzedniej. Najlepiej jest je po prostu wypunktować. Znów, tak jak poprzednio, najistotniejsze są wnioski o charakterze uniwersalnym.

Literatura

[1] *http://home.agh.edu.pl/ horzyk/lectures/miw/KNN.pdf* [dostęp 22.03.2020]
[2] *https://pl.wikipedia.org/wiki/Odleg%C5%82o%C5%9B%C4%87_Minkowskiego* [dostęp 22.03.2020]
[3] *https://en.wikipedia.org/wiki/Canberra_distance* [dostęp 22.03.2020]
[4] *https://ftims.edu.p.lodz.pl/pluginfile.php/132368/mod_folder/content/0/ksr-wyklad-2009.pdf?forcedownload=1* [dostęp 22.03.2020]